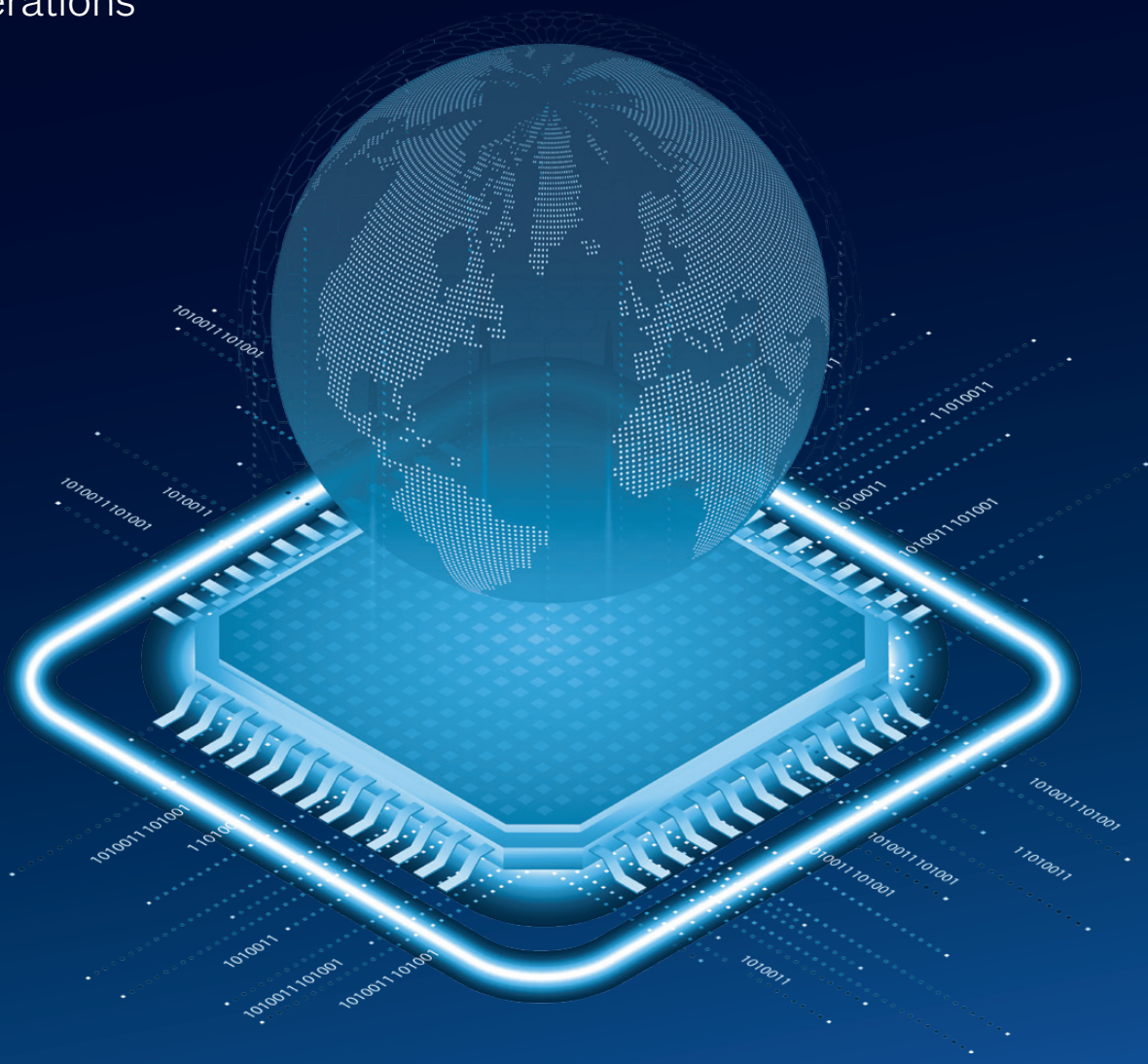
McKinsey & Company

McKinsey on Semiconductors

Creating value, pursuing innovation, and optimizing operations



Number 9, March 2024

McKinsey on Semiconductors is written by experts and practitioners in McKinsey's Semiconductors Practice along with other McKinsey colleagues.

To send comments or request copies, email McKinsey_on_Semiconductors@McKinsey.com.

Cover Image:
© Kulpreya Chaichatpornsuk/Getty Images

Table of contents

Introduction

Semiconductors are once again making headlines. Although demand for chips and devices was down in the first half of 2023, it is expected to recover through 2024. What's more, the long-term outlook appears very promising, thanks to the growth of artificial intelligence, electric vehicles, and other innovations. The global market for semiconductors could reach more than $1 trillion by 2030, up from $600 billion in 2021.

While these projections provide reason for cheer, they also raise important questions about whether the industry is poised to meet demand and accelerate technological advances. Multiple factors are increasing uncertainty and complicating strategic decisions, including geopolitical tensions, the ongoing race for technology leadership, and the ever-present fear of overbuilding capacity. Change is occurring so rapidly that even the best analysts have difficulty making concrete predictions. Living in a multiscenario world is becoming the "new normal" for boardrooms.

This issue of *McKinsey on Semiconductors* provides a snapshot of the industry's prospects and strategies for keeping a competitive edge while navigating uncertainty. The topics covered include a potential explosion in semiconductor demand arising from the generative AI infrastructure build-out, strategies for building fabrication plants (fabs) in "new" regions that are outside existing ecosystems, advanced chip packaging's appeal for premium customers, and technology innovations that appeal to the automotive industry, which is one of the semiconductor sector's fastest-growing markets. We also examine two issues that are crucial during this expansion period: the need to increase sustainability at fabs and strategies for attracting critical talent amid fierce competition for employees with strong technology skills.

As you read these articles, we hope you will find novel approaches to your top challenges, as well as new opportunities for innovation. We would be happy to elaborate on any topics covered in these articles, or on other areas of interest, as you chart your company's path forward.

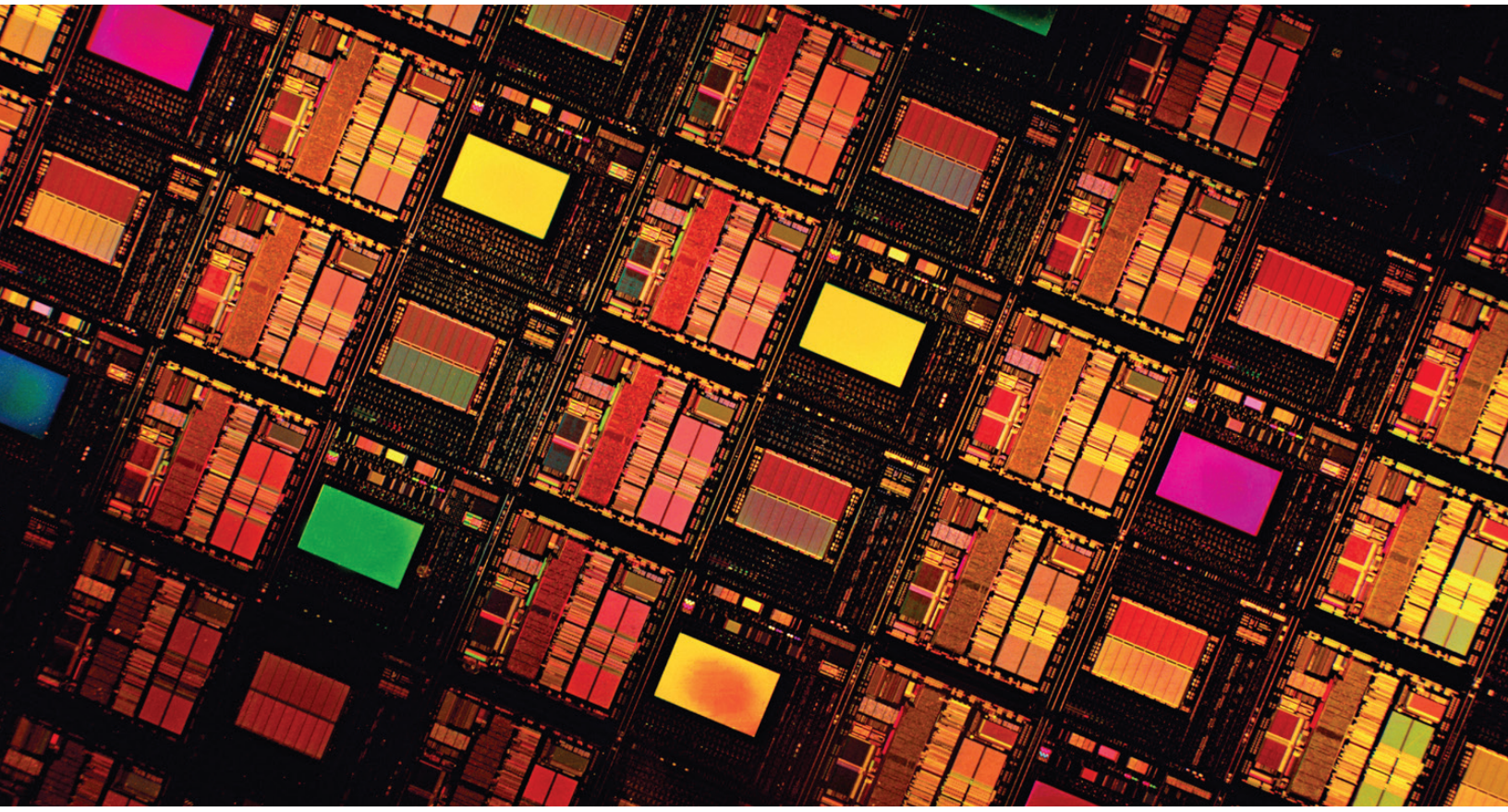
Ondrej Burkacky
Senior partner,
Munich

Bill Wiseman
Senior partner,
Seattle

Generative AI: The next S-curve for the semiconductor industry?

The surge of interest in and use of generative AI translates to higher demand for semiconductors, pushing the industry to innovate faster and produce more capable and efficient chips.

This article is a collaborative effort by Ondrej Burkacky, Mark Patel, Klaus Pototzky, Diana Tang, Rutger Vrijen, and Wendy Zhu, representing views from McKinsey's Semiconductor Practice.



As generative AI (gen AI) applications such as ChatGPT and Sora take the world by storm, demand for computational power is skyrocketing. The semiconductor industry finds itself approaching a new S-curve—and the pressing question for executives is whether the industry will be able to keep up.

Leaders are responding by committing substantial capital expenditures to expand data centers and semiconductor fabrication plants (fabs) while concurrently exploring advancements in chip design, materials, and architectures to meet the evolving needs of the gen AI–driven business landscape.

To guide semiconductor leaders through this transformative phase, we have developed several scenarios for gen AI's effect in the B2B and B2C markets. Every scenario involves a massive increase in compute—and thus wafer—demand. These scenarios focus on the data centers while acknowledging that implications for edge devices such as smartphones exist but on a much smaller scale.

The demand scenarios, developed from McKinsey analysis, are based on the wafer output that the semiconductor industry could potentially deliver, given constraints such as capital and equipment. While even scenarios that are more ambitious are plausible, the implications for the required number of fabs and the energy supply necessary for the data centers will make them unlikely.

This article will discuss the estimated wafer demand of high-performance components, including logic, memory, data storage chips, and the corresponding number of fabs needed to supply them. Equipped with this information, industry stakeholders can strategically plan and allocate resources to address the burgeoning demand for compute power, ensuring the scalability and sustainability of their operations in the years to come.

Components of gen AI compute demand

The surge in demand for AI and gen AI applications comes with a proportional increase in compute demand. However, it is essential for semiconductor leaders to understand the origins of this demand and how gen AI will be applied. We expect to see two different types of applications for gen AI: B2C and B2B use cases. Within both the B2C and B2B markets, the demand for gen AI can be categorized into two main phases: training and inference. Training runs usually require a substantial amount of data and are compute-intensive. Conversely, inference usually requires much lower compute for each run of a use case.

To empower semiconductor leaders to navigate the intricacies and demands of these markets, we outline six use case archetypes for B2B compute demand and their corresponding compute cost to serve and concurrent level of gen AI value creation.

Six B2B use case archetypes for gen AI application and workload

McKinsey analysis estimates that B2C applications will account for about 70 percent of gen AI compute demand because they include the workload from basic consumer interactions (for example, drafting emails) and advanced user interactions (for example, creating visuals from text). B2B use cases are expected to make up the other approximately 30 percent of the demand. These include use cases such as advanced content creation for businesses (for example, gen AI–assisted code creation), addressing customer inquiries, or generating standard financial reporting.

B2B applications across industry verticals and functions fall into one of six use case archetypes:

- coding and software development apps that interpret and generate code
- creative content–generation apps that write documents and communication (for example, to generate marketing material)
- customer engagement apps that cover automated customer service for outreach, inquiry, and data collection (for example, addressing customer inquiries via a chatbot)
- innovation apps that generate product and materials for R&D processes (for example, designing a candidate drug molecule)

— simple concision apps that summarize and extract insights using structured data sets (for example, to generate standard financial reports)

— complex concision apps that summarize and extract insights using an unstructured or large data set (for example, to synthesize findings in clinical images such as MRI or CT scans)

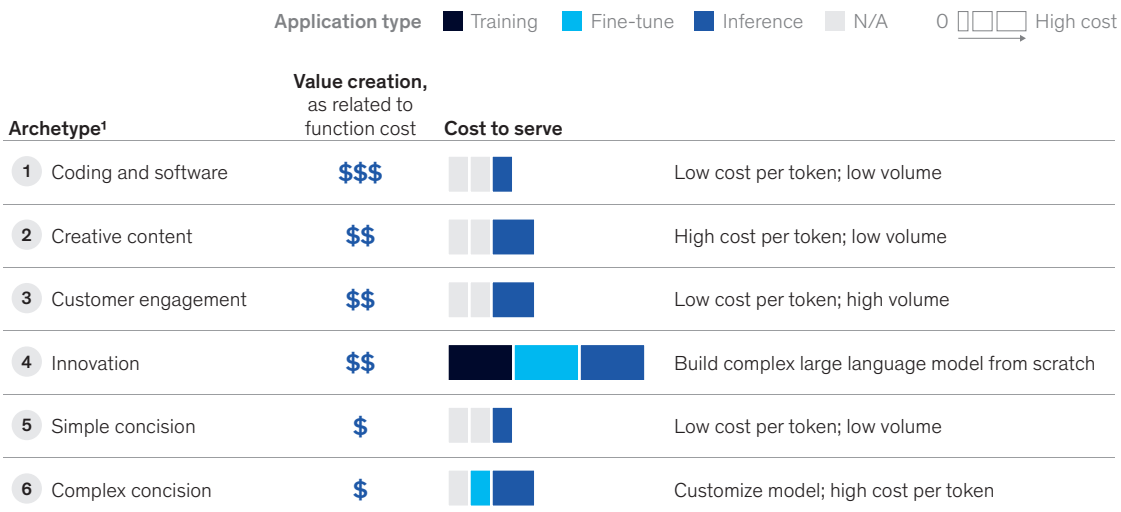
McKinsey has organized these six diverse and complex B2B use cases according to their compute cost to serve and concurrent gen AI value creation (Exhibit 1). By defining the cost to serve and value creation, decision makers can more adeptly navigate the specifics of B2B use cases and make well-informed choices when adopting them. At its core, the analysis of compute cost to serve comprises training, fine-tuning, and inferencing costs. The analysis also encompasses a hyperscaler's infrastructure as a service (IaaS) margin, which includes compute hardware, server components, IT infrastructure, power consumption, and estimated talent costs. Gen AI value creation is gauged through metrics such as productivity improvement and labor cost savings.

Gen AI demand scenarios

As organizations navigate the complexities of adopting gen AI, strategic utilization of these archetypes becomes imperative. Factors such as the economics of gen AI adoption, algorithm efficiency, and continual hardware advancements at both component and system levels further influence adoption of gen AI and technological progress. Three demand scenarios—base, conservative, and accelerated—represent the possible outcomes of gen AI demand for B2B and B2C applications. The base scenario is informed by a set of required assumptions, such as consistent technological advancements and rapid adoption, supported by business models that cover the capital and

Exhibit 1

B2B use cases are defined by their value creation and cost to serve.



[1]The six B2B application archetypes include the following: 1) coding and software development apps that interpret and generate code; 2) creative content–generation apps that write documents and communication (for example, to generate marketing material); 3) customer engagement apps that cover automated customer service for outreach, inquiry, and data collection (for example, addressing customer inquiries via a chatbot); 4) innovation apps that generate product and materials for R&D processes (for example, designing a candidate drug molecule); 5) simple concision apps that summarize and extract insights using structured data sets (for example, to generate standard financial reports); and 6) complex concision apps that summarize and extract insights using unstructured or large data sets (for example, to synthesize findings in clinical images such as MRI or CT scans).
Source: "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023; McKinsey analysis

McKinsey & Company

operating costs of gen AI training and inference. The conservative and accelerated adoption scenarios represent adoption upside and downside, respectively.

McKinsey analysis estimates that by 2030 in the base scenario, the total gen AI compute demand could reach 25x10^{30} FLOPs (floating point operations), with approximately 70 percent from B2C applications and 30 percent from B2B applications (Exhibit 2).

B2C compute demand scenarios

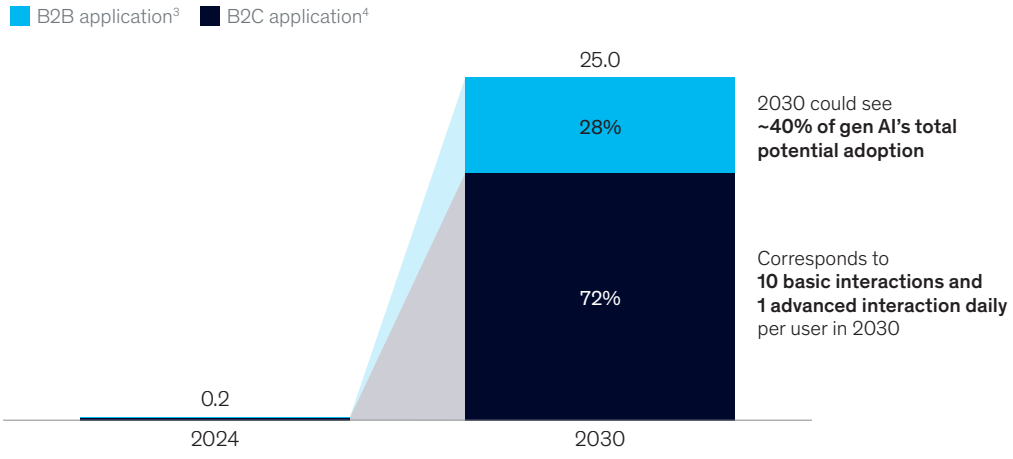
B2C compute demand is driven by the number of consumers who engage with gen AI, their level of engagement, and its compute implication. Specifically, B2C inference workloads are determined by the number of gen AI interactions per user, the number of gen AI users, and FLOPs per basic and advanced user interaction. Training workloads are determined by the number of training runs per year, the number of gen AI model providers, and FLOPs per training run by different gen AI models (for example, a state-of-the-art model such as GPT-4 in 2023 and smaller or prior generations of models). For all scenarios, it is essential that companies can develop a sustainable business model.

Base adoption. By 2030, the expected average number of daily interactions per smartphone user (with one interaction being a series of prompts) is ten for basic consumer applications, such as creating an email draft. The other expected average number is for advanced consumer applications, such as creating longer texts or synthesizing complex input documents. By using current numbers from online and application-based search queries, McKinsey analysis estimates the number of

Exhibit 2

In our base scenario, realized demand of generative AI is about 70 percent for B2C and 30 percent for B2B.

Total annual FLOP[1] demand for B2C and B2B applications, in QFLOPs[2]



[1]FLOP = floating point operation.
[2]QFLOPs (quettaFLOPs) = 10^{30} FLOPs.
[3]In 2030, 5 archetypes are expected to be adopted widely because cost to serve is lower than willingness to pay: coding and software, creative content, customer engagement, innovation, and simple concision. One archetype is expected not to be adopted at scale: complex concision.
[4]In 2030, 90% of B2C use cases are basic queries (eg, drafting an email) and 10% are complex (eg, creating a visual from text).

McKinsey & Company

interactions to be approximately twice the forecast daily number of online search queries (approximately 28 billion) in 2030. The underlying assumptions that will enable the base B2C scenario are steady technological advancements, favorable regulatory developments, and continuously growing user acceptance.

Conservative adoption. This scenario could involve cautious adoption from consumers due to ongoing concerns related to data privacy, regulatory developments, and only incremental improvements in the technology, which would lead to half the number of interactions of the base case.

Accelerated adoption. This scenario suggests a high degree of trust in the technology and widespread user acceptance. Drivers for this scenario could be attractive new business models, substantial technological advancements, and compelling user experiences. These drivers could lead to a higher adoption rate (150 percent) of the number of interactions for consumer applications in the base case.

B2B demand scenarios

The adoption of gen AI use cases in the B2B sector is significantly influenced by the sufficiency and cost of semiconductor chip supply. Enterprises must be capable of rationalizing their investment in compute infrastructure, ensuring that the cost of service is lower than the company's willingness to pay. For these B2B demand scenarios, McKinsey analysis assumes that the willingness to pay corresponds to approximately 20 percent of the total value creation.

In the context of B2B use cases, McKinsey analysis indicates that of the six use case archetypes, only five are economically viable for a broad adoption (Exhibit 3). The sixth archetype, complex concision, is not expected to be adopted broadly due to limited value creation compared to its cost through administrative labor cost savings, coupled with a significant consumption of compute power in analyzing complex and unstructured data inputs.

Base adoption. The base scenario assumes a midpoint adoption rate spanning eight to 28 years, indicating that B2B use cases achieve 90 percent adoption in 18 years.[1] Furthermore, McKinsey analysis assumes that businesses will realize value beginning in 2024. Securing investments for manufacturing capacity, manufacturing wafers, provisioning compute capacity, and training people to use new services all take time. As such, we assume a lead time of approximately two years in the manufacturing of wafers before value can be captured. This business realization is expected to produce approximately 25 percent of value captured by 2030 for the economically viable use cases. In this scenario, we assume the additional value from all small-scale improvements in labor productivity follow the same overall ratio as the calculated value potential from the six use case archetypes.

Conservative adoption. This scenario assumes an approximately 90 percent adoption rate over 28 years, yielding only approximately 15 percent in value capture by 2030. This deceleration could be attributed to a confluence of factors, including—but not limited to—regulatory constraints, data privacy concerns, and data processing challenges.

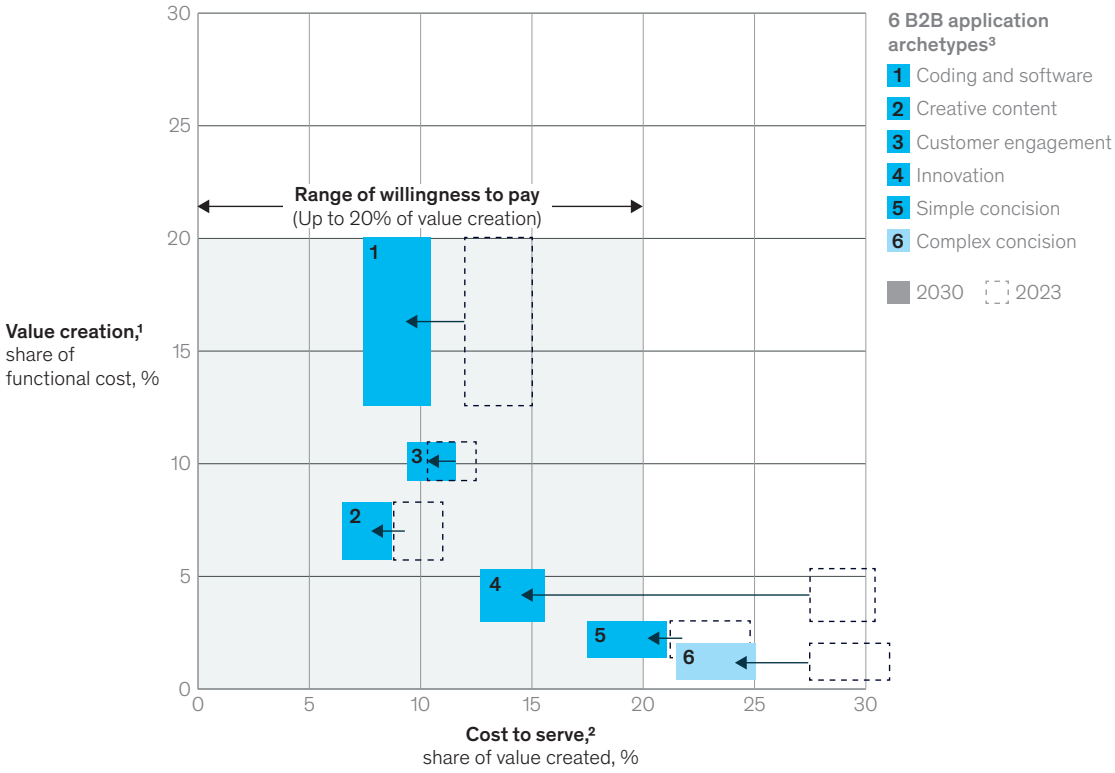
Accelerated adoption. This scenario assumes an approximately 90 percent adoption rate in about 13 years. This acceleration is contingent upon catalysts such as attractive business models, rapid technological advancement, or favorable regulations. For example, disruptive hardware architectures will substantially reduce the cost to serve. Additionally, enhancements to the process of software validation may significantly boost the efficiency of gen AI solutions. Factors such as these may expedite the adoption curve and cause a notable uptick in gen AI implementation in the semiconductor industry by 2030.

[1] "Harnessing automation for a future that works," McKinsey Global Institute, January 12, 2017.

Exhibit 3

We estimate only five out of six use case archetypes will be economically viable and assumed to be widely adopted by 2023.

Economics per B2B archetype



Note: Use cases where a portion of variable costs (eg, inference per token) are higher will see larger shifts between 2023 and 2030.
[1]Based on estimated cost saving or revenue uplift across approximately 60 use cases, as identified in "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.
[2]2023 and 2030 costs to serve are estimated based on compute cost (capital expenditures for training, fine-tuning, and inferencing costs, as well as operating expenditures, including power), estimated talent cost, and assumed hyperscaler infrastructure as a service (IaaS) margin.
[3]The six B2B application archetypes include the following: 1) coding and software development apps that interpret and generate code; 2) creative content–generation apps that write documents and communication (for example, to generate marketing material); 3) customer engagement apps that cover automated customer service for outreach, inquiry, and data collection (for example, addressing customer inquiries via a chatbot); 4) innovation apps that generate product and materials for R&D processes (for example, designing a candidate drug molecule); 5) simple concision apps that summarize and extract insights using structured data sets (for example, to generate standard financial reports); and 6) complex concision apps that summarize and extract insights using unstructured or large data sets (for example, to synthesize findings in clinical images such as MRI or CT scans).
Source: "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023; McKinsey analysis

McKinsey & Company

Gen AI data center infrastructure and hardware trends

Along with considering scenarios for gen AI compute demand, semiconductor leaders will need to adapt to changes in underlying hardware and infrastructure, mainly to data center infrastructure, servers, and semiconductor chips.

Data center infrastructure

Gen AI applications typically run on dedicated servers and in data centers. At first glance, AI data centers might look similar to traditional data centers, but there are considerable differences (see sidebar "Components of an AI server").

Rack densities—that is, the power consumed by a cabinet of servers—demonstrate the biggest difference between traditional and AI data centers. General-purpose data centers have rack power densities of five to 15 kilowatts (kW), whereas AI training workloads can consume 100 kW—or, in some cases today, up to 150 kW. This number is expected to increase, with some experts estimating power densities of up to 250 kW or even 300 kW in the next few years.[2]

Additionally, as rack power density rises, rack cooling will switch from air-based cooling to liquid cooling. Direct-to-chip liquid cooling and full-immersion cooling will also require new server and rack designs to accommodate for additional weights.

Servers

In response to the increasing demand for computational power, servers will employ high-performance graphics processing units (GPUs) or specialized AI chips, such as application-specific integrated circuits (ASICs), to efficiently handle gen AI workloads through parallel processing. Today, infrastructure for gen AI training and inference is expected to bifurcate as inference's compute demand becomes more specific to the use case and requires much lower cost to be economical.

Training. Training server architecture is expected to be similar to today's high-performance cluster architectures in which all servers in a data center are connected to high-bandwidth, low-latency connectivity. The prevailing high-performance gen AI server architecture uses two central processing units (CPUs) and eight GPUs for compute. In 2030, most training workloads are expected to be executed using this type of CPU+GPU combination. A transition to system-in-a-package design for GPUs and AI accelerators is also expected, with both architectures expected to coexist.

Inference. Current inference workloads run on infrastructure that is similar to the training workload. As gen AI consumer and business adoption increases, the workload is expected to shift to mostly inference, which favors specialized hardware due to lower cost, higher energy efficiency, and faster or better performance for highly specialized tasks. In 2030, we expect more inference-specific AI servers using a combination of CPUs and several purpose-built AI accelerators that use ASICs.

Gen AI wafer demand on the semiconductor industry

McKinsey analysis estimates the wafer demand of high-performance components based on compute demand and its hardware requirement: logic chips (CPUs, GPUs, and AI accelerators), memory chips (high-bandwidth memory [HBM] and double data rate memory [DDR]), data storage chips (NAND ["not-and"] chips), power semiconductor chips, optical transceivers, and other components. In the following sections, we will look more closely at logic, HBM, DDR, and NAND chips. Beyond logic and memory, we anticipate that there will be an increase in demand for other device types. For instance, power semiconductors will be in higher demand because gen AI servers consume higher amounts of energy. Another consideration is optical components, such as those used in communications, which are expected to transition to optical technologies over time. We have already seen this transition for long-distance networking and backbones that reduce energy consumption while increasing data transmission rates. To spur innovation in almost all areas of the industry, it is necessary to combine these new requirements with the high level of investment anticipated (see sidebar "Pursuing innovation in semiconductors to capture generative AI value").

Logic chips

Logic chip demand depends on the type of gen AI compute chip and type of server for training and inference workloads. As discussed earlier, by 2030, we anticipate the majority of gen AI compute demand in FLOPs to come from inference workloads. Currently, there are three types of AI

[2] Charlotte Trueman, "Stack Infrastructure to support AI workloads requiring up to 300kw-per-rack," DatacenterDynamics, January 8, 2024.

Components of an AI server

AI data centers and servers differ from traditional models. There are nine components of the AI server that are most relevant to semiconductor leaders (exhibit).

— *CPU (central processing unit).* The CPU manages system-level functions, coordinates data flow, and executes tasks that require a more generalized computing approach. Collaboration between CPUs and specialized processors ensures a balanced and efficient operation, optimizing the utilization of each component's strengths within the AI server.

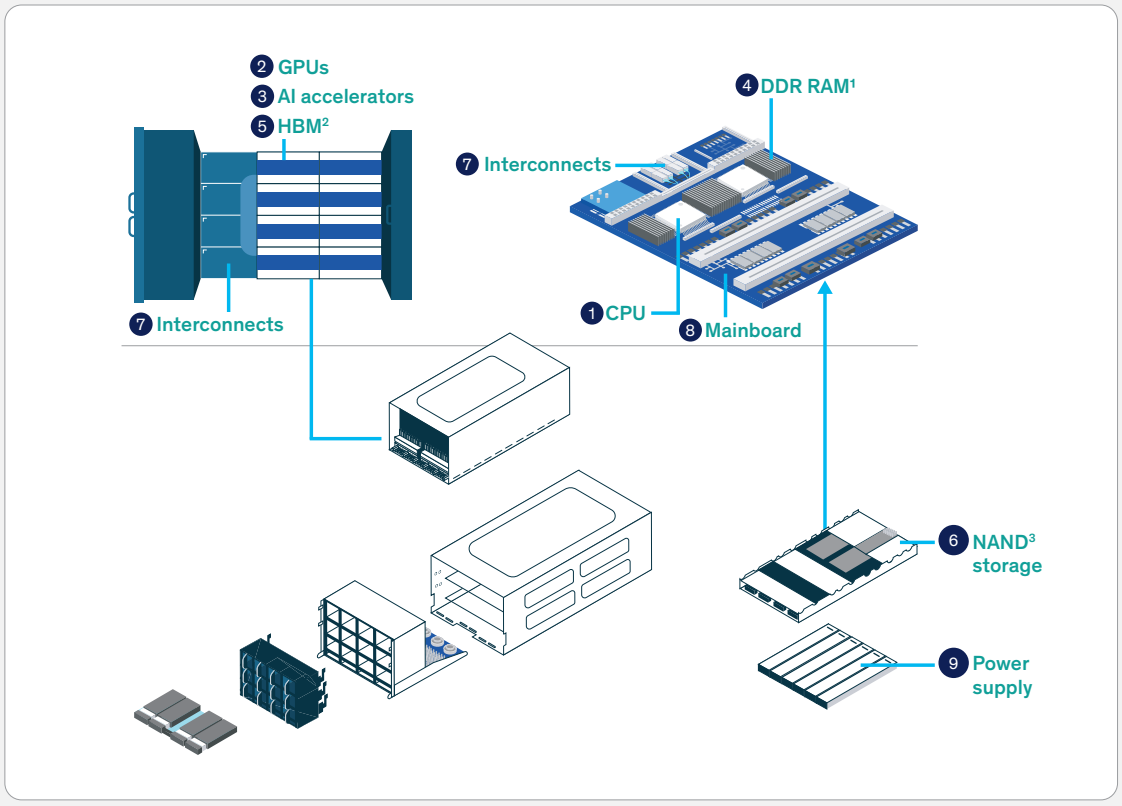
— *GPU (graphics processing unit).* The GPU is a specialized processor designed to handle complex mathematical computations in parallel, making it an essential component in AI data centers for accelerating training and inference compute.

— *AI accelerator.* This is a specialized semiconductor component designed to accelerate AI workloads by performing high-speed computations and optimizing the cost and performance of AI algorithms in data centers.

Exhibit

An AI server is made up of numerous integral components.

Illustrative breakdown of one AI server within the racks of the server room



[1]Double data rate random-access memory.
[2]High-bandwidth memory.
[3]"Not-and" memory.

McKinsey & Company

— *DDR RAM (double data rate random access memory).* A variant of dynamic random-access memory (DRAM), DDR memory provides high-speed, volatile memory, facilitating rapid data access for enhanced overall system performance.

— *HBM (high-bandwidth memory).* A variant of DRAM, HBM is specifically built for very high-bandwidth use cases, such as AI training and inference, achieving speeds of more than ten times the standard DRAM.

— *NAND ("not-and") storage.* This is used to store the operating system, model, user input, and other components.

— *Interconnects.* Equipped with optical transceivers, interconnects enable seamless communication between compute components, ensuring efficient data exchange.

— *Mainboard.* The mainboard serves as the central hub, coordinating the collaboration of various components, all powered by a reliable power supply unit and maintained at optimal conditions by cooling fans. Encased in a well-structured chassis, these components collectively form the sophisticated architecture essential for meeting the computational demands of generative AI within a dedicated data center environment.

— *Power supply unit.* The AI server is equipped with several power supply units with redundancy to reduce risk of failure.

servers that can manage inference and training workloads: CPU+GPU, CPU+AI accelerator, and fusion CPU+GPU. Today, CPU+GPU has the best availability and is used for inference and training workloads. In 2030, AI accelerators with ASIC chips are expected to serve the majority of workloads because they perform optimally in specific AI tasks. On the other hand, GPU and fusion servers are ideal for handling training workloads due to their versatility in accommodating various types of tasks (Exhibit 4).

In 2030, McKinsey estimates that the logic wafer demand from non–gen AI applications will be approximately 15 million wafers. About seven million of these wafers will be produced using technology nodes of more than three nanometers, and approximately eight million wafers will be produced using nodes equal to or less than three nanometers. Gen AI demand would require an additional 1.2 million to 3.6 million wafers produced using technology nodes equal to or less than three nanometers. Based on current logic fab planning,[3] it is anticipated that 15 million wafers using technology nodes equal to or less than seven nanometers can be produced in 2030. Thus, gen AI demand creates a potential supply gap of one million to about four million wafers using technology nodes equal to or less than three nanometers. To close the gap, three to nine new logic fabs will be needed by 2030 (Exhibit 5).

DDR and HBM

Gen AI servers use two types of DRAM: HBM, attached to the GPU or AI accelerators, and DDR RAM, attached to the CPU. HBM has higher bandwidth but requires more silicon for the same amount of data.

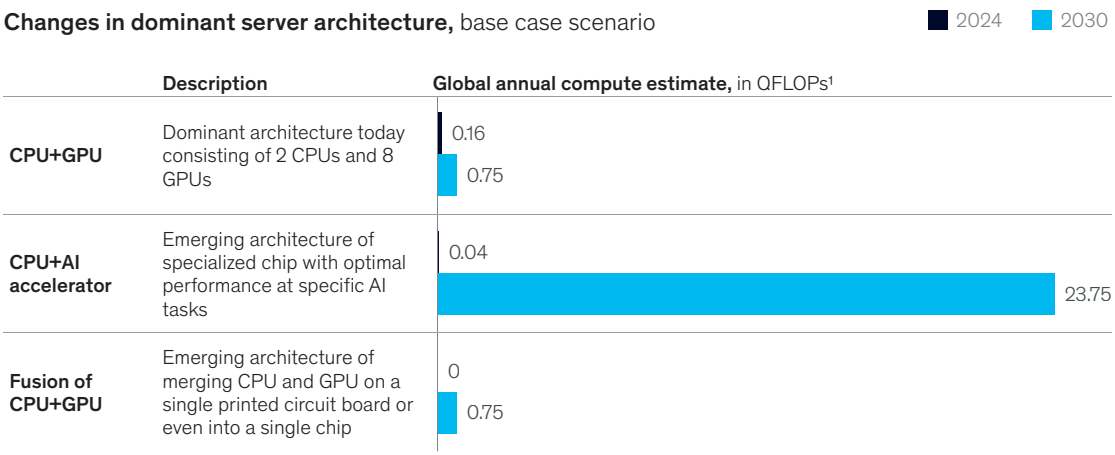
As transformer models grow larger, gen AI servers have been expanding memory capacity. However, the growth in memory capacity is not straightforward, posing challenges to hardware and software design. First, the industry faces a memory wall problem, in which memory capacity and bandwidth are the bottleneck for system-level compute performance. How the industry will tackle the memory wall problem is an open question. Static

[3] We expect a small free capacity—that is, unused or available wafers—of around 0.5 million wafers in 2030, according to current fab planning and non–gen AI logic wafer demand. However, this free capacity is unlikely to be equal to or less than three-nanometer nodes, required by gen AI demand. Therefore, our logic supply shortage estimate did not consider free capacity.

Exhibit 4

Server architecture is estimated to shift toward CPUs with AI accelerators by 2030.

Changes in dominant server architecture, base case scenario



[1]FLOP = floating point operation. QFLOPs (quettaFLOPs) = 10^{30} FLOPs.

McKinsey & Company

Pursuing innovation in semiconductors to capture generative AI value

Even though the field of generative AI is emerging, we have seen an uptick in innovative technologies and solutions in the past two to three years. To spur innovation, large amounts of global investment are needed across the value chain in all three scenarios. If all players invest in innovation, their efforts could reduce costs, optimize compute efficiency, or increase capacities to meet demand. Examples of this could include the following:

- new algorithm designs to reduce computational requirements, in terms of both number of operations and memory demand—for example, as seen in the invention of different transformer models, which represented a new approach to designing algorithms aimed at decreasing computational demands
- new chip architectures that achieve higher performance using the same area of silicon (several start-ups have already developed such an architecture)
- increased memory density of chips to increase their storage capacity (for example, by using data compression similar to Linux's zram but implemented on the chip)
- improved high-speed networks between servers to provide faster access to the memory of other servers, thereby reducing the need for storing local duplicates of data
- optimized software or compilers to improve system-level infrastructure compute efficiency

random-access memory (SRAM) is tested in various chips to increase the near-compute memory, but its high cost limits wide adoption. For example, future algorithms may require less memory per inference run, slowing down total memory demand growth. Second, AI accelerators are lighter in memory compared to CPU+GPU architecture and may become more popular by 2030 when inference workloads flourish. This could mean a potentially slower growth in memory demand.

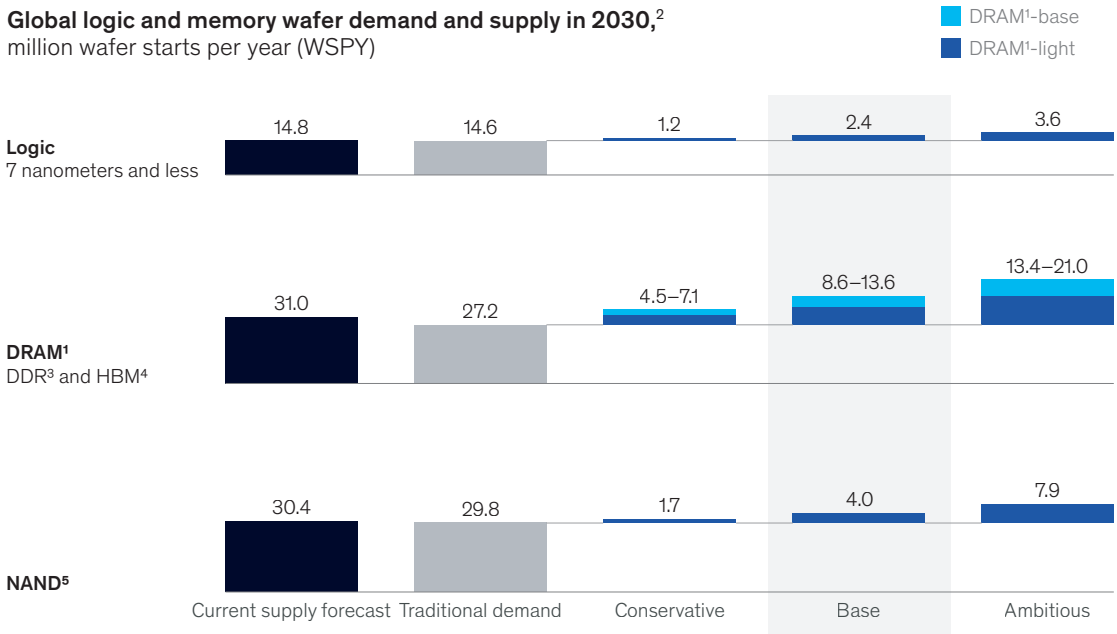
Given these uncertainties, we consider two DRAM demand scenarios in addition to the base, conservative, and accelerated adoption scenarios: a "DRAM light" scenario, in which AI accelerators remain memory-light compared to GPU based systems, and a "DRAM base" scenario, in which AI accelerator–based systems catch up to GPU-based systems in terms of DRAM demand.

By 2030, we expect DRAM demand from gen AI applications to be five to 13 million wafers in the DRAM light scenario, translating to four to 12 dedicated fabs. In the DRAM base scenario, DRAM demand would be seven to 21 million wafers, translating to six to 18 fabs. The wide range of values reflects the challenges associated with reducing the memory requirements per device.

Exhibit 5

By 2030, generative AI will increase demand for wafers significantly.

Global logic and memory wafer demand and supply in 2030,[2]
million wafer starts per year (WSPY)



[1]Dynamic random-access memory.
[2]Two DRAM demand scenarios are considered. DRAM base case: AI accelerators and CPU+GPU systems have same memory content per server. DRAM light case: CPU+AI accelerator systems have lower memory content per server (50% less) than CPU+GPU systems.
[3]Double data rate memory.
[4]High-bandwidth memory.
[5]NAND = "not-and," a type of memory.
Source: *World fab forecast*, SEMI, December 12, 2023; McKinsey analysis

McKinsey & Company

NAND memory

NAND memory is used for data storage—for instance, for the operating system, user data, and input and output. In 2030, NAND demand will likely be driven by dedicated data servers for video and multimodel data. This data will require substantial storage (for example, for training on high-resolution video sequences and retrieving data during inference). We expect the total NAND demand to be two to eight million wafers, corresponding to one to five fabs. Given that the performance requirement for NAND storage of gen AI will be the same as in current servers, fulfilling this demand will be less challenging compared to logic and DRAM.

Other components

The rising compute demand will create additional demand for many other chip types. Two types are particularly noteworthy:

High-speed network and interconnect. Gen AI requires high-bandwidth and low-latency connectivity between the servers and between the various components of the servers. A larger amount of network interfaces and switches are required to create all the connections. Today, these interlinks are mostly copper-based, but optical connectivity is expected to gain share with rising bandwidth and latency requirements.

Power semiconductors. AI servers need a large amount of electricity and might consume more than 10 percent of global electricity in 2030. This requires many power semiconductors within the server and on the actual devices.

The surge in demand for gen AI applications is propelling a corresponding need for computational power, driving both software innovation and substantial investment in data center infrastructure and semiconductor fabs. However, the critical question for industry leaders is whether the semiconductor sector will be able to meet the demand. To meet this challenge, semiconductor leaders should consider which scenario they believe in. Investment in semiconductor manufacturing capacity and servers is costly and takes time, so careful evaluation of the landscape is essential to navigating the complexities of the gen AI revolution and developing a view of its impact on the semiconductor industry.

Ondrej Burkacky is a senior partner in McKinsey's Munich office, where **Klaus Pototzky** is an associate partner; **Mark Patel** is a senior partner in the San Francisco office; **Diana Tang** is an associate partner in the Silicon Valley office, where **Rutger Vrijen** is a partner; and **Wendy Zhu** is an associate partner in the Denver office.

The authors wish to thank Stefan Burghardt, Orhan Celiker, Sebastian Göke, Yang Han, Demi Liu, Lorenzo Mambrini, and Paul Wittmer for their contributions to this article.

Exploring new regions: The greenfield opportunity in semiconductors

Three factors—supply chain security, sustainability, and subsidies—feature prominently as semiconductor companies expand into new countries or regions.

This article is a collaborative effort by Ondrej Burkacky, Matteo Mancini, Mark Patel, Giulietta Poltronieri, and Taylor Roundtree, representing views from McKinsey's Semiconductor Practice.

By now, it's old news: semiconductor demand is growing. What's new, however, is how several global trends—including the rise of artificial intelligence, vehicle electrification, and autonomous driving—will broaden demand and take it to even greater heights over the next decade.

Already, many incumbents and new entrants in semiconductor manufacturing are expanding their operations to capture the increasing opportunities along the entire value chain, including those related to wafer manufacturing, chemical supply, packaging, capital equipment, and other areas. Globally, companies plan to invest about $1 trillion in semiconductor fabs through 2030. Most investment is concentrated in Asia and the United States, but funding for European projects is also increasing.

For some semiconductor companies, expansion efforts may involve building fabs in regions or countries where they have not previously operated. In the past, companies that explored such opportunities sought locations with established semiconductor ecosystems that met some basic requirements: sufficient and stable energy and water supplies, a pool of potential employees with technical skills, the right infrastructure, and a solid transportation network. Well-known ecosystems that fall into this category include Taiwan's Hsinchu Science Park and Germany's Silicon Saxony. (For more information on these locations, see sidebar, "Thriving semiconductor ecosystems.")

When semiconductor companies consider expansion today, they still restrict their search to locations that meet their basic requirements. But simply meeting these requirements is not enough to guarantee investment. Instead, semiconductor companies now focus on three *S's*—sustainability, supply chain security, and subsidies—as they select new sites. Their shifting priorities reflect changes occurring in the world at large, including growing concern about climate change, geopolitical issues that are disrupting or slowing shipments, and economic uncertainty.

Given the value of the semiconductor industry, as well as its benefits to local economies, much is at stake as companies expand, both for the businesses themselves and for the regions or countries where they establish new sites. Here's a look at the industry's growth potential and the factors that may determine where new fabs are built over the next decade.

A thriving market that encourages expansion

The global market for semiconductors is projected to reach $1 trillion by 2030, up from $600 billion in 2021 (Exhibit 1). Although the wireless communication and computing sectors are currently undergoing some disruptions, such as lower demand for mobile phones in certain countries, they are expected to experience the strongest long-term growth, followed by the automotive and industrial sectors.

No single country or region dominates any segment of the supply chain, with the notable exception of Asia and its strong manufacturing hubs (Exhibit 2). What's more, no region or country has strong capabilities in every segment of the value chain, so

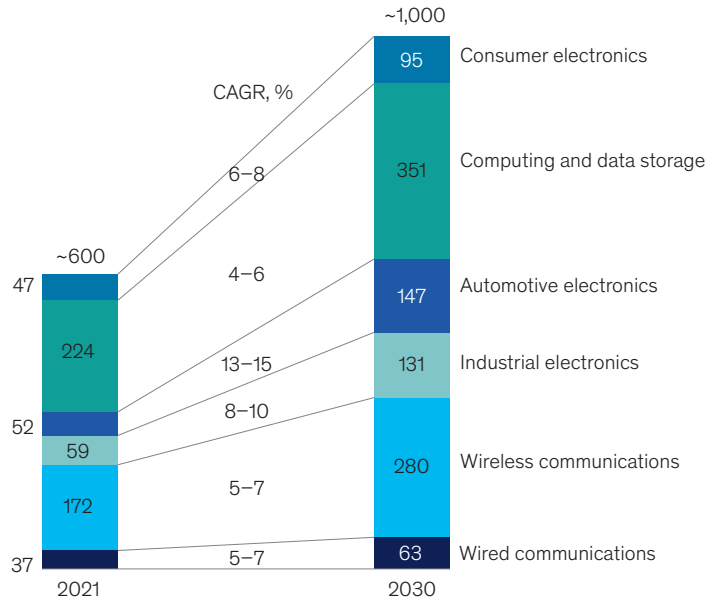
Thriving semiconductor ecosystems

Strong semiconductor ecosystems can be found worldwide, and they come in different shapes and sizes. One of the most outstanding is Hsinchu Science Park in Taiwan. Within just three square miles, Hsinchu contains more than 150 semiconductor companies and suppliers, more than 600 manufacturers, three universities, and more than 160,000 highly skilled full-time employees. Similarly, Silicon Saxony in Germany, which is the largest semiconductor cluster in Europe, contains more than 400 industry actors, universities, and research centers. Through coherent policies and robust ecosystem building, Saxony has more than doubled the number of employees in the semiconductor industry over the past 20 years. The industry projects that there will be about 100,000 workers in Saxony's semiconductor industry by 2030.

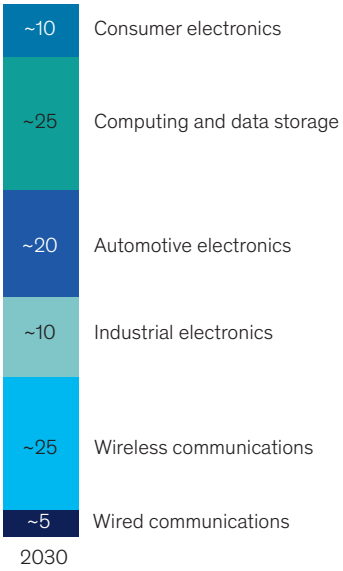
Exhibit 1

The semiconductor market is expected to reach $1 trillion in value by 2030.

Global semiconductor market, $ billion



Growth contribution per vertical, %

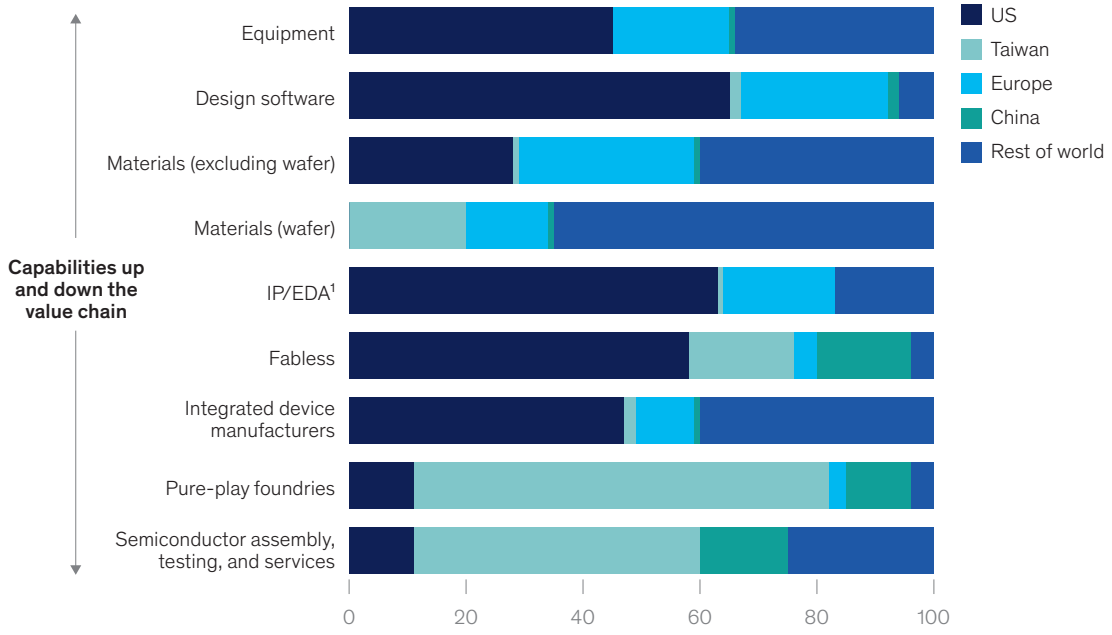


McKinsey & Company

Exhibit 2

While manufacturing is concentrated in a few locations, other stages of the value chain involve companies in multiple countries.

Sales based on company headquarters location, % share



[1]Intellectual property/electronic design automation.
Source: Gartner

McKinsey & Company

the end-to-end process for creating a semiconductor involves a global effort. A plant in Japan might cut silicon ingots into wafers, which are then shipped to the United States for fabrication into semiconductors. The next leg of the process could take them to Malaysia for sorting, cutting into dies, assembly, packaging, and testing. Finally, they might be sent to Singapore for incorporation into a finished product as a chip.

The dispersion of expertise gave semiconductor companies some freedom during past expansion efforts, since they could typically find multiple ecosystems that met their needs. Although they traditionally gravitated to areas with large, established semiconductor ecosystems, a few bucked the trend by venturing into new regions, as California-based Intel did in Oregon in 1974, Arizona in 1980, and Ireland in 1989. These moves conferred various advantages, such as lower utility prices and, in the case of Oregon, greater land availability.

Today, many semiconductor companies still focus on locations with large ecosystems—new fabs are now being built in Dresden, which is at the center of the Silicon Saxony semiconductor hub, for example—but there's a growing trend toward considering other countries or regions because they score high for sustainability, supply chain security, and subsidies. Here again, the global dispersion of expertise could be an advantage. Since no region or country dominates any step of the value chain, with the exception of Asia's pure-play foundries, companies will not be competing against a single ecosystem that dominates the industry.

To see how investment is shifting to new regions and countries, consider what's now occurring in the United States. While the country saw little fab construction over the past few decades, the value of US-based semiconductor projects that are now under way, announced, or under consideration is estimated to range from $223 billion to over $260 billion through 2030.[1] Companies are also more likely to investigate opportunities in US states that have not traditionally attracted the greatest semiconductor investment, with Intel building facilities in Ohio and Skywater planning to expand in Indiana.

New prerequisites and investment calculations

What's behind the shift that is prompting semiconductor companies to focus on the three S's, and what do they stand to gain? And what do these trends mean for regions and countries that want to attract more semiconductor investment? To answer these questions, we examined recent developments related to supply chains, sustainability, and subsidies.

A secure location that minimizes supply chain risks

With the recent pandemic, global economic uncertainty, and the war in Ukraine, executives' risk perceptions are rapidly shifting. According to a recent McKinsey survey of CEOs in advanced industries, many executives now view geopolitical dynamics as the most important challenge to their businesses.[2] In response, many leaders are now actively monitoring supply chain risks and developing strategies to prevent disruptions. One strategy that is now receiving much attention involves localizing semiconductor manufacturing to prevent disruptions and increase resilience.

Governments often welcome nearshoring efforts because they want to ensure a steady chip supply for local companies, including automakers and other businesses that depend on semiconductors. They also realize that access to chips is essential for many government security platforms.

Sustainability and decarbonization as a clear priority

Motivated by both voluntary and mandatory targets, all semiconductor companies are attempting to reduce their emissions. These efforts may help them meet emerging regulatory guidelines and satisfy the needs of their most important end customers.

[1] These figures were current as of January 2023.
[2] McKinsey Global Resilience Survey, July 2023, n = 331.

Many of these companies have set ambitious emissions reduction targets. Microsoft, for instance, wants to be carbon negative by 2030. To meet these goals, end customers must not only mitigate or eliminate their own emissions but they must also address Scope 3 upstream emissions, which include those that arise from the suppliers that provide them with products or components.[3] If semiconductor companies do not take proactive steps to reduce emissions, their customers might instead purchase carbon offsets—and they might attempt to pass the associated costs to suppliers in the form of price reductions or margin erosion.

Some semiconductor companies have already set emissions reduction goals, and their strategies often involve transitioning to renewable sources because greater than one-third of a typical fab's emissions arise from energy usage. Intel, which wants to reach net-zero greenhouse gas emissions by 2040, hopes to achieve 100 percent use of renewable electricity by 2030. The ability to shift to renewable energy will differ by location. While Singapore has a thriving semiconductor ecosystem, for instance, little land is available for building renewable-energy systems. Other locations face challenges because regulatory frameworks are still nascent.

If semiconductor companies in the United States or other countries with relatively high production costs can rely entirely on renewable sources, their energy costs could be two to four times lower than those in many Asian countries. This decrease might help offset some of their other expenses. Eventually, fabs in locations with little renewable energy might purchase it from other countries, which could raise overall energy costs above current rates. Singapore, for instance, wants to import 4 gigawatts of low-carbon electricity—equivalent to about 30 percent of its electricity supply—from neighboring countries by 2035.[4]

Subsidies

The European Union and the United States have increased the subsidies offered to semiconductor companies over the past few years. In another big shift, some countries without strong semiconductor ecosystems are actively trying to encourage the growth of such ecosystems within their borders. India and Spain are among the countries that have announced new programs over the past 12 to 18 months; these programs are designed to attract semiconductor investment and could have a major impact on site selection for new fabs.

Governments often welcome nearshoring efforts because they want to ensure a steady chip supply for local companies, including automakers and other businesses that depend on semiconductors.

[3] Scope 3 emissions also include those that arise from an end customer's product after purchase.
[4] "Regional power grids," Energy Market Authority, Government of Singapore, last updated December 18, 2023.

The subsidies offered today are often higher than they were in the past and may include new incentives. For instance, subsidies in the United States have historically involved state and local programs, such as property or sales tax abatements, but policy shifts have expanded the potential benefits. Consider some recent government-sponsored efforts to reshore high-tech manufacturing:

— *United States.* Currently, the United States only manufactures about 12 percent of the world's chips, and none are the most advanced varieties. The CHIPS and Science Act allocates over $50 billion for direct funding, federal loans, and loan guarantees designed to expand American semiconductor research and manufacturing. If successful, it would reduce dependence on foreign suppliers.

— *European Union.* Member countries have agreed to provide $47 billion in public funding aimed at doubling the European Union's share of global chip output to 20 percent by 2030.

— *Japan.* This country only has a 10 percent share of the global semiconductor market, down from about 50 percent in the 1980s. The Japanese government has announced $6.8 billion in public investment to expand domestic semiconductor production.

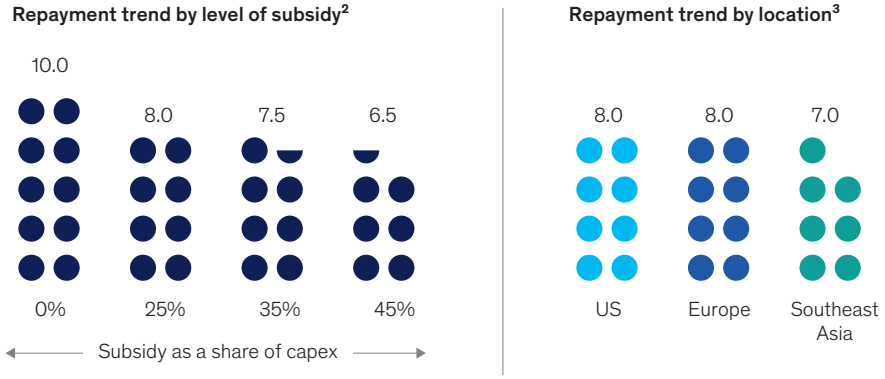
The increased subsidies have important implications because expansion is so costly. (The average cost of constructing and equipping a new fab is approaching $10 billion and could exceed that in some cases.) A recent McKinsey analysis reveals that subsidy levels have a greater impact than location on reducing the payback period for fab investments (Exhibit 3). For instance, a subsidy equal to 45 percent of the required investment will reduce the payback period to 6.5 years, compared with 10.0 years for unsubsidized facilities.

One downside to increased subsidies: they are raising the cost of entry for regions that want to attract investment from semiconductor companies, and the hurdles could become even higher if they continue to rise. Regions that have historically had

Exhibit 3

Recent analysis suggests that subsidies play the largest role in reducing the payback period for fab construction.

Payback period for a model fab,[1] years to repay



[1]Assuming a 28nm fab with a capacity of 500,000 wafers per annum; payback period calculated without discounting cash flows.
[2]Assuming a US location; subsidy as share of capex.
[3]Assuming 25% capex subsidy. Locations differentiated through regional labor and utility costs, as well as regional differences in capex (ie, building costs).

McKinsey & Company

A recent McKinsey analysis reveals that subsidy levels have a greater impact than location on reducing the payback period for fab investments.

a significant semiconductor presence could be at the greatest disadvantage.

The power of strong, local ecosystems

Semiconductor companies are not the only group that stands to gain from greenfield expansion, especially if it helps countries create ecosystems similar to Hsinchu Science Park. Such developments can offer broad rewards for regional economies, because semiconductors enable growth in a variety of realms. For example, they are essential to the growth of the Internet of Things, which is expected to have a global market worth of $4.4 trillion by 2030, and the robotics market, which is expected to be valued at $120.0 billion. If a region or country has many technology companies, a strong local supply of semiconductors could help them thrive, resulting in more jobs and a stronger regional economy.

Economic data clearly show the benefits of nurturing the semiconductor industry. It is the second-most-profitable industry in the world and thus has a major impact on GDP.[5] It also accounts for the second-highest amount of R&D spending, thus contributing to the creation of many highly skilled jobs.[6] Finally, the semiconductor industry also has strong economic multipliers, with investments estimated to increase its current value to GDP by threefold within six years. Employment multipliers are also strong, with every new job within the semiconductor industry expected to sustain over five new jobs in other industries.

Governments and companies could foster ecosystem development if they consider working together and establishing joint goals. But such "competitive cooperation," as some semiconductor companies' executives call it, is challenging in the best of times and even more so in these uncertain days, where agendas, goals, and long-term plans can be subject to change. Can these actors collaborate on developing an agenda and shared goals? Can appropriate governance controls be installed to resolve potential conflicts and remain in compliance with all local laws? Are sufficient supply chain protections and skilled labor available and willing to be deployed together? What mutual efforts are needed to build a capable workforce and ensure that the necessary infrastructure is available?

These are hard questions. Nevertheless, committed companies and countries that successfully address

[5] Based on R&D spending as a percentage of sales.
[6] Long-term implied economic profit based on July 21, 2022, market valuations from the EU Industrial R&D Investment Scoreboard and IC Insights.

If a region or country has many technology companies, a strong local supply of semiconductors could help them thrive, resulting in more jobs and a stronger regional economy.

them may foster the development of major semiconductor ecosystems that deliver returns that would far exceed those associated with a single new semiconductor company. In other words, the resulting juice may justify the squeeze of the additional effort.

Within the semiconductor industry, the macroeconomic sands are shifting. The demand outlook is exciting, geopolitical dynamics are changing, decarbonization is increasingly at the forefront, and countries are offering unprecedented incentives for greenfield investment. The emerging opportunities will help both incumbents and new companies that want to enter the fray. In all cases, subsidies, sustainability, and supply chain security will be among the most important considerations when selecting a location for greenfield building or expansion. If companies, governments, and other stakeholders successfully cooperate to create new semiconductor ecosystems, the advantages will extend far beyond individual countries or businesses. The entire industry—and the world as a whole—could benefit from the innovations that emerge.

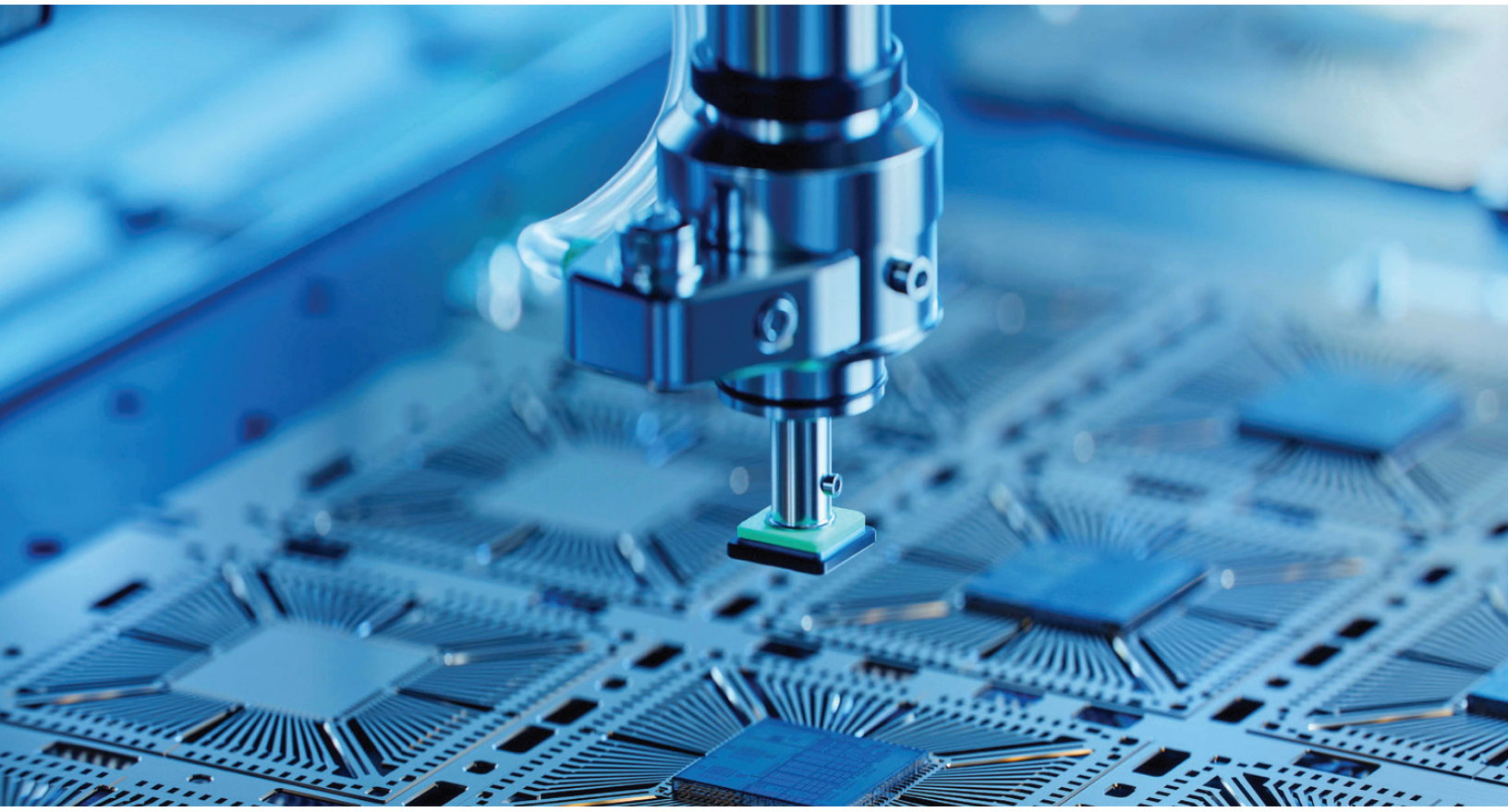
Ondrej Burkacky is a senior partner in McKinsey's Munich office, **Matteo Mancini** is a senior partner in the Riyadh office, **Mark Patel** is a senior partner in the Bay Area office, **Giulietta Poltronieri** is a partner in the Milan office, and **Taylor Roundtree** is an associate partner in the Atlanta office.

The authors would like to thank Sebastian Göke, Sean Planchard, and Ricardo Reina for their contributions to this article.

Advanced chip packaging: How manufacturers can play to win

As the benefits of Moore's law reach their limits, advances in chip performance rely more on the back end of production, including packaging.

by Ondrej Burkacky, Taeyoung Kim, and Inji Yeom



Semiconductor wafers are the basis of the integrated circuits so crucial to most of today's technology. The wafers' packaging—whether metal, plastic, ceramic, or glass—connects them to their environment and protects them from chemical contamination and damage from light, heat, and impacts. Compared with the front-end process of designing and fabricating wafers, the back-end process of packaging has been undervalued for two reasons: First, it's still possible to package wafers using old-generation equipment. Second, packaging is mostly done by outsourced semiconductor assembly and test companies (OSATs) that compete largely based on low labor costs, rather than other sources of differentiation.

This model may change with the introduction of advanced packaging, which uses sophisticated technology and aggregates components from various wafers, creating a single electronic device with superior performance. Introduced around 2000, advanced packaging is now gaining significant momentum as the next breakthrough in semiconductor technology.

Advanced packaging is helping to meet the demand for semiconductors that run emerging applications now going mainstream—for example, 5G, autonomous vehicles and other Internet of Things technologies, and virtual and augmented reality. These applications require high-performance, low-power chips that can rapidly process massive quantities of data. Despite Moore's law, which in 1965 posited that the number of transistors on a microchip would double every couple of years, node advancement is now reaching its limits. As a result, technical advances on the front end of chip manufacturing are slowing, and the economically viable maximum size of a die, and thus its performance, are becoming more limited. New approaches in back-end technology that combine multiple chips offer a promising solution. Advanced-packaging techniques that have arisen over the past two decades—including 2.5-D, 3-D, fan-out, and system-on-a-chip (SoC) packaging—promise to fill the void by supplementing the wire-bonding and flip-chip technologies of the previous half century.

Because advanced packaging offers a higher-value opportunity than traditional back-end packaging, major players and fast followers (organizations that imitate competitors' innovations) are developing and commercializing various forms of the technology to win premium customers. In this article, we describe how the market is evolving and suggest how manufacturers can take advantage of the opportunities becoming available.

Key advanced-packaging technologies

Three major advanced-packaging technologies have become commercially available since 2000, supplementing the two technologies that prevailed during the previous half century (Exhibit 1).

Traditional packaging techniques

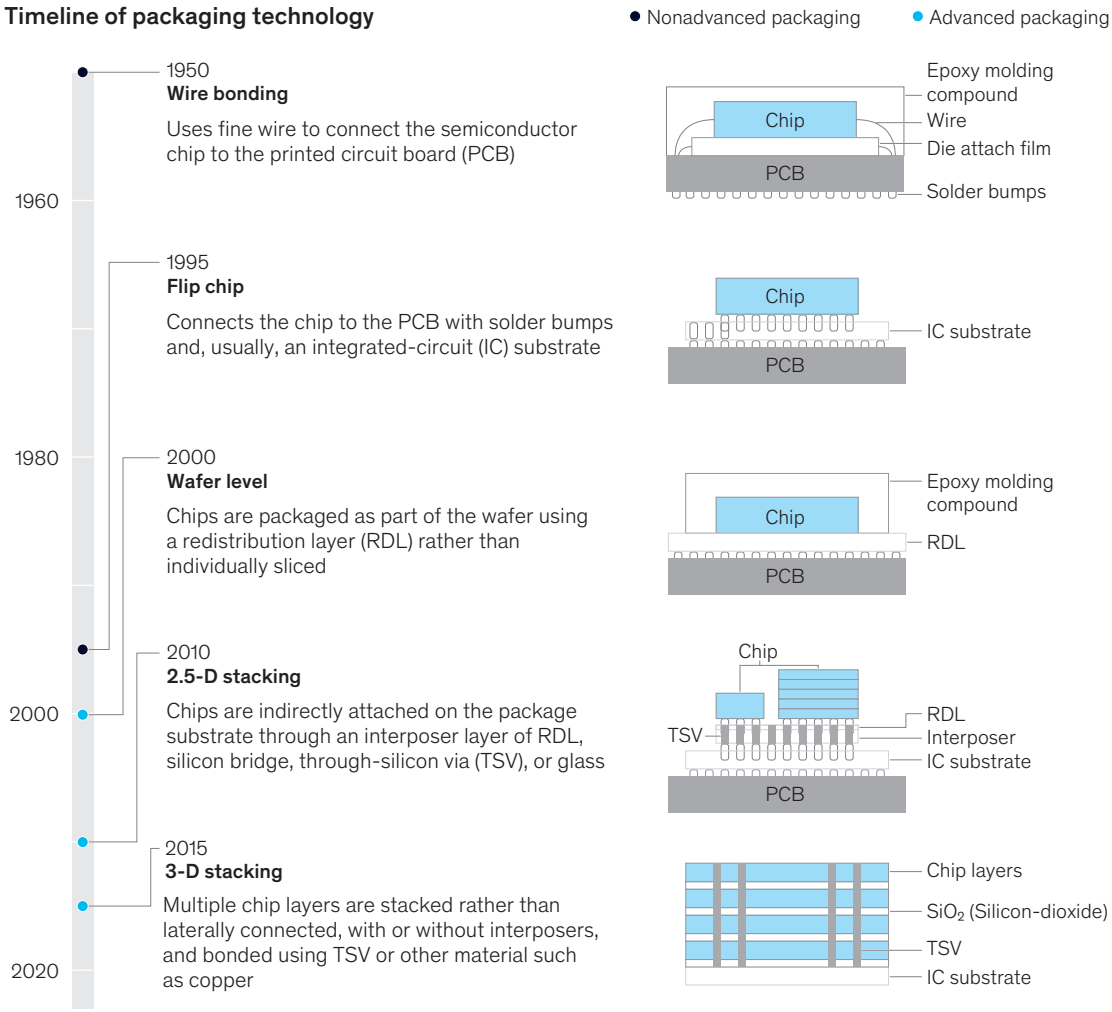
Developed in the 1950s and still in use today, wire-bond technology is an interconnection technique that attaches the printed circuit board (PCB) to the die—the silicon square that contains the integrated circuit—using solder balls and thin metal wires. It requires less space than packaged chips and can connect relatively distant points, but it can fail in high temperatures, high humidity, and temperature cycling, and each bond must be formed sequentially, which adds complexity and can slow manufacturing. The wire-bonding market is expected to be valued at about $16 billion by 2031, with a CAGR of 2.9 percent.[1]

The first major evolution in packaging technology came in the mid-1990s with flip chips, which use a face-down die, the entire surface area of which is used for interconnection through solder "bumps" that bond the PCB with the die. This results in a smaller form factor, or hardware size, and a higher signal-propagation rate—that is, faster movement of signals from the transmitter to the receiver. Flip-chip packaging is the most common and lowest-cost technology currently in use, mainly for

[1] *Wire bonding market forecast report*, 2021–2031, Transparency Market Research, November 2021.

Exhibit 1

Packaging technology for semiconductors has evolved quickly since 2000.



McKinsey & Company

central processing units, smartphones, and radio-frequency system-in-package solutions. Flip chips allow for smaller assembly and can handle higher temperatures, but they must be mounted on very flat surfaces and are not easy to replace. The current flip-chip market is around $27 billion, with a projected CAGR of 6.3 percent, which should bring it to $45 billion by 2030.[2]

Wafer-level packaging

While traditional packaging "dices" the silicon wafer into individual chips first and then attaches the chips to the PCB and builds the electrical connections, wafer-level packaging makes the electrical connections and molding at the wafer level, then dices the chips using a laser. The greatest difference between wafer-level chip-scale packaging (WLCSP) and flip

[2] "Flip chip market: Information by packaging technology (3D IC, 2.5D IC), bumping technology (copper pillar, solder bumping), and region—forecast till 2030," Straits Research, accessed April 2, 2023.

chips in terms of chip configuration is that WLCSPs have no substrate between the die and the PCB. Instead, redistribution layers (RDLs) replace the substrate, leading to a smaller package and enhanced thermal conduction.

Wafer-level packaging is divided into two types: fan-in and fan-out. In fan-in wafer-level packaging, used mainly for low-end mobile phones that require rudimentary technology, the RDLs are routed toward the center of the die. In the fan-out version, which was introduced in 2007, the RDL and solder balls exceed the size of the die, so the chip can have more inputs and outputs while maintaining a thin profile.[3] Fan-out packaging comes in three types: core, high density, and ultrahigh density. Core, which is used mostly for automotive and network applications that don't require high-end technology—such as radio frequency and infotainment chips—accounts for less than 20 percent of the almost $1.5 billion fan-out packaging market. High and ultrahigh density are mostly used for mobile applications and are expected to expand to some network and high-performance computing applications. The world's largest maker of WLCSPs is the Taiwan Semiconductor Manufacturing Company (TSMC).

The past decade saw the development of stacked WLCSP, which allows for multiple integrated circuits in the same package and is used for both heterogeneous bonding, which integrates logic and memory chips, and memory-chip stacking. In 2.5-D stacking, two or more chips are laid side by side with an interposer connecting one die to another. There are several categories of 2.5-D stacking, based on the kind of interposer it uses:

— Silicon interposers are the only type that requires TSV, or through-silicon via—a vertical electrical connection that passes through the silicon die or wafer. Silicon interposers use a stable technology that has been on the market for more than ten years, but the cost of silicon is high and requires front-end technology and manufacturing capability. TSMC's CoWoS-S (chip on wafer on substrate) dominates the market.

— Silicon bridges are relatively new. Because they use smaller amounts of silicon than traditional silicon interposers, they are thinner, which reduces power consumption and increases design flexibility. Their advantage over traditional silicon interposers is that they can enable more advanced system-level integration, so they are used for high-performance computing (HPC) such as AI. Representative technologies include Intel's EMIB (embedded multi-die interconnect bridge) and TSMC's CoWoS-L.

— Redistribution layers can also function as interposers. The greatest strength of this technology is that the photolithography process that creates RDLs allows for fine patterning, which improves speed gain and heat dissipation. TSMC's CoWoS-R (chip-on-wafer-on-substrate RDL) is set to begin mass-volume production.

— Glass is also rising as a next-generation material for interposers. It offers low cost and low power loss in high-frequency bandwidths, but it may not be marketable for some time.

In 3-D stacking, multiple chips are placed face down on top of one another, with or without an interposer. There are two main types of 3-D stacking. The most common type is TSV with micro-bumps (μ-bumps). The newer alternative, bumpless hybrid bonding, forms interconnections using a dielectric bond and embedded metal; it is just being explored by memory players.

[3] Karen Heyman and Laura Peters, "Fan-out packaging gets competitive," *Semiconductor Engineering*, August 18, 2022.

How will the market evolve?

The advanced-packaging market is driven by the end applications of its various technologies (Exhibit 2). Since the mid-2010s, fan-out wafer-level packaging has dominated, with about 60 percent market share. Fan-out packaging is cheaper than stacking and is engineered for high heat resistance and a small form factor. These attributes make it appropriate for mobile applications, which are likely to generate most of its demand.

Apple uses fan-out advanced packaging for its application processors, graphic chips, and 5G and 6G modem chips. It is the largest user of the technology, consuming most of the volume produced by TSMC. Other top fabless players—that is, companies that design and sell hardware and chips but outsource their manufacture—are also using fan-out technology in mass-produced chips.

Most of the growth in HPC and network applications is likely to come from AI chips, edge computing, and network chips in consumer devices, which require the small form factor and affordable cost that fan-out packaging can offer.

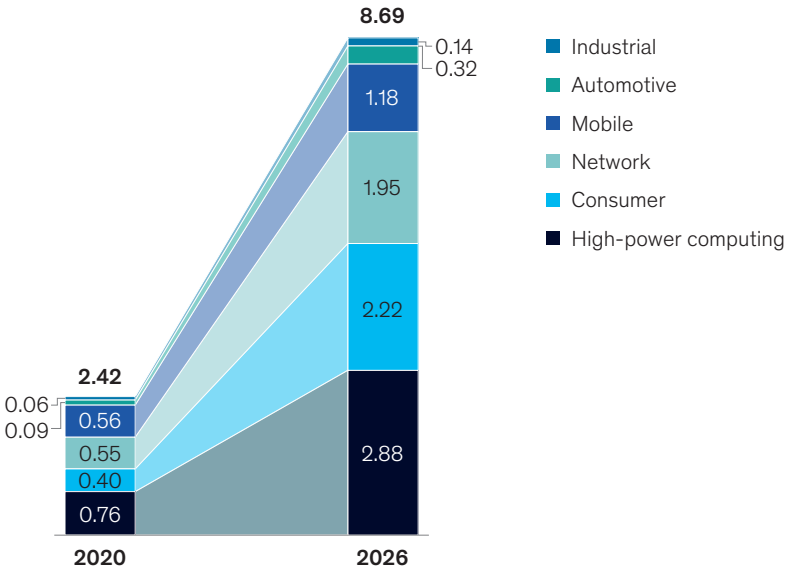
The most likely driver of growth in 2.5-D stacking could be HPC applications, which are in high demand for data centers. Although less than 20 percent of data-center capacity used 2.5-D stacking in 2022, the rate could increase to 50 percent in the next five years. For mobile applications, 2.5-D packaging is considered too costly, but this may change with the arrival of the next generation, which will feature less expensive silicon bridges, RDLs, and glass interposers.

For 3-D packaging, memory—the dominant application for 3-D stacking—and SoC use are

Exhibit 2

The advanced-packaging market is spurred by end applications.

Advanced-packaging sales, by end application, $ billion



Source: Yole report; McKinsey analysis

McKinsey & Company

expected to grow at a CAGR of roughly 30 percent. Increasingly, 3-D stacked memory is being incorporated with logic chips for high-performance products that require high bandwidth, including high-bandwidth memory (HBM) and processing in memory with HBM (PIM-HBM). Substantial demand for 3-D stacked memory will likely come from data-center servers, which require high capacity and high speed, and graphics accelerators and network devices, which require the maximum possible bandwidth for memory and processing.

HPC systems, specifically CPUs, will drive demand for 3-D SoC chips. Major players started adopting hybrid bonding in 2022, and fast followers may join the market soon. OSATs, lower-tier foundries, and integrated device manufacturers (IDMs) are unlikely to enter the market, given the high technology barrier.

Key market-winning capabilities

Market growth relies heavily on end customers, such as automotive OEMs and home appliance manufacturers. More end customers are seeking advanced-packaging providers because of the growing need for fast, reliable computing for applications such as autonomous vehicles. For semiconductor manufacturers—especially logic IDMs and foundries—advanced packaging could be a key selling point.

To acquire and retain high-value fabless customers, manufacturers need to be comfortable codeveloping advanced-packaging solutions. While fabless players take full ownership of the chip-planning process before at-scale production begins, there is room for manufacturers to add value. Joint development often occurs during the chip-architecture design stage and initial shuttle runs for design validation (Exhibit 3). The need for such cooperation is expected to increase because of the demand for higher-performance chips and the increased complexity of chip designs created by packaging.

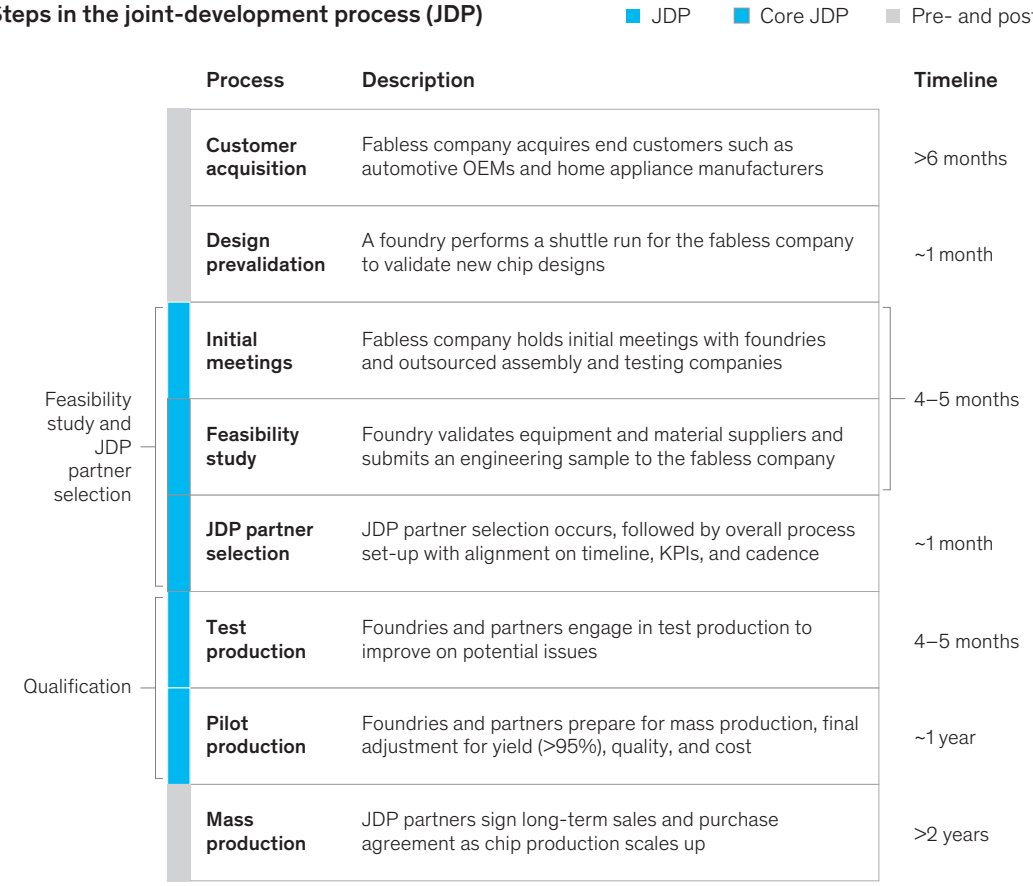
In 2016, TSMC released innovative integrated fan-out (InFO) wafer-level systems, mainly for wireless applications, in close collaboration with its lead customer. More recently, derivatives of that, such as InFO AiP (antenna in package) and InFO PoP (package on package), have been released to expand into other applications for networking and HPC.

More end customers are seeking advanced-packaging providers because of the growing need for fast, reliable computing for applications such as autonomous vehicles.

Exhibit 3

The joint-development process in advanced packaging can attract high-value fabless customers.

Steps in the joint-development process (JDP)



McKinsey & Company

Fast followers may have a hard time catching up with market leaders, because huge technology investments would be required to assure customers of the volume to support products. In addition, although fast followers may have R&D-level packaging technology for fan-out and 2.5-D, they have little or no production experience, which is essential for high production yield. To overcome this, packaging players would need to acquire anchor customers from the initial stages of development. Positioning their companies as willing to help manufacture products for advanced packaging from the design stage would be key to acquiring customers.

Advanced packaging requires changes in the architecture of end-user software and hardware, so packaging design should be considered during the initial architecture stage, when support from back-end providers can lower the burden of adopting advanced packaging. Once a customer selects an advanced-packaging vendor, it will likely commit to that vendor for future projects as well.

To acquire design capabilities, companies can partner with or invest in a design house. Design houses play a critical role across the entire chip-making process, from intellectual-property (IP) development to design and production. Additionally,

owning an IP pool can help customers meet their design needs quickly and allow them to avoid redundant designs and resources. Design houses should be able to offer front- and back-end services. Front-end services include register-transfer-level design and high-level description of the functions required; back-end design includes logic testing and place and route.

Another potentially important value proposition for the chip manufacturer is securing design capabilities and providing turnkey solutions—from design to wafer manufacturing, packaging, and testing. This type of offering provides customers with a one-stop shop.

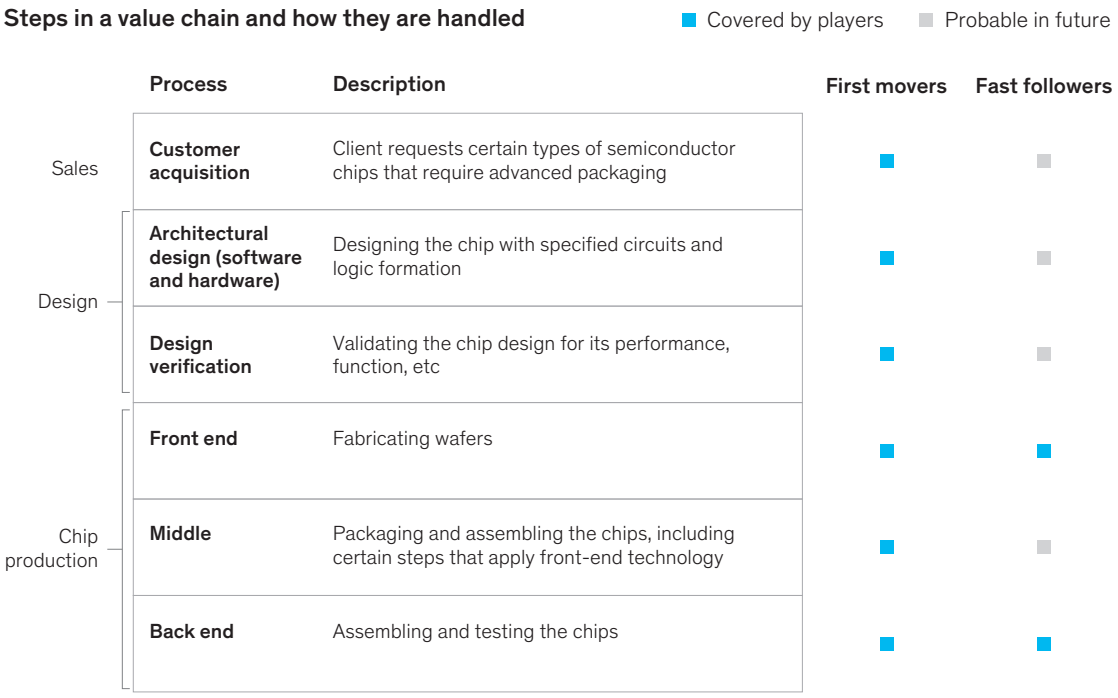
In terms of manufacturing, the two key technological capabilities manufacturers need to master for 2.5-D and 3-D packaging are, respectively, interposers and hybrid bonding. For 2.5-D, manufacturers must be able to handle emerging interposer solutions using novel materials and manufacturing methodologies, including silicon, RDL, and glass. For 3-D, the latest technology, hybrid bonding, requires chemical mechanical planarization to polish various substances with equal flatness and prevent dishing, as well as high interconnect accuracy through disk-to-wafer capabilities in both equipment and know-how.

Implications for manufacturers

Key players in advanced packaging include logic and memory IDMs, foundries with leading or mature node capabilities, and OSATs. Exhibit 4 shows the capabilities currently handled by first movers and fast followers.

Exhibit 4

Different areas of the advanced-packaging value chain are handled by different players.



McKinsey & Company

First movers
First movers have entered the market and are in mass-volume production based on their logic-packaging capabilities. They are actively developing use cases with existing customers and applying cutting-edge advanced-packaging technologies. While these major players are advanced in R&D and manufacturing, they may seek partnerships with followers to stabilize volume as they face rapidly expanding demand.

Fast followers
Many fast followers are striving to take a share of the advanced-packaging market but have not mastered the design or manufacturing capabilities or built a sufficient customer base, especially for high-end solutions.

Foundries that have mature node capability but lack advanced packaging could benefit substantially from finding synergies within their current product portfolios. While advanced logic chips with nodes smaller than ten nanometers have the greatest need for advanced packaging, it is critical for fast followers to find opportunities to capture the mature-node market. Some of the areas where advanced packaging can be adapted to enhance the performance of mature-node legacy chips are radio-frequency transceiver chips for network applications, advanced driver-assist systems (ADAS), and infotainment chips for automotive applications.

Another option is to partner with logic providers to develop design and manufacturing solutions for specific applications that use both mature and leading-edge nodes. The feasibility of this tactic would largely depend on the end-application demand and logic providers' needs.

OSATs
OSATs' capabilities in the high-end advanced-packaging market are limited. Rather than trying to compete directly with high-end solutions, they can offer comparatively low-end solutions or seek to collaborate in certain value-chain areas with players capable of high-end advanced packaging. Leading OSATs are actively investing to expand the range of advanced packaging they offer. Some can already handle core and HD-level fan-out packaging, but 2.5-D and 3-D stacking mainly remain in R&D.

First movers are in mass-volume production based on their logic-packaging capabilities. While these players are advanced in R&D and manufacturing, they may seek partnerships with followers to stabilize volume as they face rapidly expanding demand.

Another option for OSATs is to partner with players capable of 2.5-D and 3-D stacking. While these partners work on core processes—including through-silicon via, RDL lithography, and hybrid bonding—the OSATs could offer solutions for the mid- to back-end processes, including wafer thinning and bumping.

Although foundries and IDMs are developing advanced-packaging capabilities, they will likely use advanced packaging only to attract high-end customers that require state-of-the-art technology and, therefore, will not disrupt the entire OSAT business. They are not expected to expand into core and fan-out advanced packaging, given the significant differences in operating margin compared with front-end manufacturing, though they may make the leap into more profitable advanced 2.5-D or 3-D packaging.

Memory IDMs

Logic capability is essential for advanced packaging, but 3-D stacking technology can still present opportunities for memory IDMs, as top players are using it to enhance performance in memory chips that include basic-level logic chips. IDM players can also differentiate themselves by using the technology to customize memory for key clients' advanced-packaging chips.

Another scenario for memory IDMs is to develop logic capabilities, particularly in design or manufacturing, to enable synergies with advanced packaging. This would, however, require substantial investment and a risky leap across the value chain.

The advent of advanced packaging has changed the competitive landscape for chip manufacturers. Packaging is no longer a commodity process, and the majors have moved first to make advanced packaging a strategic part of their offerings. Other manufacturers risk being commoditized if they don't find a way to incorporate advanced packaging into their strategies and offerings. The advanced-packaging market offers many disruptive opportunities, as well as challenges that will likely go beyond business as usual.

Ondrej Burkacky is a senior partner in McKinsey's Munich office, and **Taeyoung Kim** is a consultant in the Seoul office, where **Inji Yeom** is an associate partner.

The authors wish to thank Hawon Baeg, Harald H. Bauer, Steve Park, Rutger Vrijen, and Bill Wiseman for their contributions to this article.

McKinsey & Company

GSA Where Leaders Meet

The future of automotive computing: Cloud and edge

The rise of 5G and edge computing will create new opportunities along the automotive supply chain. How can semiconductor companies and other stakeholders capture it?

This article is a collaborative effort by Philip Arejola, Ondrej Burkacky, Johannes Deichmann, Gourav Ganguly, Asif Khan, and Martin Wrulich, representing views from McKinsey's Advanced Industries and Technology, Media, and Telecommunications Practices.

Executive summary

The ACES trends—autonomous driving, connectivity, electrification, and shared mobility—are transforming the automotive industry. Even greater changes may be in store because 5G technology is expected to provide the bandwidth, low latency, reliability, and distributed capabilities that better address the needs of connected cars. These benefits could contribute to greater use of edge applications within the automotive sector and lead to the development of new automotive use cases. While most current automotive applications now rely solely on one workload location, they may later use some combination of edge computing with onboard or cloud processing that delivers higher performance.

These developments will have major implications for companies along the entire automotive value chain. We estimate the total value created by connected-car use cases to increase from about $65 billion in 2020 to $450 billion to $500 billion by 2030. Over that same period, the percent of value enabled by 5G and edge will increase from about 5 to 30 percent. While short-term value unlock would largely be from the enhancement of existing cases, in the longer term, value creation will be driven by enabling new and advanced use cases.

The shared realization that no player can go it alone could lead to the emergence of open and closed ecosystems (as well as hybrid models), and companies will find new opportunities in both hardware and software. This trend could encourage OEMs and suppliers to define technology standards anchored on technology stack control points. One particular opportunity is in the orchestration layer for end-to-end workload balancing that supports northbound and southbound interfaces. Developing this standard would require the players across the value chain to pool their domain expertise behind a common goal of defining an end-to-end capability.

All companies along the value chain—semiconductor players, tier 1 suppliers, OEMs, communication system suppliers, and hyperscalers—could increase value capture in the evolving automotive landscape, but they would first benefit from reviewing all aspects of their strategy, including products, capabilities, organizational and operational structures, and go-to-market models. For many players, this may be a good time to become comprehensive solution providers that deliver far more than hardware and software.

Introduction

As the connected-car ecosystem evolves, it will affect multiple value chains, including those for automotive, telecommunications, software, and semiconductors. In this report, we explore some of the most important changes transforming the sector, especially the opportunities that may arise from the growth of 5G and edge computing. We also examine the value that semiconductor companies might capture in the years ahead if they are willing to take a new look at their products, capabilities, organizational and operational capabilities, and their go-to-market approaches.

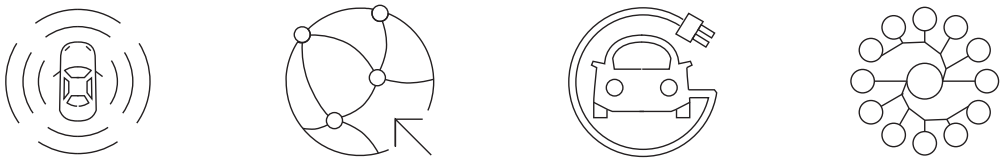
A new age of vehicle software and electronics

Four well-known technology trends have emerged as key drivers of innovation in the automotive industry: autonomous driving, connectivity, electrification, and shared mobility—such as car-sharing services (Exhibit 1). Collectively, these are referred to as the ACES trends, and they will have a significant impact on computing and mobile-network requirements. Autonomous driving may have the greatest effect, since it necessitates higher onboard-computing power to analyze massive amounts of sensor data in real time. Other

Exhibit 1

ACES trends are shaping the future of the automotive industry and are enabled by advancements in semiconductors and software.

4 Automation, connectivity, electrification, and shared mobility (ACES) trends



Autonomous driving
OEMs and suppliers are fueling the trend toward autonomous driving with different concepts and technological enablers

Connectivity
Intelligent communication within and outside of the car is a key enabler for autonomous technologies

Electrification
Electric vehicles are becoming more standard and continuing to increase in range

Shared mobility
Ownership model of cars is evolving to "renting and sharing" as customer preferences shift

McKinsey & Company

autonomous technologies, over-the-air (OTA) updates, and integration of third-party services will also require high-performance and intelligent connectivity within and outside of the car. Similarly, increasingly stringent vehicle safety requirements require faster, more reliable mobile networks with very low latencies.

With ACES functions, industry players now have three main choices for workload location: onboard the vehicle, cloud, and edge (Exhibit 2).

To ensure that use cases meet the thresholds for technical feasibility, companies must decide where and how to balance workloads across the available computing resources (Exhibit 3). This could allow use cases to meet increasingly strict safety requirements and deliver a better user experience. Multiple factors may need to be considered for balancing workloads across onboard, edge, and cloud computing, but four may be particularly important. The first is safety, since workloads essential for passenger safety require extremely

Advances in computing and connectivity are expected to enable many new and advanced automotive use cases.

Exhibit 2

Primary processing of connected-car workloads may occur onboard the vehicle, in edge computing, or in the cloud.

Computing location examples

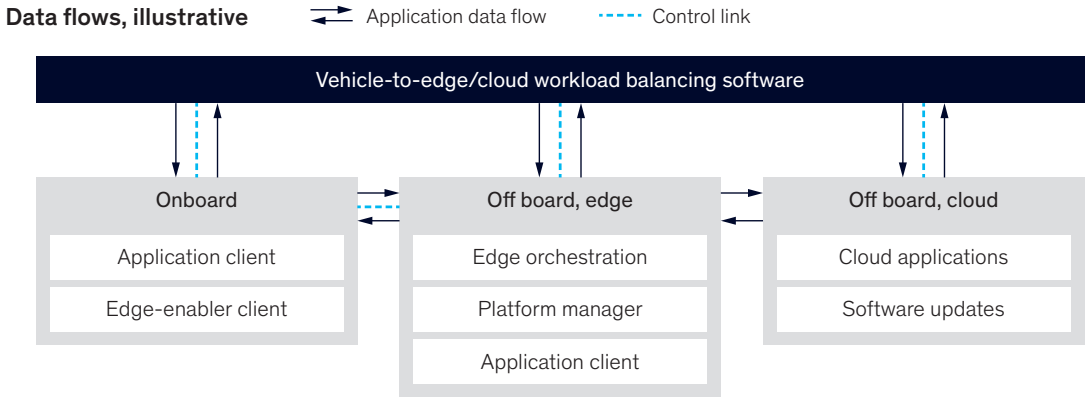
Onboard
- autonomous emergency-braking system
- forward-collision warning
- onboard critical computing for cooling management, seat belt/airbag activation, etc

Off board, edge
- smart traffic management systems
- intersection collision warning
- situational-awareness system

Off board, cloud
- over-the-air updates
- infotainment
- in-car office

Exhibit 3

Balancing connected-car computing workloads is a critical piece of the puzzle, and software can help determine the best workload location.



McKinsey & Company

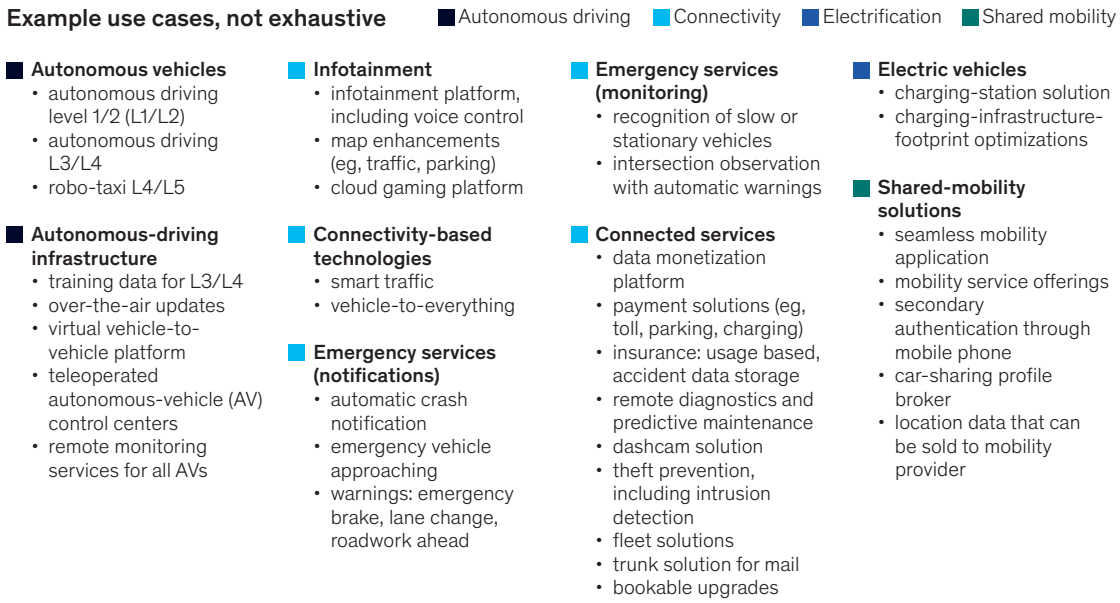
fast reaction times. Other considerations include latency, computing complexity, and requirements for data transfer, which depend on the type, volume, and heterogeneity of data.

Connected-car use cases today typically rely on either onboard computing or the cloud to process their workloads. For example, navigation systems can tolerate relatively high latency and may function better in the cloud. OTA updates are typically delivered via a cloud data center and downloaded via Wi-Fi when it is least disruptive, and infotainment content originates in the cloud and is buffered onboard to give users a better experience. By contrast, accident prevention workloads such as autonomous emergency-braking systems (AEBS) require very low latency and high levels of computing capability, which, today, may mean that they are best processed onboard the vehicle.

Advances in computing and connectivity are expected to enable many new and advanced use cases (Exhibit 4). These developments could alter where workloads are located. Of particular significance, the rollout of 5G mobile networks could allow more edge processing. Given the importance of these interrelated technologies, we

Exhibit 4

Advances in connectivity and computing may unlock many new use cases for the automotive sector.



Source: McKinsey Center for Future Mobility

McKinsey & Company

explored their characteristics in detail, focusing on automotive applications.

The benefits of 5G and edge computing

5G technology is expected to provide the bandwidth, low latency, reliability, and distributed capabilities that better address the needs of connected-car use cases. Its benefits to automotive applications fall into three main buckets:

— *Enhanced mobile broadband (EMBB):* 5G may provide faster, more uniform user experiences with speeds reaching ten gigabits per second, five to ten times faster than 4G technology. This may enhance high-bandwidth use cases such as in-car infotainment, vehicle teleoperation, and real-time human-machine-interface rendering.

— *Massive Internet of Things (IoT):* By enabling up to a million connections per square kilometer, 5G networks could efficiently support a large number of concurrent connections from cars on the road, connected infrastructure end points, and end-user devices. This may eliminate the possibility that cars and other devices get disconnected from the mobile network inadvertently because of a large number of connections.

— *Ultra-low-latency communications (URLLC):* 5G latency can theoretically go down to one millisecond—five to 15 times better than 4G. This means 5G can combine high speed with high reliability, eliminating the need for trade-offs between the two. This is important for object tracking in autonomous vehicles, the protection and control of smart-grid critical infrastructure, and remote-control and process automation for applications including aviation and robotics.

These benefits could contribute to greater use of edge applications within the automotive sector.

Workloads that are not safety-critical—infotainment and smart traffic management, for example—could start to shift to the edge from onboard or in the cloud. Eventually, 5G connectivity could reduce latency to the point that certain safety-critical functions could begin to be augmented by the edge infrastructure, rather than relying solely on onboard systems.

Most current automotive applications today tend to rely exclusively on one workload location. In the future, they may use some combination of edge computing with onboard or cloud processing that delivers higher performance. For instance, smart traffic management systems may improve onboard decision making by augmenting the vehicle's sensor data with external data (for example, other vehicles' telemetry data, real-time traffic monitoring, maps, and camera images). Data could be stored in multiple locations and then fused by the traffic management software. The final safety-related decision will be made onboard the vehicle. Ultimately, large amounts of real-time and non-real-time data may need to be managed across vehicles, the edge infrastructure, and the cloud to enable advanced use cases. In consequence, data exchanges between the edge and the cloud must be seamless.

Shifting industry dynamics and new opportunities

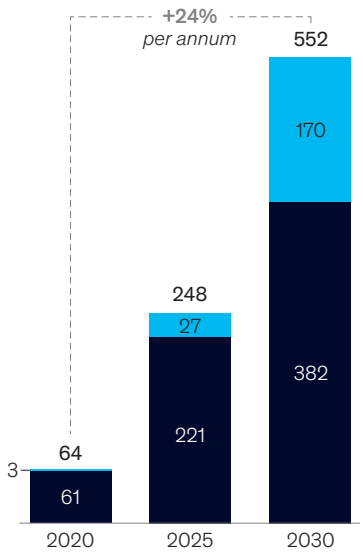
The evolving automotive value chain will open many new opportunities for those within the industry and external technology players. The total value created by connected-car use cases could reach more than $550 billion by 2030, up from about $64 billion in 2020 (Exhibit 5).

Increased connectivity opens up opportunities for players across the automotive value chain to improve their operations and customer services. Take predictive maintenance in cars as an example. Aftermarket maintenance and repair provision now predominantly involve following a

Exhibit 5

The evolving automotive value chain offers a significant upside for all stakeholders.

Value creation from 5G and edge, accelerated scenario, $ billion



	CAGR, %	
	2020–25	2025–30
Enabled	59	44
Enhanced	29	12

McKinsey & Company

fixed interval maintenance schedule or reactive maintenance/repair. There is little visibility around the volume of vehicles that need to be serviced in a particular period, leading to inefficiencies in service scheduling, replacement parts ordering, and inventory, among others. Predictive maintenance using remote car diagnostics could improve the process by giving OEMs and dealers an opportunity to initiate and manage the maintenance process.

The pace of rollout of advanced connected-car use cases is highly contingent on the availability of 5G and edge computing. A variety of factors are converging to accelerate this. Demand is rising for these critical enablers, fueled by a proliferation of consumer and industry use cases. In the short term, value may be generated through enhancements to services already available with 4G, including navigation and routing, smart parking, centralized and adaptive traffic control, and monitoring of drivers, passengers, or packages.

We expect that greater 5G and edge availability may expand the list of viable use cases (technically and financially), boosting edge value exponentially. Looking to 2030, about 30 percent of our value estimate may be enabled by 5G and edge (from 5 percent in 2020), largely consistent with our cross-sectoral report on advanced connectivity.

Value creation could be accelerated by traditional players moving into adjacencies and by new entrants from industries not traditionally in the automotive value chain, such as communication system providers (CSPs), hyperscalers, and software developers. Players such as Intel, Nvidia, and the Taiwan Semiconductor Manufacturing Company are adding automotive-software capabilities, leading to greater synergies and vertical-integration benefits. In addition to accelerating value creation, new entrants may compete for a greater share of the total value.

Automotive-hardware value chains are expected to diverge based on the type of OEM. Traditional auto manufacturers, along with their value chains, are expected to see a continuation of well-established hardware development roles based on existing capabilities. Automobiles, components, devices, and chips for applications ranging from cars to the cloud may continue to be primarily manufactured by the companies that specialize in them. Nontraditional or up-and-coming automotive players could codevelop vehicle platforms with the established car OEMs and use OEMs' services or contract manufacturers such as Magna Steyr for the traditional portions of the value chain.

Established players may seek to increase their share by expanding their core businesses, moving up the technology stack, or by growing their value chain footprints. For instance, it is within the core business of semiconductor players to create advanced chipsets for automotive OEMs, but they could also capture additional value by providing onboard and edge software systems or by offering software-centric solutions to automotive OEMs. Similarly, to capture additional value, hyperscalers could create end-user services, such as infotainment apps for automotive OEMs or software platforms for contract manufacturers.

Emerging ecosystem archetypes

As players make strategic moves to improve their position in the market, we can expect two types of player ecosystems to form. In a closed ecosystem, membership is restricted and proprietary standards may be defined by a single player, as is the case with Volkswagen, or by a group of OEMs. Open ecosystems, which any company can join, generally espouse a democratized set of global standards and an evolution toward a common technology stack. In extreme examples—where common interfaces and a truly open standard exist—each player may stay in its lane and focus on its core competencies.

Hybrid ecosystems will also exist. Players following this model are expected to use a mix of open and closed elements on a system-by-system basis. For example, this might be applied to systems in which OEMs and suppliers of a value chain have particular expertise or core competency.

Exhibit 6 describes the advantages and disadvantages of each ecosystem model.

Value chain dynamics

Companies in the emerging connected-car value chain develop offerings for five domains: roads and physical infrastructure, vehicles, network, edge, and cloud. For each domain, companies can provide software services, software platforms, or hardware (Exhibit 7).

As automotive connectivity advances, we expect a decoupling of hardware and software. This means that hardware and software can develop independently, and each has its own timeline and life cycle. This trend may encourage OEMs and suppliers to define technology standards jointly and could hasten innovation cycles and time to market. Large multinational semiconductor companies have shown that development time can be reduced by up to 40 percent through decoupling and parallelization of hardware and software development. Furthermore, the target architecture that supports this decoupling features a strong middleware layer, providing another opportunity for value creation in the semiconductor sector. This middleware layer may likely be composed of at least two interlinked domain operating systems that may handle the decoupling for their respective domains. Decoupling hardware and software, which is a key aspect of innovation in automotive, tilts the ability to differentiate offerings heavily in favor of software.

New opportunities. In the software layer, companies could obtain value in several different ways. With open ecosystems, participants will have broadly adopted interoperability standards with relatively common interfaces. In such cases, companies may remain within their traditional domains. For instance, semiconductor players may focus on producing chipsets for specific customers across the domains and stack layers, OEMs concentrate on car systems, and CSPs specialize in the connectivity layer and perhaps edge infrastructure. Similarly, hyperscalers may capture value in cloud/edge services.

Exhibit 6

Two types of connected-car ecosystems are emerging, as well as a hybrid model.

Ecosystem models

Closed
Selected participants follow strategic group-defined standards
- Participants define a standard for a new technology layer or control point
- Higher control over end-to-end user experience
- Affects speed of innovation
- Risk of being restricted to certain technologies or processes

Hybrid
Companies select their chosen ecosystem on a system-to-system basis according to their strategic planning horizon
- Companies assert a degree of control over quality and interface
- Flexible sourcing could mitigate risks, except for IP critical systems
- Need to integrate multiple systems has implications on time-to-market

Open
Participants follow a democratically chosen set of global standards
- Standards help companies to reach critical scale faster
- Innovation cycles benefit from "openness"
- Transparent value chain selection based on core competency
- Time-to-market contingent on pace of standardization
- Standardized interfaces could lead to commoditization

McKinsey & Company

Exhibit 7

Companies in the emerging connected-car value chain can develop offerings in five domains.

Value opportunities, not exhaustive

	Roads and infrastructure	Vehicle	Network	Edge	Cloud
Software services	• security • orchestration	• security • orchestration • human–machine interface/driver interface • application layer	• security • orchestration	• security • orchestration • applications, analytics, visualizations	• security • orchestration • applications, analytics, visualizations
Software platform	• security • road asset-management layer • open data platform • device management • data ingest and aggregation layer	• platform • middleware • base vehicle operating system (OS)	• network orchestration • connectivity management, provisioning, and life cycle management	• multiaccess edge platform • multiaccess edge-computing management	• cloud platform services • virtual infrastructure, virtual machines, and OS
Hardware	• security • roadway infrastructure • roads, traffic, infrastructure, signage, and CCTV[1] cameras • monitoring sensors • refueling, recharging infrastructure	• electrical and electronic hardware • vehicle hardware	• 5G management and orchestration • 5G core • 5G distributed and central units • active equipment • passive equipment	• physical infrastructure	• physical infrastructure

[1]Closed-circuit television.

McKinsey & Company

In closed ecosystems, by contrast, companies may define proprietary standards and interfaces to ensure high levels of interoperability with the technologies of their members. For example, OEMs in a closed ecosystem may develop analytics, visualization capabilities, and edge or cloud applications exclusively for their own use, in addition to creating software services and platforms for vehicles. Sources of differentiation for vehicles could include infotainment features with plug-and-play capabilities, autonomous capabilities such as sensor fusion algorithms, and safety features.

While software is a key enabler for innovation, it introduces vulnerabilities that can have costly implications for OEMs, making cybersecurity a priority (see sidebar, "The importance of cybersecurity," for more information). Combined, the 5G and edge infrastructure could potentially offer increased flexibility to manage security events related to prevention and response.

Hardware players could leverage their expertise to offer advanced software platforms and services. Nvidia, for instance, has entered the market for advanced driver-assistance systems (ADAS) and is complementing its system-on-a-chip AI design capabilities with a vast range of software offerings that cover the whole automated-driving stack—from OS and middleware to perception—and trajectory planning.

Some companies are also moving into different stack layers. Take Huawei, which has traditionally

The importance of cybersecurity

Open ecosystems involve many players, necessitating additional interfaces, which provides more potential entry points for attacks. Closed systems are less exposed to third-party hardware and software because they are closely controlled by the OEM at each stage of the value chain. But recent attacks on some vehicles raise questions about whether closed systems are truly less vulnerable.

Automotive cybersecurity could be reinforced with the adoption of new working practices in four main areas:

- *Managing vehicle cyber risks* by establishing governance structures and clear responsibilities for cybersecurity of vehicles and related domains.
- *Securing vehicles by design* by introducing secure engineering practices into research and development and integrating cybersecurity into supplier audits.
- *Reacting promptly to security incidents* by establishing efficient detection and secure response capabilities.
- *Providing safe and secure software updates* by creating processes for securely updating vehicle software without affecting safety.

Cybersecurity concerns affect the whole automotive value chain, and suppliers can help develop solutions. For instance, they could provide OEMs with cybersecurity-related artifacts, which tracks interactions with the system and other digital evidence.

been a network equipment provider and producer of consumer-grade electrical and electronic (E&E) equipment, and manufacturer of infrastructure for the edge and cloud. Currently, the company is targeting various vehicle stack layers, including the base vehicle operating systems, E&E hardware, automotive-specific E&E, and software and EV platforms. In the future, Huawei may develop vehicles, monitoring sensors, human–machine interfaces, application layers, and software services and platforms for the edge and cloud domains.

New opportunities and strategies along the automotive value chain

Greater automotive connectivity will present semiconductor players and other companies along the automotive value chain with numerous opportunities. In all segments, they may benefit from becoming solution providers, rather than keeping a narrower focus on software, hardware, or other components. As they move ahead and attempt to capture value, companies may benefit from reexamining elements of their core strategy, including their capabilities and product portfolio.

Semiconductor companies

The automotive semiconductor market is one of the most promising subsegments of the global semiconductor industry, along with the Internet of Things and data centers. Semiconductor companies that transform themselves from hardware players to solution providers may find it easier to differentiate their business from the competition's. For instance, they might win customers by developing application software optimized for their system architecture. Semiconductor companies could also find emerging opportunities in the orchestration layer, which may allow them to balance workloads between onboard, cloud, and edge computing.

Semiconductor companies can capitalize on their edge and cloud capabilities by building strategic partnerships with hyperscalers and edge players that have a strong focus on automotive use cases.

As semiconductor companies review their current product offerings, they may find that they can expand their software presence and produce more purpose-specific chips—such as microcontrollers for advanced driver-assistance, smart cockpit, and power-control systems—at scale by leveraging their experience in the automotive industry and in edge and cloud computing. Beyond software, semiconductor companies might find multiple opportunities, including those related to more advanced nodes with higher computing power and chipsets with higher efficiency.

To improve their capabilities related to purpose-specific chips, semiconductor players would benefit from a better understanding of the needs of OEMs and consumers, as well the new requirements for specialized silicon. Semiconductor companies can capitalize on their edge and cloud capabilities by building strategic partnerships with hyperscalers and edge players that have a strong focus on automotive use cases.

Tier 1 suppliers

Tier 1 suppliers could consider concentrating on capabilities that may allow them to become "tier 0.5" system integrators with higher stack control points. In another big shift, they could leverage existing capabilities and assets to develop operating systems, ADAS, autonomous driving, and human-machine-interface software for new cars.

To produce the emerging offerings in the automotive-computing ecosystem, tier 1 players might consider recruiting full-stack employees who see the bigger picture and can design products better tuned to end-user expectations. They might also want to think about focusing on low-cost countries and high-volume growth markets with price-differentiated, customized, or lower-specification offerings that have already been tested in high-cost economies.

OEMs

OEMs could take advantage of 5G and edge disruption by orienting business and partnership models toward as-a-service solutions. They could also leverage their existing assets and capabilities to build closed- or open-ecosystem applications, or focus on high-quality contract manufacturing. Key OEM high growth offerings could include as-a-service models pertaining to mobility, shared mobility, and batteries. OEMs, when seeking partnerships with other new and existing value chain players, need to keep two major things in mind: filling talent and capability gaps (for instance, in chip development) and effectively managing diverse portfolios.

CSPs

CSPs must keep network investments in lockstep with developments in the automotive value chain to ensure sufficient 5G/edge service availability. To this end, they may need to form partnerships with automotive OEMs or hyperscalers that are entering the space. For best results, CSPs will ensure that their core connectivity assets can meet vehicle-to-everything (V2X) use case requirements and create a road map to support highly autonomous driving.

Connectivity alone represents a small part of the overall value to CSPs, however, and companies will benefit from expanding their product portfolios to include edge-based infrastructure-as-a-service and platform-as-a-service. Evolving beyond the traditional connectivity core may necessitate organizational structures and operating models that support more agile working environments.

Hyperscalers

Hyperscalers could gain ground by moving quickly to partner with various value chain players to test and verify priority use cases across domains. They could also form partnerships with industry players to drive automotive-specific standards in their core cloud and emerging edge segment. To determine their full range of potential opportunities—as well as the most attractive ones—hyperscalers should first analyze their existing assets and capabilities, such as their existing cloud infrastructure and services. They would also benefit from aligning their cloud and edge product portfolios or by extending cloud-availability zones to cover leading locations for V2X use case rollouts and real-world testing. If hyperscalers want to increase the footprint of their cloud and edge offerings within the automotive value chain, they could consider a range of partnerships, such as those with OEMs to test and verify use cases.

The benefits of 5G and edge computing are real and fast approaching, but no single player can go it alone. There are opportunities already at scale today that are not clearly addressed in the technological road map of many automotive companies, and not everybody is capturing them.

Building partnerships and ecosystems for bringing a connected car to market and capturing value are crucial, and some semiconductor companies are already forging strong relationships with OEMs and others along the value chain. The ACES trends in the automotive industry are moving fast; semiconductor companies must move quickly to identify opportunities and refine their existing strategies. These efforts will not only help their bottom lines but also could also allow tier 1s and OEMs to shorten the time-to-market for their products and services, which would accelerate the adoption of smart vehicles—and that benefits everyone.

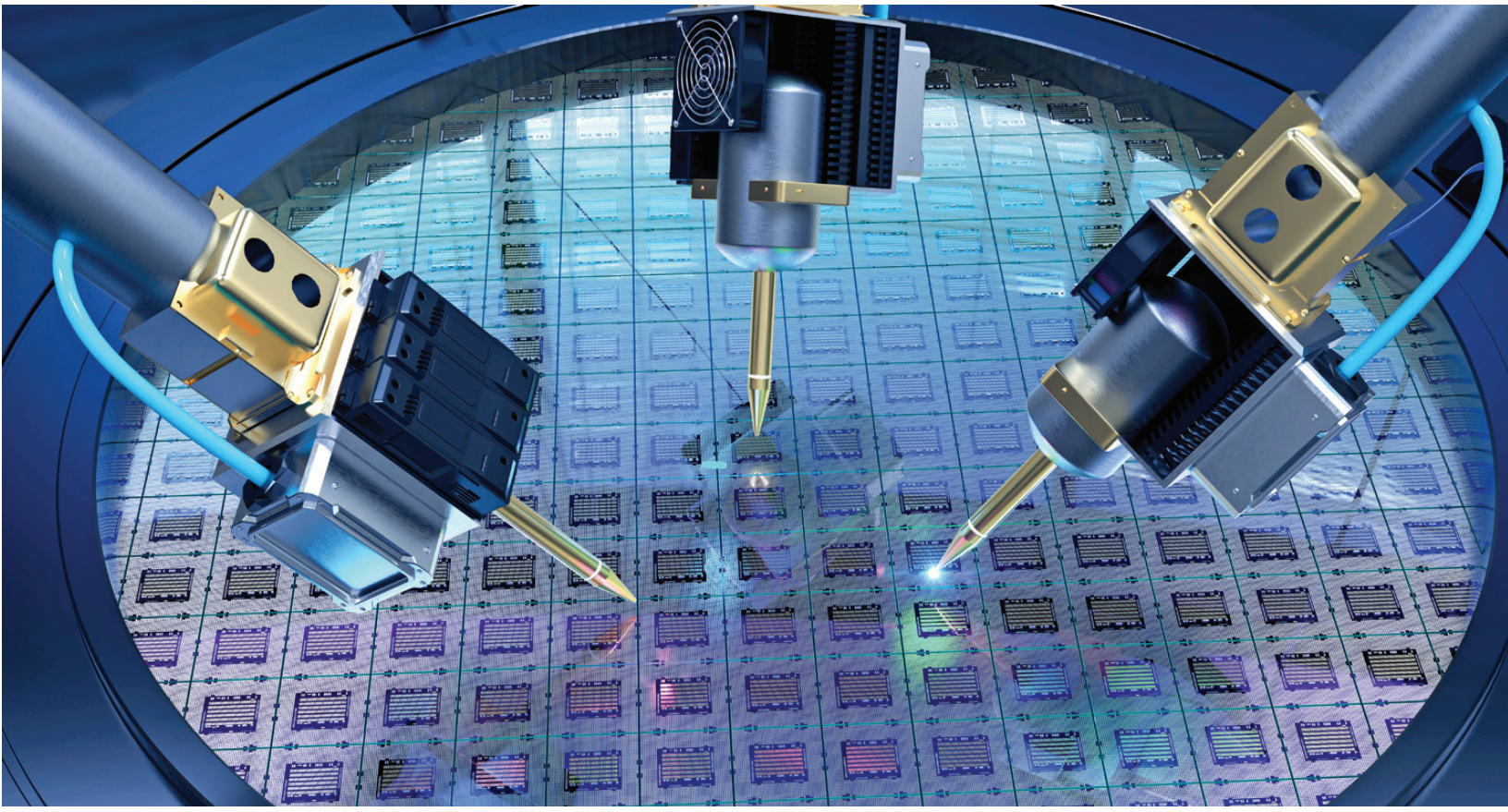
Philip Arejola is a specialist in McKinsey's Wrocław office, **Ondrej Burkacky** is a senior partner in the Munich office, **Johannes Deichmann** is a partner in the Stuttgart office, **Gourav Ganguly** is a specialist in the Mumbai office, **Asif Khan** is a consultant in the Stockholm office, and **Martin Wrulich** is a senior partner in the Vienna office.

The authors wish to thank the Global Semiconductor Alliance members who provided invaluable input, as well as Cathrine Särnqvist and Fabian Steiner for their contributions to this article.

New silicon carbide prospects emerge as market adapts to EV expansion

Rising electric-vehicle adoption is boosting demand for crucial silicon carbide power electronics components. How can semiconductor players, automotive OEMs, and others create value amid disruption?

This article is a collaborative effort by Albert Brothers, Ondrej Burkacky, Julia Dragon, Jo Kakarwada, Abhijit Mahindroo, Jwalit Patel, and Anupama Suryanarayanan, representing views from McKinsey's Semiconductor Practice.



The electric-vehicle (EV) market is estimated to grow at a 20 percent CAGR through 2030, when sales of xEVs are estimated to reach 64 million—four times the estimated EV sales volume in 2022.[1] Ensuring the EV component supply is sufficient to meet this rapid rise in estimated demand is critical, and the supply of silicon carbide (SiC) merits special consideration. Our analysis shows that compared to their silicon-based counterparts,[2] SiC metal-oxide-semiconductor field-effect-transistors (MOSFETs)[3] used in EV powertrains (primarily inverters, but also DC-DC converters and onboard chargers)[4] provide higher switching frequency, thermal resistance, and breakdown voltage. These differences contribute to higher efficiency (extended vehicle range) and lower total system cost (reduced battery capacity and thermal management requirements) for the powertrain. These benefits are amplified at the higher voltages needed for battery electric vehicles (BEVs), which are expected to account for most EVs produced by 2030.

In this article, we will examine how SiC manufacturers, automotive OEMs, and others can seize the opportunities inherent in the projected EV market growth urge to create value and gain competitive advantages.

Extensive market growth projected for EVs and SiC by 2030

Between 2018 and 2022, projections for EVs' share of the global light-vehicle market in 2030 increased 3.8 times, from around 17 million to 64 million units (Exhibit 1). This growth has been fueled by the expectation that EVs will reach total cost of ownership (TCO) parity with internal-combustion vehicles (ICEs) in many countries by 2024 or 2025,[5] as well as by the regulatory actions taken and investments made in EVs and charging infrastructure as part of the push to meet net-zero targets.

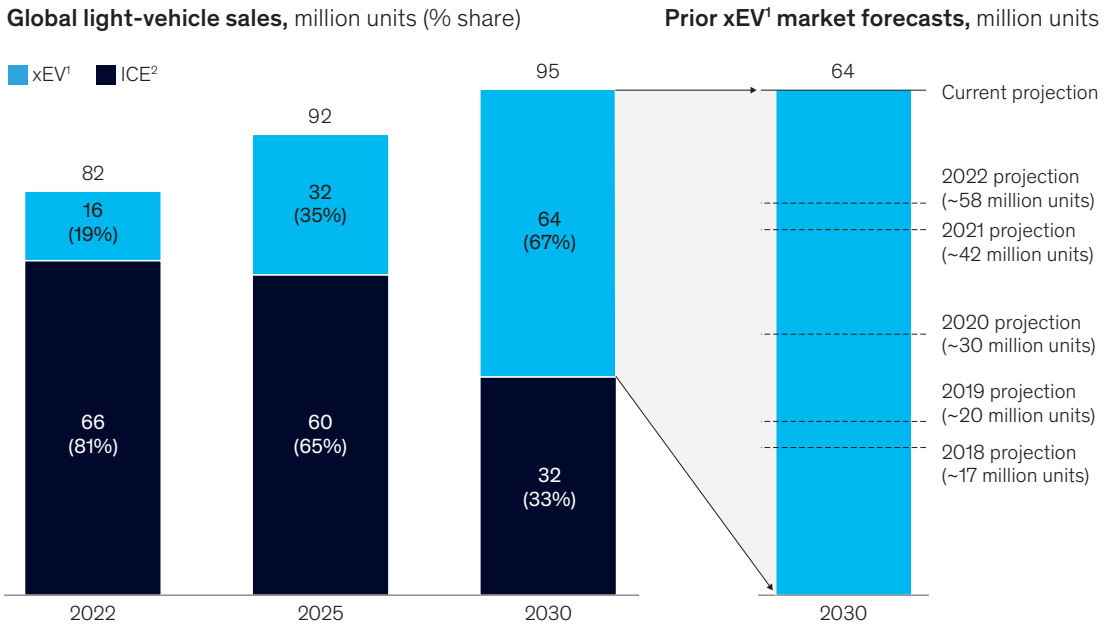
The SiC device market, valued at around $2 billion today, is projected to reach $11 billion to $14 billion in 2030, growing at an estimated 26 percent CAGR (Exhibit 2). Given the spike in EV sales and SiC's compelling suitability for inverters, 70 percent of SiC demand is expected to come from EVs. China, where

Ensuring the EV component supply is sufficient to meet this rapid rise in estimated demand is critical, and the supply of silicon carbide merits special consideration.

[1] Based on data from the McKinsey Center for Future Mobility.
[2] That is, silicon insulated-gate bipolar transistors (IGBTs).
[3] A MOSFET is an electronically controlled switch.
[4] An inverter is a device that converts DC power from the EV battery to AC supply for the EV motor.
[5] Excluding subsidies. With subsidies, TCO is already at parity between EVs and ICE vehicles.

Exhibit 1

The push to achieve net-zero objectives has accelerated the pace of electric-vehicle adoption.



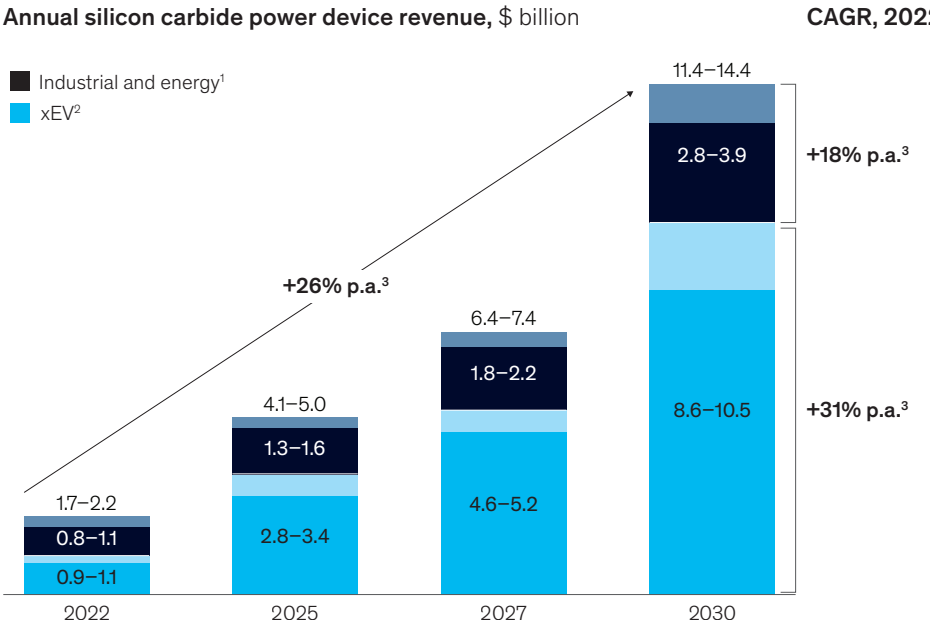
Note: Figures may not sum, because of rounding.
[1]xEV includes battery electric vehicles (BEVs), hybrid electric vehicles (HEVs), plug-in hybrid electric vehicles (PHEVs), and fuel-cell electric vehicles (FCEVs).
[2]Internal-combustion engine.
Source: McKinsey Center for Future Mobility

McKinsey & Company

The SiC device market, valued at around $2 billion today, is projected to reach $11 billion to $14 billion in 2030, growing at an estimated 26 percent CAGR.

Exhibit 2

The silicon carbide device market is estimated to grow at a CAGR of 26 percent between 2022 and 2030.



Note: Data is as of November 2022.
[1]Other applications as a share of industrial and energy include power supplies (23%), industrial applications (14%), commercial vehicles (12%), uninterruptable power supplies (12%), and military and aerospace (12%).
[2]xEV includes battery electric vehicles (BEVs), hybrid electric vehicles (HEVs), plug-in hybrid electric vehicles (PHEVs), and fuel-cell electric vehicles (FCEVs).
[3]Per annum.
Source: McKinsey Center for Future Mobility, Current Trajectory Scenario

McKinsey & Company

anticipated EV demand is highest, is projected to drive around 40 percent of the overall demand for SiC in EV production.

Across EVs, the type of powertrain—BEV, hybrid electric vehicle (HEV), plug-in hybrid electric vehicle (PHEV), 400-volt, or 800-volt—determines the benefits and relative uptake of SiC. Because of their greater efficiency needs, 800-volt BEV powertrains are most likely to use SiC-based inverters.[6] According to our analysis, by 2030, BEVs are expected to account for 75 percent of EV production (up from 50 percent in 2022), while HEVs and PHEVs will make up the other 25 percent. Furthermore, we anticipate more than 50 percent market penetration for 800-volt powertrains by 2030 (up from less than 5 percent in 2022). Accordingly, we anticipate a significant tailwind for SiC devices in the coming decade

Vertical integration: A compelling business model in the SiC market

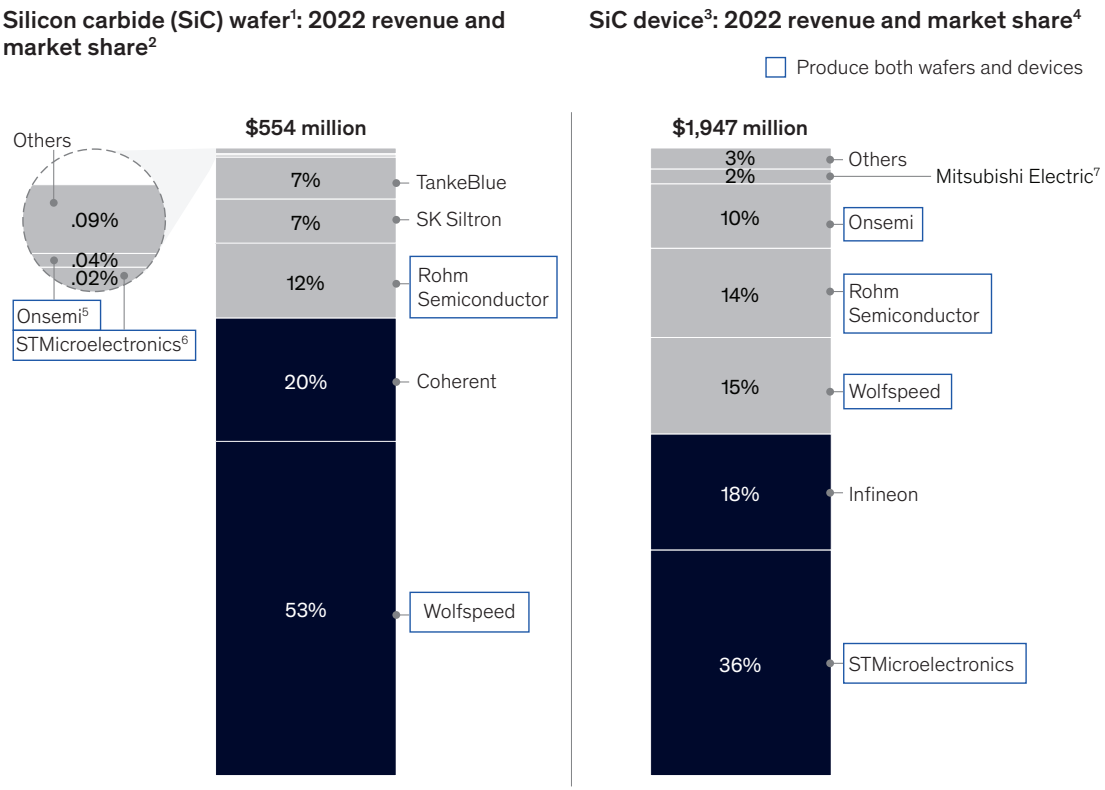
The current SiC market is highly concentrated, with only a few end-to-end leaders. Indeed, the top two companies in the SiC wafer and device markets control around 60 to 65 percent of SiC market share (Exhibit 3).

The market rewards vertical integration, as evidenced by the dominance of the mostly integrated leading players. According to our analysis, vertical integration

[6] Based on data from the McKinsey Center for Future Mobility.

Exhibit 3

The top two players served around 55 to 75 percent of device and wafer markets in 2022.



Note: Figures do not sum to 100%, because of rounding.
[1]Includes only finished raw SiC wafers.
[2]Revenues and market shares of Rohm Semiconductor, SK Siltron, TankeBlue, STMicroelectronics, and Onsemi estimated based on 2021 market shares as a proxy.
[3]Discretes and modules.
[4]Revenues and market shares of Mitsubishi Electric estimated based on 2021 market shares as a proxy.
[5]Approximate revenue equivalent for GT Advanced Technologies.
[6]Market share for Norstel.
[7]Fully integrated end to end, but captive.
Source: *Power SiC 2023*, Yole Group, August 2023; McKinsey analysis

McKinsey & Company

in SiC wafer and device manufacturing can improve yield by five to ten percentage points and margins by ten to 15 percentage points,[7] partly from lower yield loss and partly from eliminating margin stacking at each step in the process (Exhibit 4). Higher yields are achieved from better control over design and faster yield ramps with closed-loop feedback between wafer and device manufacture.

Strategically, vertically integrated manufacturers can also offer a stronger value proposition to automotive OEMs because of higher supply assurance, which is noteworthy in light of recent supply chain challenges. Similarly, vertical integration also offers wafer players a hedge against commoditization, such as has occurred in the silicon market.

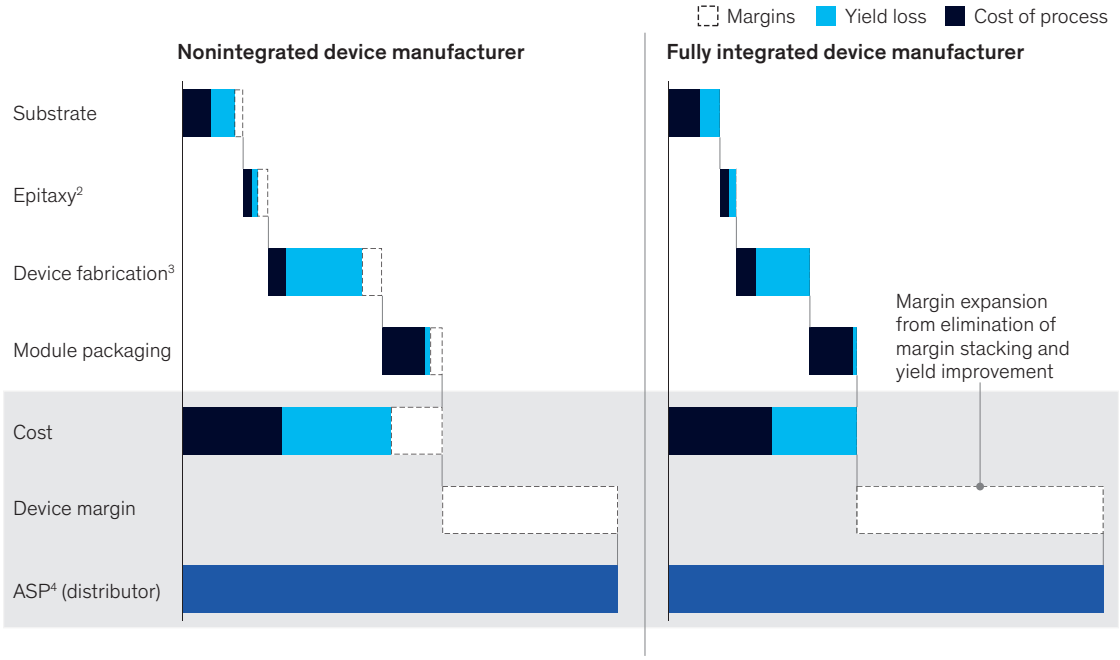
Not surprisingly, several leading manufacturers have already evolved toward vertical integration through M&A and partnerships. In particular, semiconductor device manufacturers have added upstream capacity in wafer materials manufacturing. This includes

[7] Compared to a combination of pure-play providers across these segments of the value chain.

Exhibit 4

Vertical integration in silicon carbide device manufacturing can help realize significant increases in margin and yield.

6-inch silicon carbide MOSFET[1]: Relative cost comparison by value chain step ($ per wafer), 2022



[1]Metal-oxide-semiconductor field-effect-transistor.
[2]Epitaxy yields representative of epitaxy providers and integrated device manufacturers. Leading epitaxy providers offer a yield improvement over in-house epitaxy, even for integrated manufacturers.
[3]Device fabrication includes dicing and probe test.
[4]Average selling price.
Source: *SiC transistor comparison 2021*, Yole Group, December 2021; McKinsey analysis

McKinsey & Company

the STMicroelectronics' acquisition of Norstel, Onsemi's acquisition of GT Advanced Technologies (GTAT), and the Rohm Semiconductor acquisition of SiCrystal.[8] These and other acquisitions demonstrate confidence in the operational, financial, and strategic benefits of vertical integration.

Transitioning to 8-inch wafers can offer price, margin, and market advantages

According to our analysis, a transition from the production and use of six-inch wafers to eight-inch wafers is anticipated, with material uptake beginning around 2024 or 2025 and 50 percent market penetration reached by 2030. Once technological challenges are overcome, eight-inch wafers offer manufacturers gross margin benefits from reduced edge losses, a higher level of automation, and the ability to leverage depreciated assets from silicon manufacturing. Our analysis projects the gross margin benefit of this transition to be about five to ten percentage points, depending on the level of vertical integration.

Volume production of eight-inch wafers in the United States is projected to begin in 2024 and 2025, when industry-leading manufacturers are

[8] For more on the STMicroelectronics–Norstel acquisition, see "STMicroelectronics closes acquisition of silicon carbide wafer specialist Norstel AB," STMicroelectronics, December 2, 2019; for more on the Onsemi–GTAT acquisition, see "Onsemi completes acquisition of GT Advanced Technologies," Onsemi, November 1, 2021; for more on the Rohm Semiconductor–SiCrystal acquisition, see "History of SiCrystal," SiCrystal, accessed September 5, 2023.

slated to bring capacity online.[9] Production of eight-inch wafers is expected to ramp rapidly thereafter, chiefly in response to demand and price pressures (especially from midtier-volume EV OEMs), as well as to cost savings realized by conversion to eight-inch SiC wafer fabrication.

Our analysis shows that eight-inch wafer substrates are still relatively more expensive per square inch compared to six-inch wafers, due to lower yields. However, the gap is expected to close for leading manufacturers in the coming decade because of process yield improvement and novel wafering technologies. For instance, we find that, compared to the conventional wafering technique with multi-wire saws, laser-cutting techniques have the potential to more than double the number of wafers produced from one monocrystalline boule. And advanced wafering techniques such as hydrogen splitting could further increase the output.

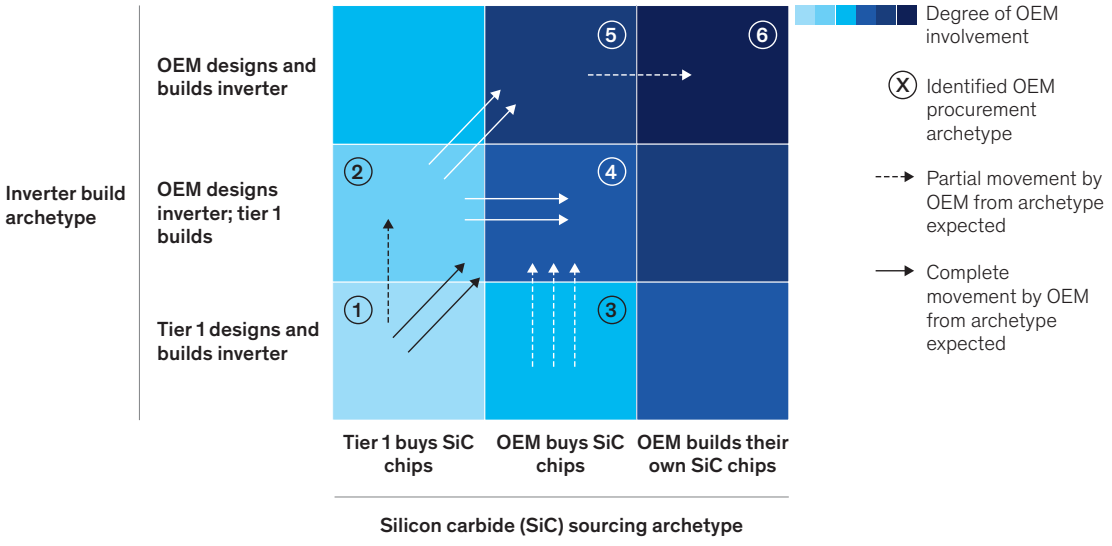
Greater involvement in the SiC value chain creates new priorities for automotive OEMs

Acute supply chain challenges, geopolitical considerations, the transition to 800-volt vehicles, and the resulting increase in demand for SiC MOSFETs have all prompted recent expansions of OEM involvement in semiconductor and SiC sourcing. Given recent supply chain disruptions and the developing SiC landscape, with anticipated major technological innovations, automotive OEMs engage in multiple sourcing models for both SiC-based EV inverters and the underlying SiC chips (Exhibit 5). Our analysis shows that, as the industry matures, preferences are likely to shift toward greater OEM involvement in sourcing SiC as well as designing inverters. This shift also manifests itself in a growing number of partnerships between SiC manufacturers and automotive OEMs.

[9] McKinsey analysis based on announcements from SiC wafer and device manufacturers.

Exhibit 5

OEM involvement in silicon carbide sourcing and component manufacturing will prompt changes across the power component value chain.



McKinsey & Company

OEMs have engaged in numerous partnerships but few exclusive agreements

Partnerships between SiC manufacturers and OEMs range from long-term supply agreements to strategic and development partnerships—and even to co-investments and joint venture agreements in manufacturing facilities. Our analysis of public announcements[10] from 18 automotive OEMs representing more than 75 percent of 2030 BEV volume found that 12 OEMs (representing more than 60 percent of 2030 BEV volume) have already announced two or more partnerships with SiC manufacturers. Five OEMs (representing around 15 percent of BEV volume) have announced one partnership, while only one OEM (representing around 2 percent of BEV volume) has not announced a partnership with a SiC manufacturer. While this analysis is limited to announced partnerships, there is a clear trend toward automotive OEMs diversifying and securing their supply chain with nonexclusive partnerships (Exhibit 6).

This high level of OEM involvement indicates that incumbent and prospective SiC manufacturers that develop deep relationships with OEMs and have automotive-specific device capabilities will be best positioned to participate in the growth of this sector. SiC manufacturers seeking to ensure share of wallet may wish to secure partnerships early, given barriers to demonstrating technical proficiency and assuring access to supply. This is particularly pertinent in light of the long-term nature of many supplier–OEM relationships. Furthermore, less-established SiC manufacturers may need to build early partnerships with OEMs to achieve a proof of concept and demonstrate assurance of supply to

[10] As of April 19, 2023.

Exhibit 6

Several supply partnerships between OEMs and silicon carbide manufacturers have been announced, but few are exclusive.

Number of announced silicon carbide (SiC) partnerships across automotive OEMs

McKinsey & Company

There is a clear trend toward automotive OEMs diversifying and securing their supply chain with nonexclusive partnerships.

be designed into automotive platforms. Our analysis shows that OEMs are likely to be open to multiple partnerships with less-established manufacturers to create new avenues of assured supply.

Chinese OEMs signal increased local sourcing, but leaders have yet to emerge

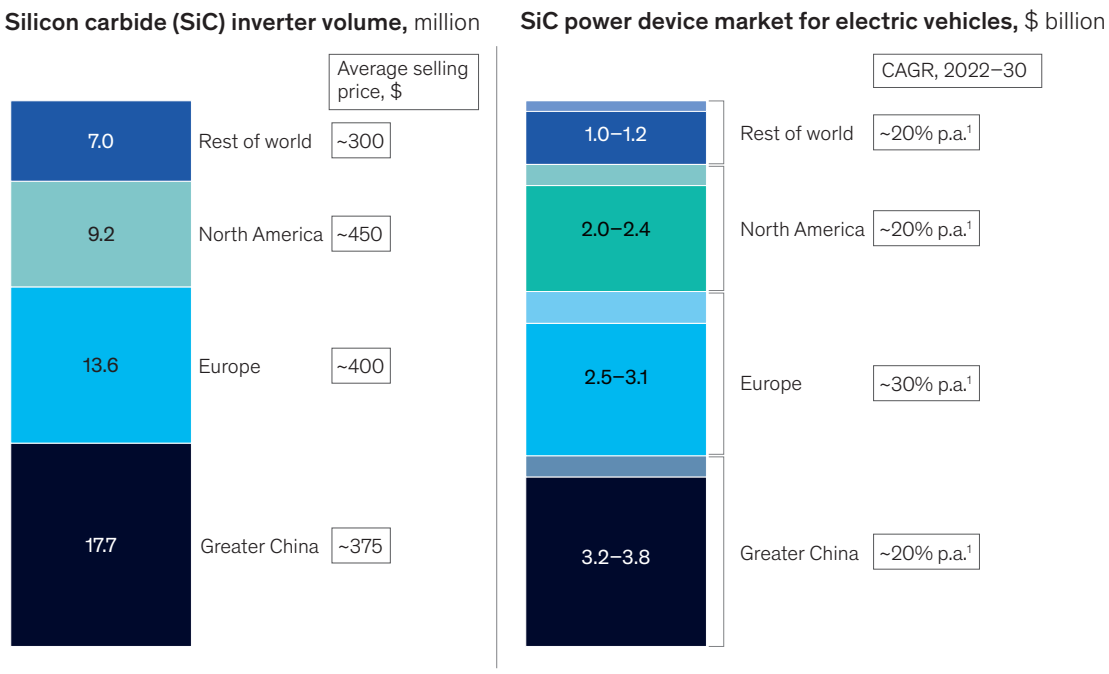
China is expected to remain the largest SiC market through 2030 (Exhibit 7), with growth driven by consumer demand and supported by popular incentives, such as EVs' exemption from license plate quotas. According to McKinsey research and analysis, this market is approximately one-third Chinese OEMs and two-thirds foreign OEMs in China, a mix that is expected to shift toward Chinese OEMs and approach a more even split by 2030.

Currently, non-Chinese SiC manufacturers supply 80 percent of the wafer market in China and more than 95 percent of the device market. However, our analysis shows that Chinese OEMs are increasingly seeking local supply sources due to geopolitical and supply assurance considerations. Given sufficient capacity and technological performance, Chinese OEMs are expected to broadly shift procurement to local suppliers, from what is currently approximately 15 percent to around 60 percent by 2030 (Exhibit 8).

This shift to local procurement in China is expected to be enabled by a rise in Chinese players across the whole SiC value chain—from equipment supply, to wafer and device manufacture, to system integration. Chinese equipment suppliers already cover all major SiC fabrication steps and have announced investments to ramp up capacity through 2027. However, clear supply leaders have yet to emerge in the Chinese ecosystem.

Exhibit 7

China is expected to remain the largest market for silicon carbide through 2030.



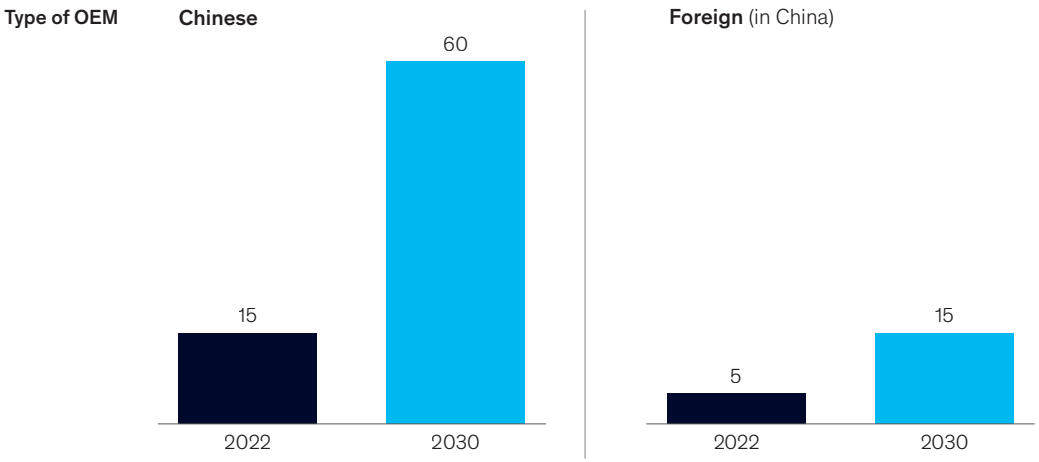
[1]Per annum.
Source: McKinsey Center for Future Mobility, Current Trajectory Scenario; Yole Group component teardown tracks; McKinsey analysis

McKinsey & Company

Exhibit 8

Chinese OEMs are expected to increasingly prefer sourcing silicon carbide locally, from 15 percent in 2022 to 60 percent by 2030.

Expected silicon carbide device share of wallet procured locally within China, %



McKinsey & Company

How stakeholders can make the most of the SiC demand surge

The accelerating adoption of EVs and the increasingly vital role of SiC in the growing EV market denotes fundamental implications for players across the SiC value chain. While there is no paramount strategy to lead with increased market share or value creation, some considerations are imperative for players to position themselves for primacy in the shifting SiC market.

Automotive OEMs and tier-one suppliers

Well-positioned automotive OEMs and tier-one suppliers will have EV and SiC adoption and timing plans that are aligned with the market and their peers. As OEM and tier-one partnerships are formed early in the development process, SiC inverter and semiconductor supply chain strategies tailored to internal capabilities and growth strategy—for example, co-development partnerships with SiC device manufacturers versus more straightforward supply agreements—are highly advantageous in securing and maintaining partnerships. With advancements in technology such as trench topologies for transistors and hybrid Si-SiC inverter designs and continued shifts in the value chain, designing a holistic sourcing strategy that takes uncertainty into account will similarly serve OEMs and tier-one suppliers well.

Semiconductor component manufacturers

Defining a SiC growth and investment strategy that keeps pace with the growing opportunity for SiC across the EV and other markets is central to any well-situated semiconductor component manufacturer's outlook. Access to the market with appropriately defined partnerships with automotive OEMs and tier-one suppliers is likewise vital, as is continued investment in technology development, capacity ramp-up execution, and cost degression—particularly in light of a transition to eight-inch wafers. Players will continue to shape and be shaped by build-buy-partner decisions across the manufacturing value chain, including those related to substrate, epitaxy, and devices.

Prospective investors in SiC

Ideally, a SiC investment thesis incorporates an assessment of reinvestments and time to maturity that is aligned with the market, value chain, and technology dynamics. It is important for investors to consider

which players are likely to emerge as leaders as the market matures, whether announced capacities are likely to come online as scheduled, and whether there are opportunities to disrupt and create substantial value with strategically chosen investments.

Governments

Incentives or ecosystem enablers can help governments support local demand for SiC for use in EVs and other applications. International frameworks that support the value chain and safeguard national interests could help support a global supply chain while fulfilling demands for localization and supply resiliency.

The adoption of EVs represents a significant opportunity for players in the silicon carbide value chain. Competitive gains will likely be realized by those companies that attend to trends and opportunities in the SiC ecosystem and quickly build key capabilities and partnerships to support their growth ambitions. The SiC value chain is dynamic and has a high degree of uncertainty. There have been significant shifts in the demand environment: changes in inverter design and the MOSFET need per inverter; the continued acceleration of EV demand; the value chain, including emerging players in China and investments in the SiC value chain by nontraditional players such as automotive OEMs; regulatory postures; and technology, including the rise of new wafering techniques improving yield. In this environment, all market participants gain strategic advantages from monitoring developments on an ongoing basis and building flexibility into their plans.

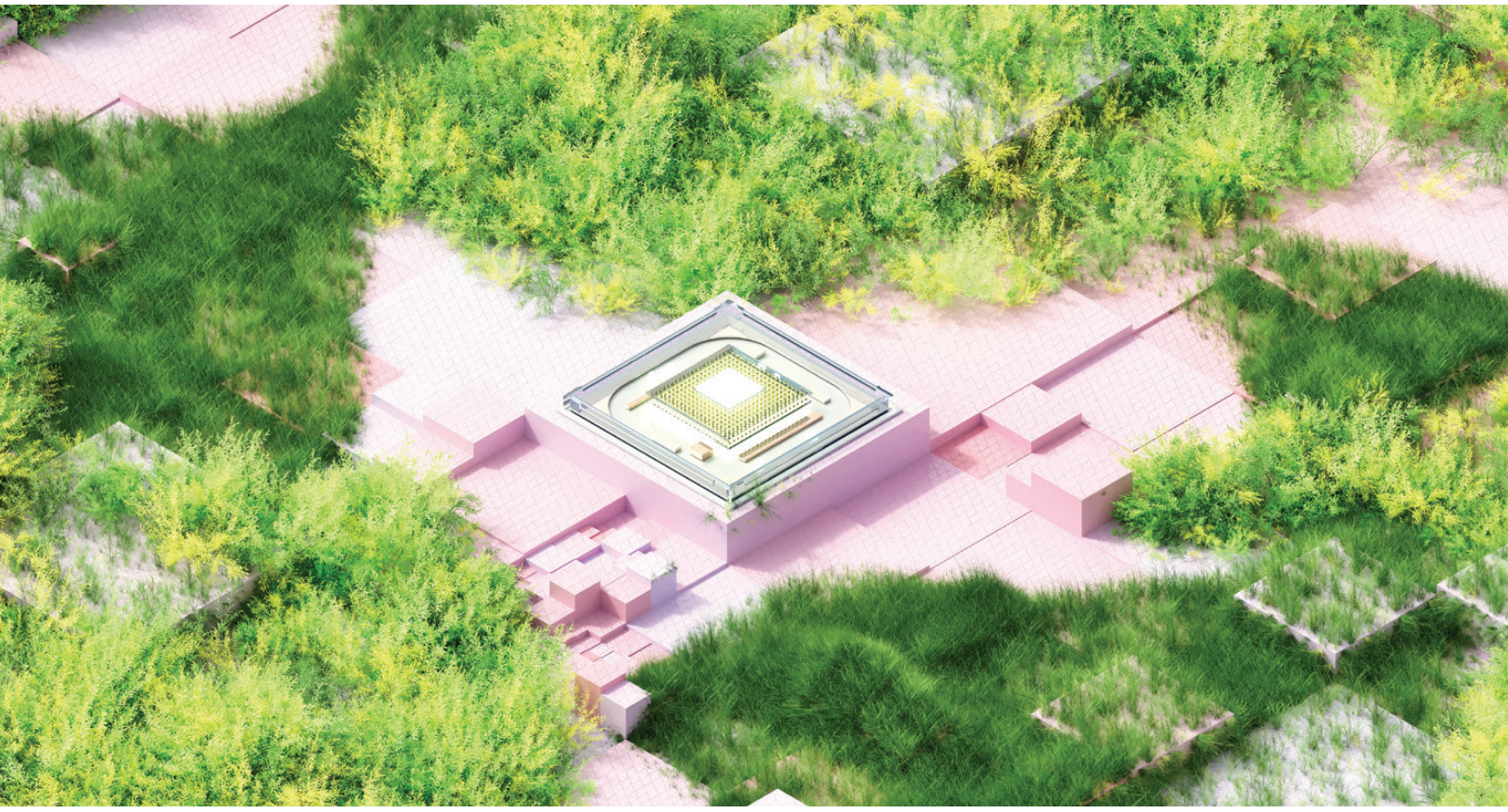
Albert Brothers is a consultant in McKinsey's New York office; **Ondrej Burkacky** is a senior partner in the Munich office; **Julia Dragon** is an associate partner in the Frankfurt office; **Jo Kakarwada** is a consultant in the Carolinas office; **Abhijit Mahindroo** is a senior partner in the Southern California office; **Jwalit Patel** is an associate partner in the Dallas office; and **Anupama Suryanarayanan** is an associate partner in the Silicon Valley office.

The authors wish to thank Michael Guggenheimer, Zachary Salyer, Dennis Schwedhelm, Brandon Strecker, Andreas Tschiesner, and members of the McKinsey Center for Future Mobility team for their contributions to this article.

Beyond the fab: Decarbonizing Scope 3 upstream emissions

As the imperative to achieve net-zero emissions grows, semiconductor companies are increasingly focused on supplier emissions.

This article is a collaborative effort by Martin Burkardt, Felix Dietrich, Sebastian Göke, Mark Nikolka, Mark Patel, Peter Spiller, and Tuisku Suomala, representing the views of McKinsey's Semiconductors Practice.



Every step of semiconductor manufacturing, from wafer manufacturing through packaging, requires fossil fuels and generates emissions. To address climate change, some semiconductor companies have recently taken major steps to decarbonize. These efforts are the first steps in the transition to more sustainable operations, and they are intensifying as the companies' major end customers establish even more ambitious emissions reduction targets. Intel, for instance, recently committed to net-zero greenhouse-gas (GHG) emissions in its global operations by 2040 and has targeted achieving 100 percent use of renewable electricity as an interim milestone in 2030. Since semiconductor companies are often among the suppliers with the greatest emissions, they will face particular scrutiny as end customers increasingly look upstream.

As semiconductor companies race to meet supplier expectations, they must broaden their decarbonization efforts. To date, most of their programs have focused on two emissions categories: those directly related to activities within their fabs (Scope 1) and those arising from the generation of purchased electricity, steam, heating, and cooling equipment (Scope 2). Since Scope 1 and 2 emissions represent only 65 percent of total GHGs from fabs, companies can only meet their customers' net-zero goals by expanding their efforts to include Scope 3 upstream emissions—those originating from the suppliers that provide services or silicon and other materials for chip manufacturing.[1]

Many semiconductor companies have hesitated to address Scope 3 upstream emissions because of the challenges encountered to date when attempting to create transparency about emissions and drive decarbonization initiatives. The problems largely arise because emissions are fragmented across hundreds of suppliers and thousands of materials. But semiconductor companies may now overcome these hurdles by applying new methodologies, leveraging automated baselining tools, and driving their Scope 3 upstream decarbonization in cross-functional programs that have top management support. In this article, we explore the way semiconductor players can decarbonize their Scope 3 upstream emissions by facilitating cooperative efforts with suppliers, improving waste management, redesigning or enhancing product specifications, and optimizing use of materials.

Since Scope 1 and 2 emissions represent only 65 percent of total greenhouse-gas emissions from fabs, companies can only meet their customers' net-zero goals by expanding their efforts to include Scope 3 upstream emissions.

[1] Scope 3 downstream emissions are related to the use of products that include semiconductors and are not discussed in this article.

Scope 3 upstream emissions: An overlooked but important category

Our analysis suggests that upstream emissions for a typical fab have three main sources (Exhibit 1):

— purchased materials (representing about 62 percent of all Scope 3 upstream emissions)

— maintenance services, spare parts, and capital expenditures for equipment upgrades (about 22 percent)

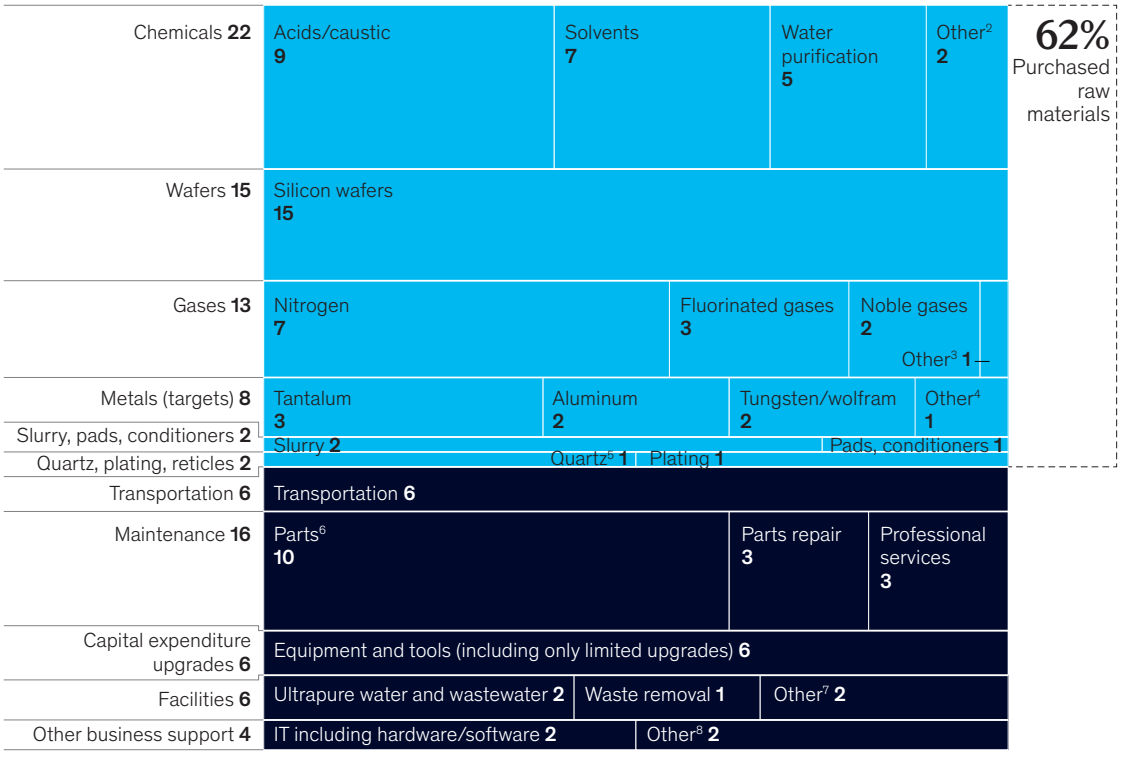
— supplier transportation, such as material deliveries (about 6 percent)

Of course, every fab may differ from the norm in some respects. For instance, some companies may already rely on chemical suppliers that prioritize renewable energy. But one factor common to many fabs is a lack of clarity about Scope 3 upstream emissions, including the amount associated with specific materials, services, or suppliers. For example, nitrogen trifluoride (NF3), which is commonly used in semiconductor fabrication, rates very high for global warming potential. Fugitive emissions—those that escape unintentionally—related to the production of NF3 may be higher than the emissions associated with the actual NF3 production process, but these are very difficult to quantify.

Exhibit 1

Purchased raw materials account for 62 percent of Scope 3 emissions.

Share of Scope 3 (upstream) emissions, typical semiconductor fabrication,[1] %



Note: Figures may not sum to 100%, because of rounding.
[1]Emissions averaged across 300-millimeter semiconductor fabs with node sizes ranging 40–90 nanometers. [2]Photoresists and specialty chemicals. [3]Deuterium, silane, and oxygen. [4]Mainly copper, gold, and titanium. [5]Including quartz reticles. [6]Including consumable materials. [7]Including accessories, building management, etc. [8]Mainly professional services.

McKinsey & Company

Many semiconductor companies may also use misleading assumptions to calculate their Scope 3 upstream emissions. Consider aluminum. While most industries can use 99 percent pure aluminum with no complications, semiconductor companies often require 99.9 percent purity—and that slight improvement requires far more energy, partly because of repetitive melting and cooling as well as the electrochemical purification required, which increases the emissions. The same pattern holds true for many other materials.

While fabs may deal with hundreds of suppliers during procurement, our analysis revealed that about six to ten suppliers will account for half of emissions for chemicals, wafers, and gases (Exhibit 2). For maintenance, spare parts, and capital expenditures for equipment upgrades, about three to five suppliers will account for over half of emissions. These patterns mean that semiconductor companies can address the majority of Scope 3 upstream emissions by focusing on a relatively small group of suppliers.

The first step in decarbonizing Scope 3 upstream emissions involves establishing a detailed and reliable baseline. This can be done by examining procurement data for Tier 1 suppliers, including the exact quantities of materials purchased. While fabs will ideally base their analysis on volume data whenever possible, this information may be unavailable or irrelevant for some categories, such as services. In such cases, they must examine spending levels instead. When estimating emissions associated with different materials and services, our approach calls for fabs to factor in the high-energy requirements needed to create semiconductor-grade quality materials (for more information, see sidebar "Our methodology for estimating Scope 3 upstream emissions").

Exhibit 2

Ten or fewer suppliers typically contribute half of emissions for chemicals, wafers, and gases.

Supplier base, typical semiconductor fabrication[1]

■ Indicative number of suppliers ☐ Number of suppliers responsible for 50% of the category's Scope 3 emissions

Purchased raw materials	Chemicals	3–5 out of 25
	Wafers	2–3 out of 10
	Gases	1–2 out of 10
	Metals (targets)	1–2 out of 5
	Slurry, pads, conditioners	1–2 out of 10
	Quartz, plating, reticles	2–3 out of 15
Maintenance and capital expenditure upgrades	Maintenance	3–5 out of ≥100[2]
	Capital expenditure upgrades	2–3 out of 20
Transport	Upstream logistics	n/a
Other	Facilities	5–6 out of ≥100[2]
	Other business support	10–15 out of ≥100[2]
	TOTAL	25–40 out of ≥350[3]

[1]Emissions averaged across 300-millimeter semiconductor fabs with node sizes ranging 40–90 nanometers. Very small longtail suppliers excluded from total number of suppliers due to negligible impact for emissions.
[2]Number of suppliers dependent (eg, on maturity of alternative parts sourcing and usage of OEM/local/3rd-party maintenance suppliers).
[3]Some suppliers are serving several categories.

McKinsey & Company

Our methodology for estimating Scope 3 upstream emissions

To establish a baseline for Scope 3 upstream emissions, we examined procurement data for a typical fab and leveraged from McKinsey's sustainability solution, Catalyst Zero, which analyzes proprietary data for more than 300,000 spending- and consumption-based emissions factors. The database also considers information on material volumes and recent improvement actions executed by semiconductor fabs and suppliers. Our analysis focused on supplier data for typical 300 millimeter semiconductor fabs with node sizes ranging 40 to 90 nanometers in Asia, Europe, and the United States.

Finding the right decarbonization pathways for Scope 3 upstream emissions

Once semiconductor companies have established emissions baselines for Tier 1 suppliers, they must identify the right set of decarbonization levers to reduce their emissions, focusing on the suppliers and materials that contribute most emissions.

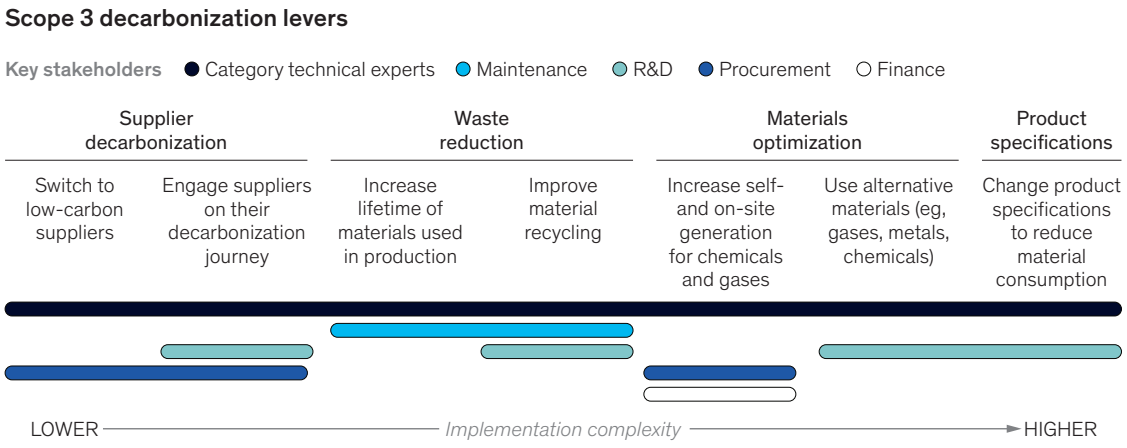
The full range of levers spans four areas and requires the involvement of different stakeholders in the organization (Exhibit 3). Semiconductor companies should consider all of these levers to develop the most effective and efficient approach to decarbonization.

Some gains may come from requiring suppliers to use renewable materials or from identifying ways to reduce waste within fabs, thereby decreasing the volume of materials ordered. Other levers involve optimizing the materials used (for instance, using metals with a lower emission footprint or changing product specifications to reduce the need for high-emissions materials). Fabs should consider the cost and decarbonization potential of all levers, but their ability to implement them will differ. The internal stakeholders or functions involved will also vary.

Top leadership commitment and involvement is essential to drive decarbonization across the different functions, including operations, technology, development, and procurement.

Exhibit 3

The full set of Scope 3 upstream decarbonization levers can be grouped into four areas.



McKinsey & Company

Supplier decarbonization

When helping suppliers decarbonize, fabs may want to focus on the top three materials categories—chemicals, wafers, and gas—because of their outsize impact on Scope 3 upstream emissions.

In many cases, fabs may be able to reduce emissions by switching from their current Tier 1 suppliers to those with lower carbon footprints. But if they want to preserve their existing relationships or have no alternatives, fabs can cooperate with their current suppliers to accelerate decarbonization programs. For instance, they might jointly agree on emissions reduction targets, identify abatement levers, and define execution road maps.

For many purchased materials, including wafers, hydrogen peroxide, nitrogen fluoride, and aluminum, energy from conventional sources will account for more than half of emissions during production. Fabs might be able to reduce this percentage by offering financial incentives or other benefits to encourage suppliers to increase their use of renewable energy. The ability to apply this lever will vary, however, since suppliers may be located in countries that offer limited access to renewable energy. For some chemicals and process gases, fabs might also collaborate with suppliers that can offer innovative production processes or synthetic routes that lower overall emissions.

Many emissions arise from Tier 2 suppliers that do not interact directly with fabs, or from suppliers that are even further upstream. These suppliers also contribute to a fab's Scope 3 upstream emissions, so they need to be part of any decarbonization efforts. Tier 1 suppliers might be best positioned to encourage Tier 2 suppliers to reduce emissions, since their relationship as end customers may give them more leverage. While it may seem difficult for suppliers to cooperate along the value chain, examples of past collaborations exist. For instance, semiconductor companies and Tier 1 suppliers of tools and chemicals worked closely together to reduce the usage of perfluorocarbons when regulators began discouraging their use.

Although semiconductor companies deal with too many suppliers to develop an individual decarbonization plan for each one, they can still drive improvement throughout the entire vendor base by implementing new procurement strategies, such as policies that give preference to suppliers that disclose their emissions or that have lower emissions than their competitors (see sidebar "Decarbonizing wafer production" for examples of concrete steps that fabs can take).

Waste reduction

Over the short to medium term, fabs can also reduce Scope 3 upstream emissions by taking steps to decrease waste within their facilities, but they must first balance trade-offs and identify potential risks. Consider wafer cleaning. A single wafer runs through more than 100 different chemical baths while being processed. Fabs might be able to increase the number of wafers processed in the same chemical bath to reduce waste, but cross-functional teams would first need to determine when reuse might decrease yield to an unacceptable level. Fabs might also investigate whether they can extend the life of machine parts by increasing predictive maintenance, or determine if they can reduce use of spare parts by specifying that they should only be replaced once specific triggers, such as particle counts, are exceeded.

Recycling—an area where few fabs undertake extensive efforts—could also decrease waste if leaders expand their current programs. For instance, they could investigate the possibility of introducing recycling programs for materials, such as ultra-high-purity aluminum, that have not been previously reused. Whenever recyclates are investigated, fabs must determine how they can remove impurities to meet the semiconductor industry's high standards. Moving recycling on-site could also decrease waste.

Materials optimization

In some cases, fabs may be able to use lower-emissions materials, chemicals, or gases during production. First, however, companies must

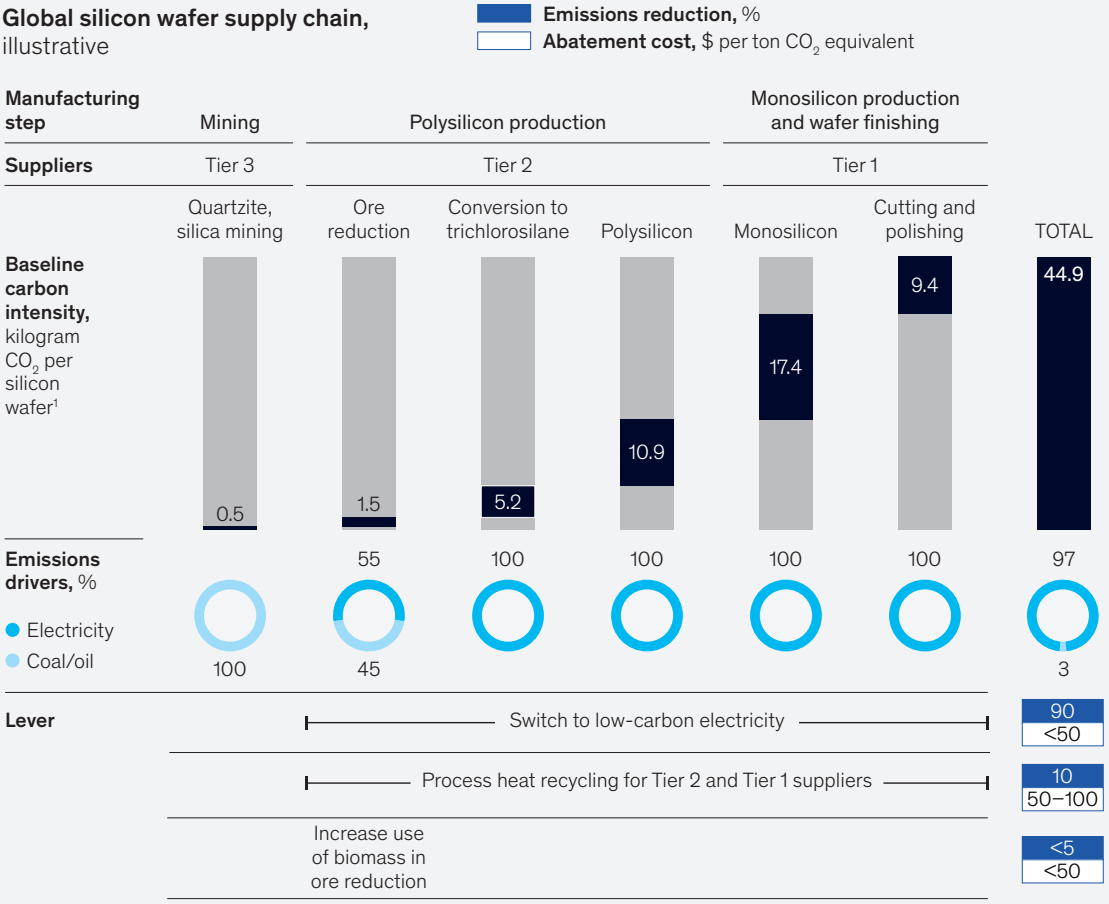
Decarbonizing wafer production

Wafers account for about 15 percent of Scope 3 upstream emissions at fabs, with 90 percent of the total resulting from the electricity required for ore to become polysilicon (a Tier 2 product), which is then transformed to monosilicon (Tier 1). To reduce emissions, semiconductor companies could investigate entirely new strategies, such as crafting monosilicon bids that favor Tier 1 suppliers that use a high percentage of green energy and engage in process heat recycling (exhibit). But they also need to look further upstream—and that is where they may encounter more difficulties. Chinese companies currently produce 80 to 90 percent of the world's polysilicon, and China's share of renewable energy is expected to remain under 50 percent through 2030, according to the McKinsey Global Energy Perspective 2022. Tier 1 suppliers might be in the best position to encourage the Tier 2 polysilicon companies to decarbonize, especially if they are critical end customers, since they will have more leverage and greater insights about polysilicon volumes. As with Tier 1 suppliers, use of green energy and process heat recycling will be important for Tier 2s, as will the greater use of biomass in the conversion of ore to polysilicon. If all efforts at decarbonization still fall short of net zero at Tier 1 companies and other upstream vendors, they might consider purchasing carbon credits.

Exhibit

Using low-carbon electricity can reduce emissions by up to 90 percent.

Global silicon wafer supply chain, illustrative



McKinsey & Company

set up R&D, quality, and engineering teams to identify better manufacturing methods, evaluate alternatives, and review product specifications. With any change in production or composition, the teams must evaluate trade-offs, such as a potential decrease in product performance, and determine what is acceptable.

To facilitate progress, semiconductor leaders may cooperate with industry organizations to develop green alternatives to emissions-heavy materials, chemicals, and gases, such as fluorinated chemicals.

Product-specification adjustments

Fabs might contemplate changing product specifications to reduce emissions—for instance, they could evaluate whether existing purity grades for certain metals are truly essential. Again, many trade-offs come into play. In some cases, loosening requirements may achieve the dual goal of reducing emissions and driving cost improvements. In other instances, however, less stringent requirements might introduce quality issues that lower yield.

Since fabs have not made any extensive efforts to reduce emissions by changing product specifications, their initial focus should involve fairly straightforward solutions, such as the use of lower-grade chemicals during less critical steps of the wafer-cleaning process. Over the longer term, they can consider more complex process changes, such as finding replacements for highly emissive materials, since these may become increasingly important as semiconductor end customers push for zero emissions throughout the supply chain.

Although the supplier landscape is fragmented, with many vendors and products, companies can still develop a viable strategy for reducing Scope 3 upstream emissions. With six to ten suppliers accounting for half of all emissions for chemicals, wafers, and gases—the top three materials categories—fabs may want to concentrate their initial efforts on this group. When implementing decarbonization levers, the effort must go beyond procurement to encompass product and operational changes. Such broad efforts may actually produce better results, since many semiconductor companies encounter difficulties when attempting to change suppliers. Because decarbonization efforts will be broad in scope, they will require the involvement of top leadership and stakeholders from all relevant groups, including operations, procurement, and R&D. Some semiconductor companies are already launching initiatives to reduce Scope 3 upstream emissions, and they could emerge as early leaders. Now it's time for others to follow their example.

Martin Burkardt is a consultant in McKinsey's Denver office; **Felix Dietrich** is a consultant in the Berlin office, where **Sebastian Göke** is an associate partner; **Mark Nikolka** is a consultant in the Munich office; **Mark Patel** is a senior partner in the Bay Area office; **Peter Spiller** is a partner in the Frankfurt office; and **Tuisku Suomala** is a consultant in the Helsinki office.

The authors wish to thank McKinsey's Catalyst Zero and research and information teams, which include Neha Chatterjee, Nitin Shetty, and Witold Waliszewski, for their contributions to this article.

How semiconductor companies can fill the expanding talent gap

Companies will need to cast a wider net, improve their employee value proposition, and get more out of their existing workforce.

This article is a collaborative effort by Scott Brugmans, Ondrej Burkacky, Katrin Mayer-Haug, Andrea Pedroni, Giulietta Poltronieri, Taylor Roundtree, and Brooke Weddle, representing views from McKinsey's Semiconductors Practice.

The semiconductor industry is at the center of a high-stakes race amid a broad recognition that chips will be the engine for the next wave of growth and innovation. From South Korea to Germany to the United States, companies have announced plans for massive new factories. In all, close to $1 trillion in investment is expected from 2023 to 2030.[1] This frenzy of global expansion could reshape the industry and disperse the balance of power around the world.

Manufacturing capacity is just one part of the formula, however. Talent will be a critical part of the equation in this evolving industry. Companies must ensure they can attract and retain a sufficient pool of talent to ensure the new capacity under construction can operate at full steam when it starts production. We have noted previously the challenges semiconductor companies face in talent attraction and retention.[2] Yet too few companies and regions have done enough to address the industry's massive shortfall of qualified workers. The convergence of an insufficient number of graduates, an aging workforce, and an industry with a poor perception among candidates means these new capital projects could be delayed or unable to run at full capacity without urgent, coordinated action.

For semiconductor companies, prioritizing talent as a top strategic objective is no longer an option—it's a necessity. Business leaders can pursue a number of actions to make the most of the existing workforce, harness previously untapped pools of workers, and fill the remaining gaps with contingent labor.

Sizing the talent challenge in semiconductors

Even before the current wave of investment, industry demand for qualified candidates had grown by leaps and bounds. Job postings for semiconductor technical roles in the European Union and United States rose at a CAGR of more than 75 percent from 2018 to 2022.[3] If the semiconductor sector does not become more attractive, the resulting talent gap for engineers will be massive: more than 100,000 each in the United States and Europe and upward of 200,000 in Asia–Pacific (excluding China).[4] Major disparities exist among countries in Asia–Pacific: for instance, India is a potential net exporter of engineering talent, while other countries, such as Japan and South Korea, face severe shortages. And since the number of new graduates hasn't kept pace with job openings, the industry faces increasing demand for talent.

The talent challenge extends across the broader ecosystem of semiconductor value chain players. For example, companies designing and manufacturing the complex, capital-intensive equipment to produce chips face similar challenges in achieving growth and adding required capabilities. In turn, the (often midsize) companies supplying individual parts for these machines also struggle to fill the talent gap—since they are typically located outside talent hubs.

Our analysis identified the primary drivers of increased demand for technical talent at semiconductor companies.

The siting of new construction far from existing talent pools

Building new fabs requires the rapid onboarding of multiple roles, including in manufacturing (process engineers and technicians, area operators, and maintenance services), facilities, quality, and industrial engineering. Skilled construction workers (pipefitters, welders, electricians, and carpenters) are also needed.[5]

To date, each region has benefited from the concentration of talent close to existing semiconductor hubs—think Silicon Valley, Taiwan, and "Silicon Saxony" in Germany. New construction

[1] McKinsey analysis of data from Gartner and the Semiconductor Industry Association (SIA), 2023.
[2] "How semiconductor makers can turn a talent challenge into a competitive advantage," McKinsey, September 7, 2022.
[3] Based on McKinsey analysis of data from the McKinsey Org Analytics data platform.
[4] McKinsey Global Semiconductor Talent Model.
[5] "Strategies for building US semiconductor fabs: Finding skilled labor," McKinsey, February 7, 2023.

in other areas likely won't be so lucky; companies could face the daunting prospect of developing their own semiconductor ecosystems to serve as a magnet for talent. These ecosystems matter because highly skilled workers appreciate having multiple employment opportunities and connecting with similarly minded people. A well-developed ecosystem can also spur cross-pollination among companies, serving to disseminate teamwork practices, tools, and culture.

A shift in required skills

Silicon-based semiconductor chips have gotten progressively more powerful for decades, in line with Moore's Law.[6] More recently, the physical limitations of existing materials have sparked a quest for the next wave of leading-edge chips.

Research into new materials (such as silicon carbide and gallium nitride), advanced packaging, specialized ASIC (application-specific integrated circuit) applications, and the increased importance of embedded software have changed the talent profile for semiconductor companies.[7] Artificial intelligence and machine learning have replaced systems architecture as the most critical skills on the European job market in 2022, and the surge of generative AI could further amplify the importance of these skills (Exhibit 1).[8] Knowledge of applications and new materials has also become more relevant over the past few years.

Thanks to these changes, in 2022 the software engineer role (especially embedded software programming) replaced design engineer as the most critical occupation in the European semiconductor industry.[9]

Persistent attraction and attrition issues

Several patterns in the semiconductor industry have created recurring challenges in drawing top talent.

Demographics and the 'gray to green' transition. The industry is staring down a wave of impending retirements. One-third of semiconductor employees in the United States are aged 55 or older.[10] The European Union fares better, with one-fifth of the workforce in this demographic, but it also has a significant proportion of engineering and manufacturing employees close to retirement age.[11] According to the Association of Electrical and Digital Industry (ZVEI) in Germany and the Federation of German Industries (BDI), about one-third of the country's semiconductor workforce will retire in the next decade.[12] The shortfall of STEM graduates to replace these retirees could leave a yawning labor gap.

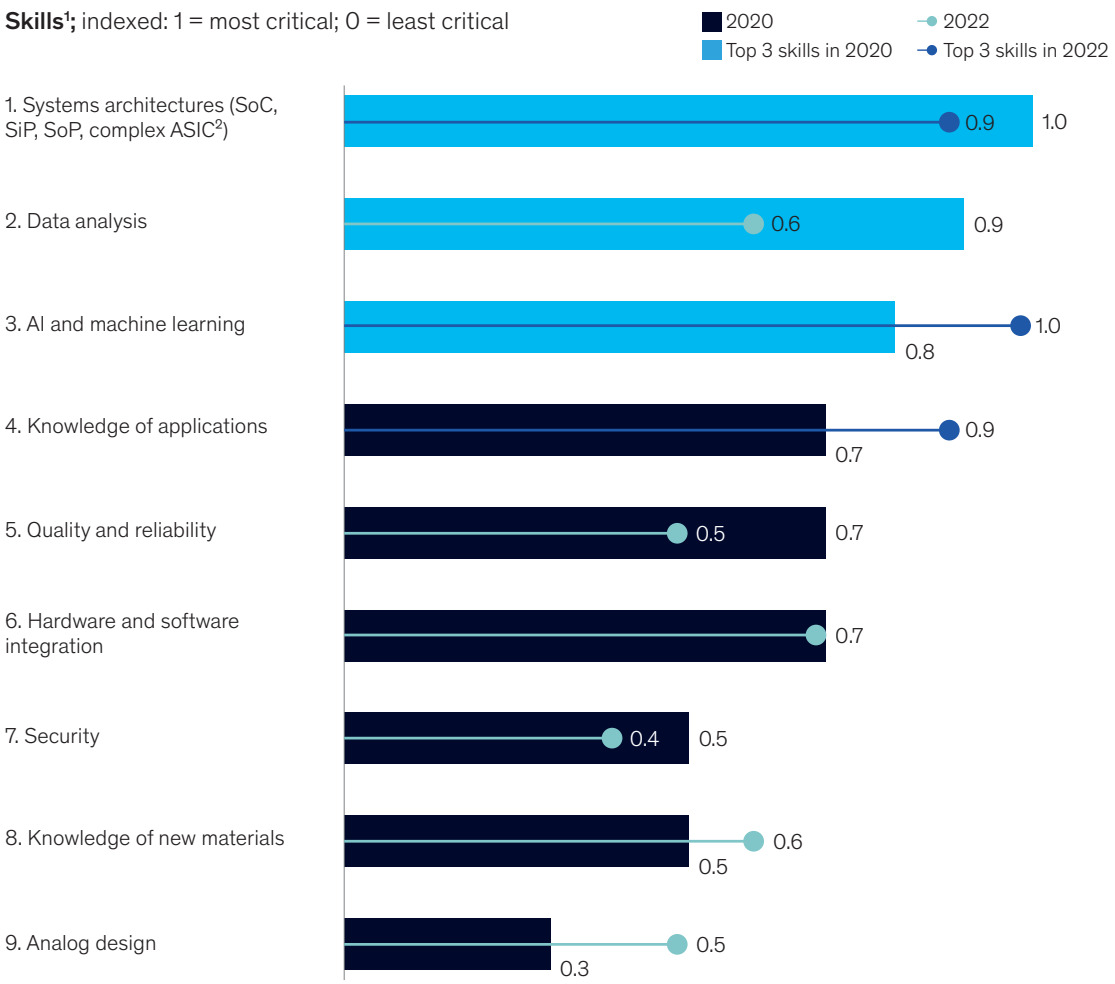
Subpar branding. The semiconductor industry faces a branding and marketing challenge in attracting technology talent. Surveys of both employers and college students indicate a lack of enthusiasm for semiconductor brands. Among senior executives, about 60 percent believe semiconductor companies have weak brand image and recognition compared with other, higher-profile tech companies. Meanwhile, students show more interest in working at consumer-oriented tech companies, which they believe can offer more-exciting jobs, higher compensation, and better development prospects.[13]

Employees with 'itchy feet.' An increasing number of employees in advanced electronics and semiconductors are at least somewhat likely to leave their current job in the next three to six months—53 percent this year versus 40 percent in 2021.[14] These employees cited an absence of career development and advancement (34 percent) and lack of workplace flexibility (33 percent) as the top reasons for looking for opportunities with another company (Exhibit 2). This is made worse by the fact that many of those who quit don't just quit a

[6] "How semiconductor makers," September 7, 2022.
[7] Léo Saint-Martin, *METIS skills strategy*, SEMI, November 18, 2021; Olivier Coulon, Jean-Charles de La Roncière, and Léo Saint-Martin, *Yearly monitoring report 2022*, SEMI, 2022; Ondrej Burkacky, Marc de Jong, and Julia Dragon, "Strategies to lead in the semiconductor world," McKinsey, April 15, 2022; "Cracking the complexity code in embedded systems development," McKinsey, March 25, 2022.
[8] *METIS skills strategy*, November 18, 2021.
[9] Ibid.
[10] McKinsey analysis of data from the Bureau of Labor Statistics, 2023.
[11] McKinsey analysis of data from Eurostat, 2023.
[12] Sabine Köhne-Finster, Susanne Seyda, and Dirk Werner, *Shortage of skilled workers in professions in the semiconductor industry*, Cologne Institute for Economic Research, March 7, 2023.
[13] "How semiconductor makers," September 7, 2022.
[14] McKinsey Great Attrition/Great Attraction Survey, March 2023, n = 667.

Exhibit 1

AI and machine-learning skills have recently replaced system architecture knowledge on the European job market.



[1]The skills most sought-after by companies and those most difficult to find on the European job market.
[2]System-on-a-chip, system in package, system on package, application-specific integrated circuit.
Source: Léo Saint-Martin, *METIS skills strategy*, SEMI, November 18, 2021; *Yearly monitoring report 2022*, SEMI, 2022; McKinsey analysis

McKinsey & Company

company but leave an industry altogether. Indeed, McKinsey's Great Attrition/Great Attraction Survey found that just 36 percent of respondents in industrials who had quit their jobs from April 2020 to April 2022 took another job in the same industry (compared with 45 percent in technology, media, and telecommunications).[15] Other respondents moved to a different industry or did not return to the workforce due to retirement.

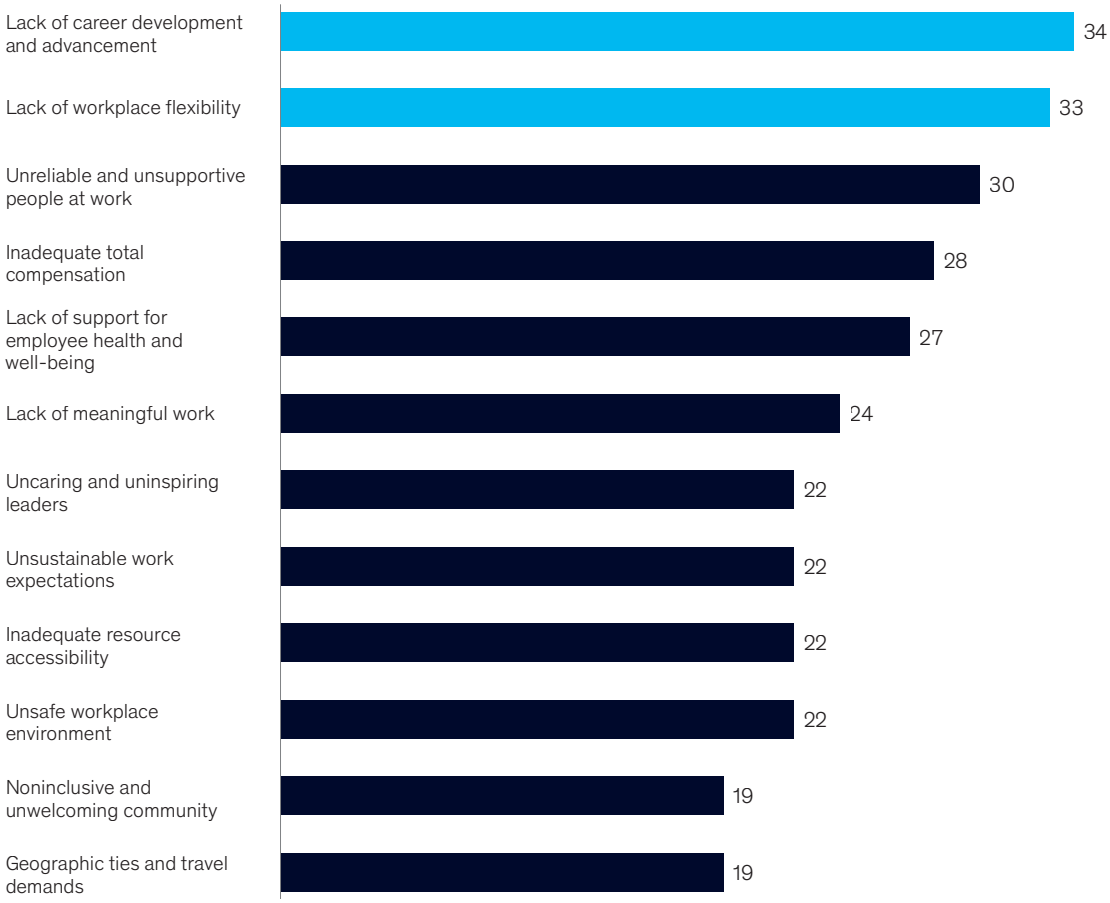
Another challenge for semiconductor companies is that employee satisfaction in the industry still lags behind that of tech and automotive players. The proximity of a company's business model to software

[15] Aaron De Smet, Bonnie Dowling, Bryan Hancock, and Bill Schaninger, "The Great Attrition is making hiring harder. Are you searching the right talent pools?," *McKinsey Quarterly*, July 13, 2022.

Exhibit 2

In advanced electronics and semiconductors, employees plan to leave jobs in search of career development opportunities and workplace flexibility.

Reasons behind current employees planning to leave, % of respondents[1]



[1]Represents the number of people who selected factors below as top 3 reasons they left their previous jobs.
Source: McKinsey Great Attrition Great Attraction Survey, n = 746

McKinsey & Company

is a contributing factor: for example, workers in foundry, materials, and outsourced semiconductor assembly and test (OSAT) positions report low scores, while those in intellectual property, electronic design automation (EDA), and fabless have the highest employee satisfaction scores.[16]

Key actions to attract and retain semiconductor talent

Despite the uphill battle for tech talent and the widening gaps between supply and demand, semiconductor companies can take several actions to reverse these trends.

[16] McKinsey Great Attrition/Great Attraction Survey, March 2023, n = 667.

Tackle reasons for current attrition

The past several years have created new expectations among employees for how, where, and when they work. Semiconductor companies that prioritize the fundamentals can ensure their workforce environment meets those expectations.

Reinforce a nontraditional career trajectory for advancement. Many companies take a traditional approach to career paths: when employees distinguish themselves with exemplary work, their reward is becoming a manager and taking on additional responsibilities for their team. It's critical for organizations to recognize that not all high-performing employees aspire to manage people—nor do they all have the people skills needed to excel in these roles.[17] In fact, two-thirds of developers have no ambition to become people managers.[18] To complement traditional people leadership career paths, companies should consider defining an expert path that allows individual contributors to rise through the company ranks. Equally important, companies should seek to be clear about the expectations of managers who lead teams. To ensure that these employees take their people leader tasks as seriously as they do driving content, companies should provide them with the time, training, skills, and tools to do so.

For example, a leading semiconductor company has defined three parallel career paths: management, technical (in which fellow is the highest role), and nontechnical support functions (such as finance, sales and marketing, and HR). This initiative boosted the motivation and overall retention of employees who are interested in career advancement but who want to continue as senior individual contributors. In addition, early talent identification and succession planning play an important role in improving career trajectories and enable companies to develop the future leaders they need to support their growth.

Give power to 'the middle.' Middle managers can find themselves mired in administrative tasks rather than focusing on the work that makes an organization run, such as nurturing talent. On average, just 28 percent of their time is focused on talent and people management.[19] Since they lack the necessary support and resources to manage their teams more effectively, they spend the majority of their time on individual contributor work.

Companies that restructure their organizations to free up middle managers can create "force multipliers," who make their direct reports much better. Actions to support this goal include optimizing team structures and reviewing roles to limit unnecessary layers and processes. Indeed, the top factor contributing to a middle manager's negative experience is organizational bureaucracy, cited by 44 percent of respondents.[20] Companies could also invest in enhancing the people skills of middle managers while improving their overall experience and ensuring they have the right degree of accountability and autonomy. For instance, a McKinsey survey found that providing middle managers with decision-making authority was the top factor in creating a positive environment for them.[21] Ideally, companies regularly review their operating models to ensure that decision-making authority lies in the most optimal position and that interfaces between departments are well defined.

One biotech start-up reviewed its organization and discovered that more than half of its managers had three or fewer direct reports. To optimize its structure, the company increased the number of employees under each manager by transitioning some people managers into expert roles that were better suited to their strengths. These shifts improved the efficiency of more than 200 teams with no reduction in head count.[22]

[17] "Cracking the code on digital talent," McKinsey, April 20, 2023.
[18] Sven Blumberg, Ranja Reda Kouba, Suman Thareja, and Anna Wiesinger, "Tech talent tectonics: Ten new realities for finding, keeping, and developing talent," McKinsey, April 14, 2022.
[19] "Stop wasting your most precious resource: Middle managers," McKinsey, March 10, 2023.
[20] Ibid.
[21] Emily Field, Bryan Hancock, Stephanie Smallets, and Brooke Weddle, "Investing in middle managers pays off—literally," McKinsey, June 26, 2023.
[22] Ibid.

Improve workplace flexibility. Despite the vast changes to workplace schedules and arrangements since the onset of the pandemic, most companies have only begun to scratch the surface for how to adapt to employee expectations. Technology can support a more strategic approach to on-site and remote work, but most organizations are still struggling to strike the right balance in creating true hybrid-work models.[23] Many have mastered basic capabilities, such as advanced workplace technologies that enable synchronous and asynchronous communication seamlessly from anywhere. When it comes to more dynamic practices, companies are missing opportunities to test new work arrangements (such as hybrid) and codify lessons learned, as well as to gauge hybrid versus full-remote experiences. Many still struggle with balancing an employee's desire for remote work with the risk of lower efficiency and a weaker connection to the company and its culture.

However, leaders of hybrid teams will also need to adapt their leadership methods and approaches to successfully lead their hybrid teams compared with fully on-site teams.

Identify and access untapped talent pools

Semiconductor companies could start to address skill gaps by considering several often-overlooked talent pools. For example, women account for only 17 percent of tech roles in the semiconductor industry, compared with 32 percent in social media and 23 percent in industrials.[24] McKinsey's Women in the Workplace research found that women leaders are significantly more likely than men leaders to leave their jobs in pursuit of more flexibility or to work for a company that is more committed to employee well-being and diversity, equity, and inclusion (Exhibit 3).

Our research suggests that, to become more appealing to women, companies could focus on providing work options. The most important factors when employees choose an employer are the opportunity to work remotely and to have greater control over both location and schedules, along with healthcare benefits (including mental-health benefits).[25]

Retired people who would like to work but aren't currently doing so—20 percent of respondents in a recent survey of high-income economies—could also help fill the gap.[26] Out-of-work older adults who are interested in securing a job cite barriers such as the lack of attractive opportunities, difficulties in landing a job, and societal barriers (such as mandatory retirement policies and cultural norms).[27] By defining different paths for older adults with previous experience in the sector, companies could create a fast track to help them reenter the workplace in areas in which they have special competencies (such as semiconductor R&D).

Adapting role requirements to focus on an individual's skills rather than their credentials (such as college degrees)[28] could also help companies find and attract a broader pool of candidates who are better suited to fill these positions in the long term.[29] For example, process engineers in fabs are responsible for process stability and recipe optimization and development, a role that requires knowledge of areas such as quality assurance and statistical control, continuous improvement of processes, and the ability to develop new processes, support new product introduction, and lead process-related customer meetings. In some cases, a former operator's shift leader could have these skills and perform the process engineer role despite not having a degree.

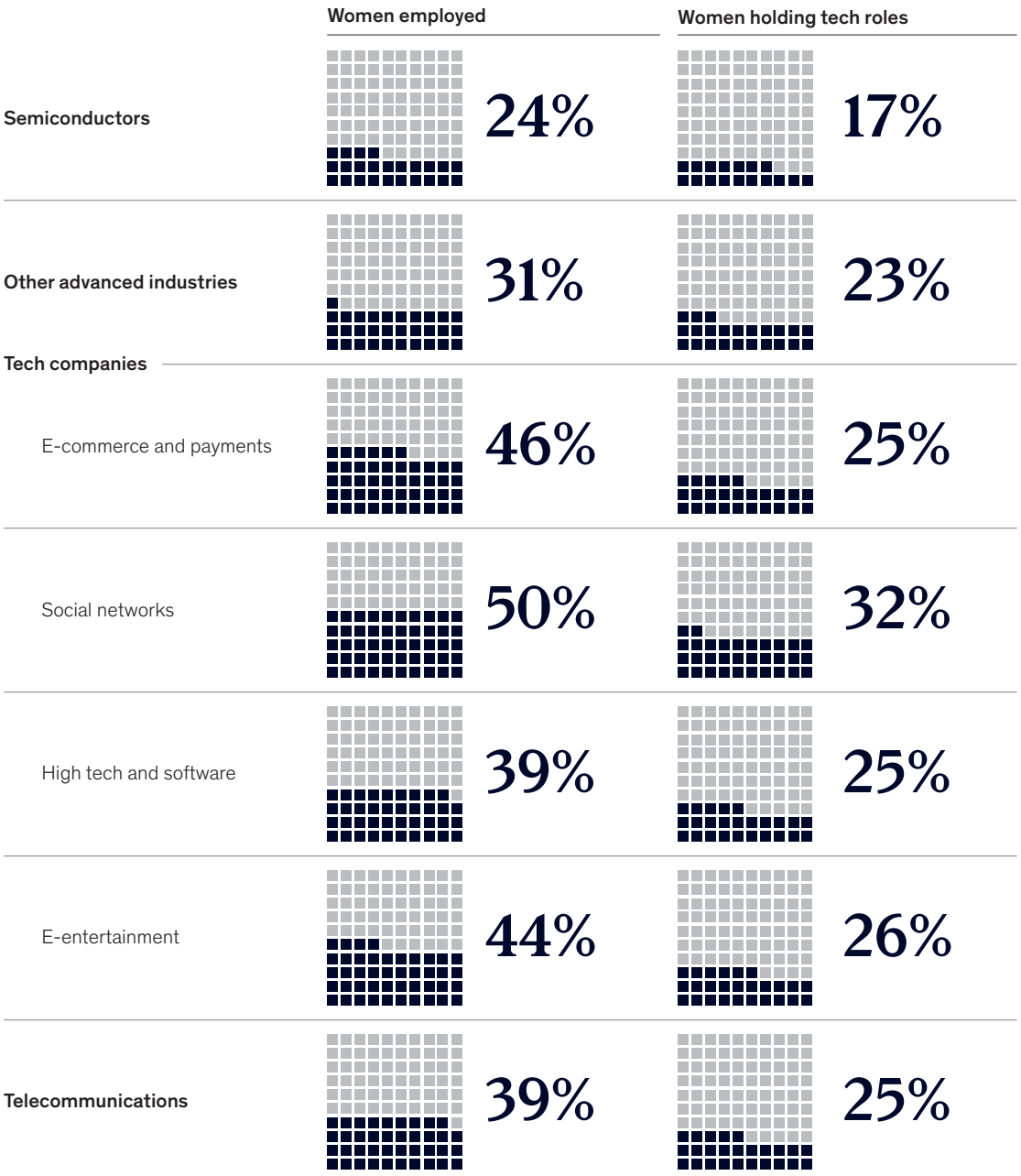
Generative AI could help to accelerate that shift through its capacity to tag abilities in unstructured data—essentially piecing together a candidate's skills based on descriptions of their experience in

[23] Phil Kirschner, Adrian Kwok, and Julia McClatchy, "Is your workplace ready for flexible work? A survey offers clues," McKinsey, June 1, 2023.
[24] Sven Blumberg, Melanie Krawina, Elina Mäkelä, and Henning Soller, "Women in tech: The best bet to solve Europe's talent shortage," McKinsey, January 24, 2023.
[25] "Women in the Workplace 2023," McKinsey, October 5, 2023.
[26] "Age is just a number: How older adults view healthy aging," McKinsey, May 22, 2023.
[27] Ibid.
[28] "Generative AI and the future of work in America," McKinsey Global Institute, July 26, 2023.
[29] "Taking a skills-based approach to building the future workforce," McKinsey, November 15, 2022.

Exhibit 3

Semiconductors lag behind advanced industries, tech companies, and telecommunications in gender balance in Europe.

% of employees, n = >1 million profiles



Source: Analysis of 2022 data from Eightfold AI by McKinsey and Eightfold AI

McKinsey & Company

previous roles.[30] These capabilities could expand talent pools to include workers in adjacent industrial sectors: for example, workers in clean-room manufacturing (such as chemicals and pharma) and heavy capital equipment (for example, military maintenance and power generation) have skills that are transferable to roles in a fab.

HR vendors are now integrating generative AI into talent acquisition. One global HR technology company uses these tools to generate contextually relevant job descriptions, highlight external and internal candidates who are a good fit, send personal emails, and provide succession planning for high-performing employees. These tools can also identify upskilling and reskilling opportunities and flag employees who could be flight risks.

In general, gen AI can help to elevate the HR function in semiconductor companies. This need is critical, as many companies are seeking to hire unprecedented numbers of new employees quickly, often in locations without existing semiconductor ecosystems.

Enhance storytelling related to semiconductors

The faster pace of technological innovation compels companies to ensure their workforce's skills and capabilities are keeping up. Organizations that make upskilling and continuous learning a part of their culture can gain a recruiting advantage. A focus on career development and well-being can be particularly attractive for a younger workforce. These elements can be promoted in recruitment pitches, storytelling, and online communities to reinforce an organization's commitment to its employees.

Beyond using development opportunities as a recruiting angle, the industry could also collaborate to improve the perception of semiconductors, starting with rebranding (for example, from semiconductors to micro- and nanoelectronics). Moreover, facilitating contacts between universities and semiconductor companies and research centers could increase student exposure to the industry and its career opportunities.

It is also critical to connect the often highly specialized job of each individual worker to the significant impact the company and semiconductors have on the world. Research has found that when employees find their work to be meaningful, their performance improves by 33 percent, they are 75 percent more committed to their organization, and they are 49 percent less likely to leave.[31]

Reimagine workforce productivity

Companies could invest in building the relevant skills internally by moving past traditional, generic programs to focus on tailored learning journeys. This approach to reskilling and upskilling could be summed up as "experiences and apprenticeships, not courses," crafted specifically for the necessary roles and job families (which organizations could identify as part of a workforce planning effort).[32]

Decreasing onboarding times and accelerating time to competence are critical levers to increase productivity, so these journeys need to start the moment a new employee walks through the door. Onboarding speed can be boosted through tech-enabled levers to enhance knowledge management and new-skill development. For example, large language models could enable organizations to "assetize" existing knowledge from today's workforce quickly and easily.

In addition, companies could harness generative AI to reduce skills requirements, accelerate skills development, or both. AI and the newest frontiers of generative AI have the potential to double the productivity of software developers, enabling them to complete coding tasks up to twice as fast.[33] More concretely, generative AI can expedite manual, repetitive work (such as autofilling standard functions and documenting code functionality), jump-start the first draft of new code, and

[30] Bryan Hancock, Bill Schaninger, and Lareina Yee, "Generative AI and the future of HR," McKinsey, June 5, 2023.
[31] *People & Organization Blog*, "Making work meaningful from the C-suite to the frontline," blog entry by Timothy Bromley, Taylor Lauricella, and Bill Schaninger, McKinsey, June 28, 2021.
[32] *Operations Blog*, "Ops 4.0—The Human Factor: The need for speed in building skills," blog entry by Markus Hammer, McKinsey, July 13, 2022.
[33] "Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.

The semiconductor industry is not alone in facing a talent shortfall, but its rapid expansion in the coming years creates a greater sense of urgency.

accelerate updates to existing ones. In addition, increasing simplicity and user-friendliness of platforms (such as low-code and no-code)[34] could reduce the need for additional software developers, since people without significant experience could still be effective at writing basic code.

As the industry further matures, cost will become more important, and increasing productivity is a key driver.

Draw on outsourced labor to manage shortages

Labor shortages are likely to persist—particularly in roles that may not be needed on a long-term basis. Examples include construction, equipment installation, and specialized maintenance. For such roles, companies can rely on outsourced labor services to address critical gaps. This approach is not as simple as filling individual roles, however. Both semiconductor companies and labor providers should be prepared to work together closely. Strategic cooperation and management can ensure the proper planning and allocation of outsourced resources, prevent double booking, and maintain schedules and productivity as planned.

Other industries provide a road map for talent outsourcing in the face of workforce challenges. For example, the healthcare industry adapted to dramatic pandemic-related labor shortages by accelerating the use of travel contract labor (for example, travel nurses). This shift was enabled by an existing, mature outsourcing model characterized by multiple staffing agencies with preexisting hospital relationships and rapidly scalable travel nurse sourcing and placement services. Hospitals applied this approach to solve other forecasted demand swings, such as an increase in patients during local holidays in snowbird destinations.

The IT industry also offers valuable lessons. Its demand for outsourcing has grown due to aging software infrastructure, an embrace of remote work, and increasing technology specialization (for example, AI-driven data analysis and integration projects). The transition from relying on internal core staff for all business functions to outsourcing noncore industry functions (such as low-skilled technicians, linen services, and security) has enabled organizations to concentrate their resources on filling critical positions.

The semiconductor industry is not alone in facing a talent shortfall, but its rapid expansion in the coming years creates a greater sense of urgency. The most successful organizations will not only expand their

[34] Harald Bauer, David Ebenstein, Giulietta Poltronieri, and Jan Paul Stein, "Is industrial automation headed for a tipping point?," McKinsey, June 16, 2023.

candidate pools by being more strategic and resourceful but also implement efforts to get more from their existing workforces.

The growth of the semiconductor market and new fab builds will boost demand for talent, requiring increasingly sophisticated approaches such as hiring at scale and in a short time frame—frequently in locations without existing semiconductor ecosystems. Indeed, the industry and governments appear to recognize the need to close the gaps that have been created in rapid succession following the passage of legislation in multiple countries to support semiconductor production. Numerous programs are under way to increase the supply of skilled construction craft laborers, semiconductor engineers, and technicians. Further evaluation will be needed to understand whether existing programs are on pace to fully close the emerging talent gap.

These recent trends reinforce that it is critical for organizations to take their people strategies as seriously as their business strategies. There's no time to waste.

Scott Brugmans is an associate partner in McKinsey's Amsterdam office; **Ondrej Burkacky** is a senior partner in the Munich office; **Katrin Mayer-Haug** is a partner in the Stuttgart office; **Andrea Pedroni** and **Giulietta Poltronieri** are partners in the Milan office; **Taylor Roundtree** is an associate partner in the Atlanta office; and **Brooke Weddle** is a partner in the Washington, DC, office.

March 2024
Designed by McKinsey Global Publishing

McKinsey.com