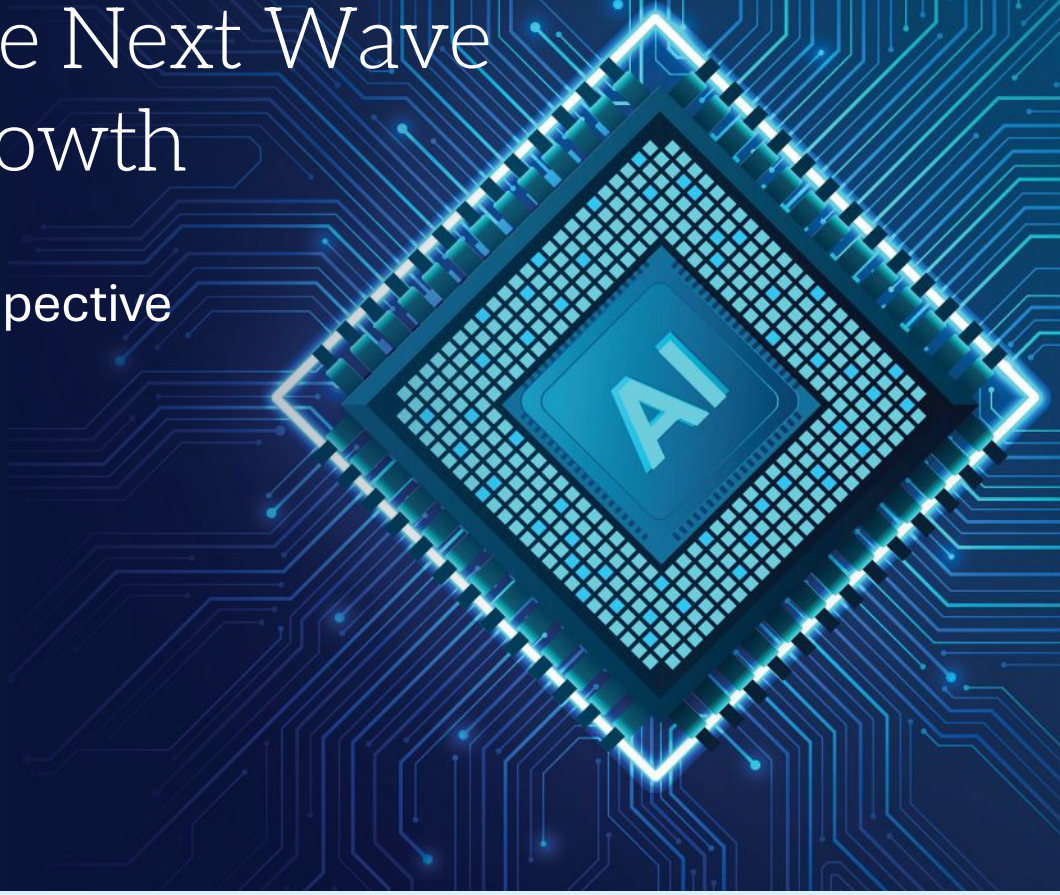
Omdia
by informa techtarget

AI Semiconductor Platforms: Powering the Next Wave of Global Growth

An Asia Pacific Perspective



In partnership with

HCLTech

Contents

Executive Summary	3
The AI Semiconductor Inflection Point	4
Taiwan’s Central Role in the Global AI Semiconductor Ecosystem	6
The Emerging AI Semiconductor Platform	9
Redefining Performance for the AI Era	10
From Hyperscale Data Centers to Distributed Intelligence	12
Building the AI Semiconductor Ecosystem	14
Strategic Priorities for Industry Leaders	16
Conclusions	18
Appendix	19

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

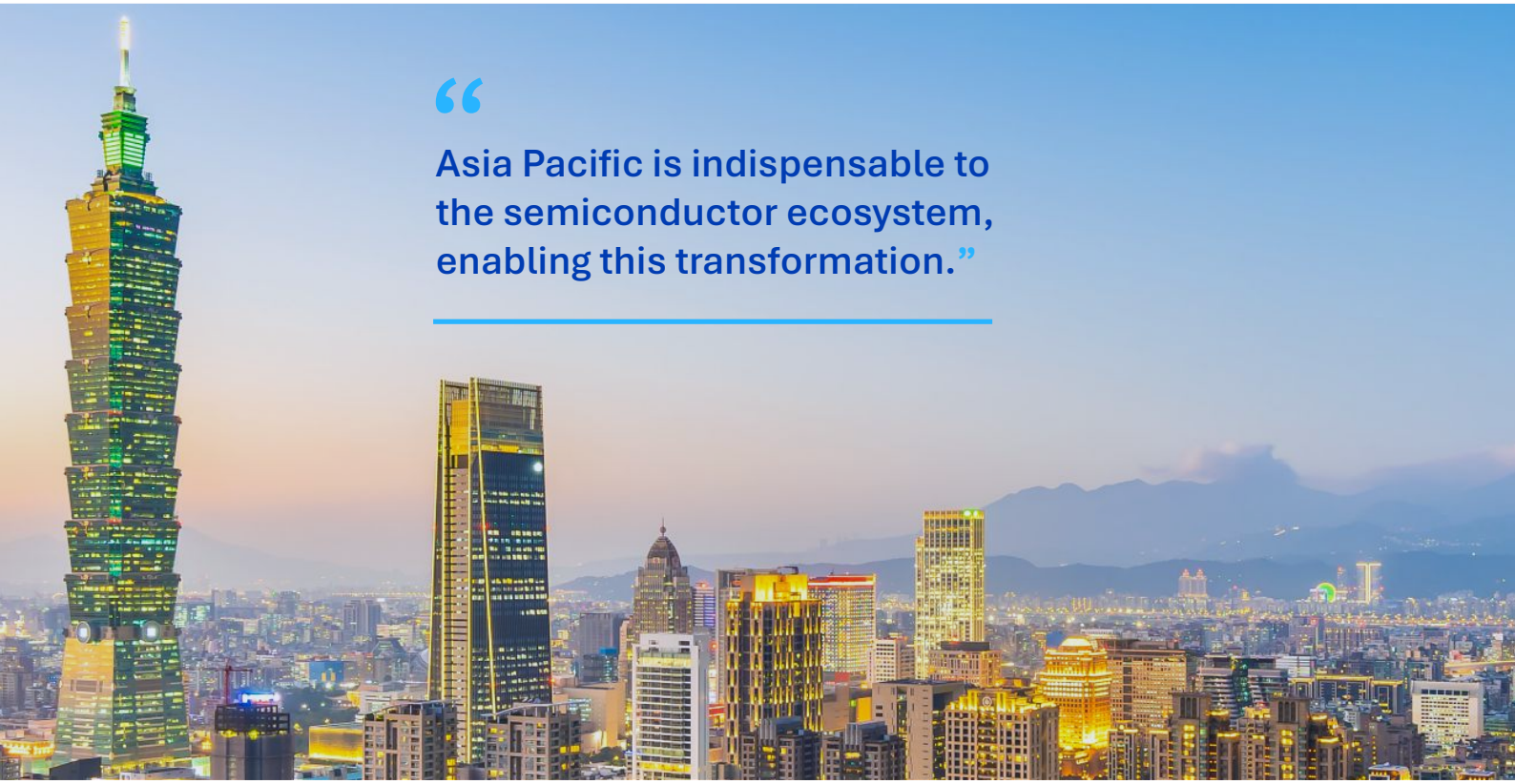
Executive Summary

Artificial intelligence (AI) is reshaping the semiconductor industry by increasing demand for advanced devices while changing how computing systems are designed, manufactured, and evaluated. Performance increasingly depends on coordination across the entire compute stack, giving rise to integrated **AI semiconductor platforms** optimized around workload-level performance rather than individual component specifications.

At the same time, AI deployment is expanding beyond hyperscale data centers into enterprise infrastructure, telecommunications networks, factories, vehicles, PCs, smartphones, and other edge devices. This broadens semiconductor demand across both leading-edge technologies and the specialty and mature-node technologies required for sensing, power management, control, and embedded intelligence.

Asia Pacific is indispensable to the semiconductor ecosystem, enabling this transformation. Taiwan combines leadership in leading-edge manufacturing and advanced packaging with specialty processes and large-scale AI system production, while Singapore, Malaysia, and Vietnam provide complementary capabilities across an increasingly distributed regional production network. China occupies a distinct position: it remains deeply connected to regional semiconductor demand and manufacturing while simultaneously building a more self-reliant technology ecosystem.

For industry leaders, the strategic implications will vary depending on where they compete. Companies may benefit from concentrating differentiation where it creates defensible value, while relying more heavily on partners where system complexity makes vertical optimization impractical. Anticipating capacity and technology bottlenecks can help preserve scalability, while greater manufacturing and geographic flexibility can improve resilience—provided the benefits justify the additional cost and complexity. Ultimately, semiconductor innovation creates sustainable advantage only when it can be translated into scalable business value.



© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

The AI Semiconductor Inflection Point

AI entered the public consciousness in late 2022, following OpenAI's release of ChatGPT. Since then, AI development and adoption have accelerated, moving from experimentation and proof-of-concept toward increasingly large-scale commercial deployment. This transition is reshaping the semiconductor industry—not only by increasing demand, but also by accelerating technology development and changing supply-chain relationships.

A Fundamental Shift in Semiconductor Demand

Conventional computing systems are designed to support a broad range of general-purpose applications. AI workloads, in contrast, require enormous numbers of mathematical operations to be carried out concurrently. This shift has created strong demand for highly parallel computing architectures and some of the most complex and valuable semiconductor devices produced today.

The scale of AI infrastructure magnifies this effect. Large AI data centers can require hundreds of megawatts of power and contain tens or hundreds of thousands of specialized chips operating as tightly coordinated computing clusters. Facilities now planned or under construction are larger still, with some planned campuses requiring gigawatt-scale power.

Generative AI (GenAI) has therefore introduced not simply another source of semiconductor demand, but demand of a different magnitude and concentration—placing unprecedented pressure on advanced semiconductor capacity and associated supply chains.

The Convergence of Enabling Technologies

The introduction of the transformer architecture in 2017 provided the foundation for today's large language models (LLMs), but GenAI's rapid development also depended on parallel advances in semiconductor technology. Leading-edge process nodes increased compute density, while specialized accelerators, high-bandwidth memory (HBM), advanced packaging, chiplets, and high-speed interconnects enabled increasingly large computing systems to operate as tightly integrated platforms.

The significance of these technologies lies increasingly in how they work together. Rather than relying primarily on improvements in any single device, modern AI systems achieve performance through the coordinated optimization of compute, memory bandwidth, data movement, packaging, connectivity, power, cooling, and software.

Demand Accelerates as AI Moves Beyond Experimentation

AI is now beginning to move beyond hyperscale infrastructure into a broader phase of industrial-scale deployment, increasingly driven by inference.

Businesses are increasingly deploying AI across a broad range of operational, technical, and customer-facing functions. As deployment expands, performance remains important, but so do cost, security, reliability, data governance, energy consumption, and integration with existing infrastructure.

Inference will consequently spread across a much wider range of environments—from enterprise infrastructure and telecommunications networks to factories, vehicles, PCs, smartphones, and intelligent edge devices. This broadening of AI deployment is expanding semiconductor demand beyond hyperscale computing into an increasingly diverse range of applications.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

From General-Purpose Computing to AI-Optimized Architectures

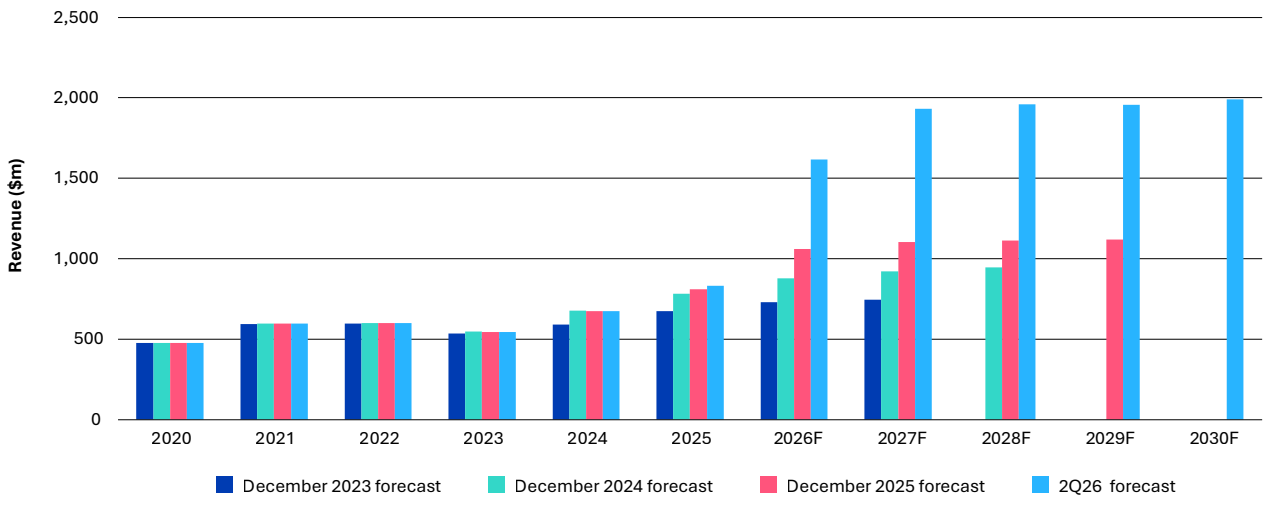
GPUs long predate the current AI era, but their highly parallel architecture made them particularly well-suited to neural-network computation. As AI workloads have expanded in scale and diversity, they have created demand for an increasingly broad range of workload-optimized devices, including GPUs, NPUs, custom AI accelerators, and other specialized architectures. The result is not simply a shift away from general-purpose CPUs, but an expansion of semiconductor content around new classes of computing workloads. The impact is visible in successive upward revisions to the semiconductor industry's long-term revenue outlook (**Figure 1**), reflecting not only greater semiconductor demand but also the exceptionally high semiconductor content and value of advanced AI systems.

Table 1. Selected Milestones in the Evolution of Foundation Models

Model name	Release year	Parameter size	Notes
BERT	2018	340 million	Early encoder-only local baseline
GPT-2	2019	1.5 billion	Easily fits on a modern smartphone today
GPT-3	2020	175 billion	Sparked the massive cloud data center boom
GPT-4	2023	1.8 trillion	Early Mixture-of-Experts (MoE) cloud model
Claude 3 Opus	2024	2 trillion	Heavy, closed-source cloud hybrid
Llama 3	2024	405 billion	Peak uncompressed open-weights baseline
GPT-5	2025	Proprietary (trillions)	OpenAI's shift to multi-tier "thinking" models
DeepSeek-R1	2025	671B (total)	Sparse MoE (37 billion active parameters)
GLM-5	2026	744B (total)	Zhipu's open MoE (40 billion active parameters)
Claude Fable 5	2026	Proprietary (multi-trillion)	Anthropic's frontier "Project Glasswing" model
GPT-5.6 Sol	2026	Proprietary (multi-trillion)	OpenAI's top-tier agentic reasoning model
Kimi K3	2026	2.8 trillion	Massive open-weights MoE

Source: Company disclosures; media reports; Omdia analysis

Figure 1: Successive Upward Revisions to the Global Semiconductor Revenue Forecast



Source: Omdia, Application Market Forecasting Tool (AMFT)

© 2026 Omdia

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Taiwan's Central Role in the Global AI Semiconductor Ecosystem

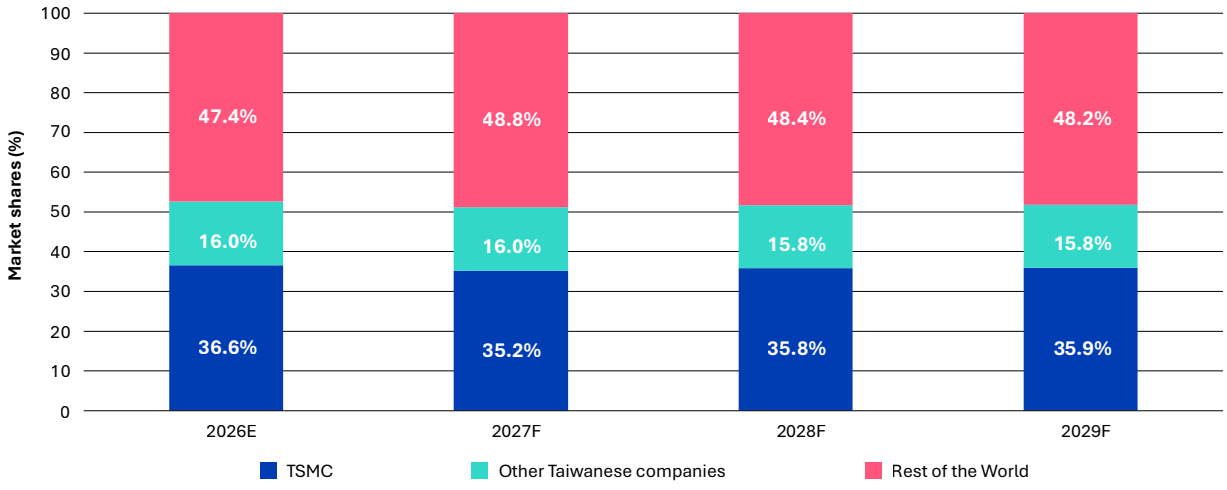
Taiwan occupies a uniquely important position in the global AI semiconductor ecosystem. Its significance begins with semiconductor manufacturing but extends across IC design, wafer fabrication, advanced packaging, substrates, printed circuit boards, power systems, thermal management, server design, rack integration, and final system assembly.

This breadth gives Taiwan a distinct advantage as AI computing becomes increasingly dependent on the integration and industrialization of complex systems rather than the production of individual semiconductor devices alone.

Taiwan's Semiconductor Manufacturing Dominance

Taiwan's most visible contribution is its leadership in semiconductor manufacturing. The foundry model supports a highly developed fabless semiconductor ecosystem by allowing chip companies to concentrate on design and product development without the enormous capital requirements of owning and operating advanced fabrication facilities. This model is particularly important to AI, where many leading accelerator and custom-silicon suppliers depend on external foundries.

Figure 2: Global Pure-Play Foundry Market Shares



Source: Omdia, Pure-Play Foundry Market Tracker

© 2026 Omdia

Collectively, Taiwan-headquartered foundries account for more than half of the global foundry-service market (**Figure 2**). Their position spans Taiwan Semiconductor Manufacturing Company (TSMC), United Microelectronics Corporation (UMC), Powerchip Semiconductor Manufacturing Corporation (PSMC), Vanguard International Semiconductor (VIS), Episil, and other specialist suppliers. At the leading edge, the concentration is even greater: TSMC manufactures more than 90% of the world's leading-edge semiconductors.

This leadership matters to AI because advanced accelerators and custom silicon rely heavily on leading-edge process technologies. Continued transistor scaling enables greater computing density and improved energy efficiency—critical advantages for data centers constrained by power, cooling, and physical space.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Taiwan's manufacturing strength, however, extends well beyond advanced logic. Its foundries support specialty technologies including BCD and high-voltage processes, analog and mixed-signal, RF, embedded nonvolatile memory, image sensors, MEMS, ultra-low-power technologies, and compound-semiconductor processes, such as GaN and SiC. These capabilities support many of the power-management, connectivity, interface, control, and sensing functions required throughout AI infrastructure.

Taiwan, therefore, possesses a dual manufacturing advantage: world-class process technology for advanced AI chips, complemented by a broad base of mature and specialty manufacturing for the devices surrounding them. This scale creates cumulative advantages in process expertise, yields, design enablement, supplier relationships, and customer trust that cannot be replicated simply by constructing new fabs.

Advanced Packaging Capabilities Critical for AI Chips

Advanced packaging has become nearly as important as wafer fabrication to AI semiconductor performance. TSMC's Chip-on-Wafer-on-Substrate (CoWoS) platform, introduced in 2012, allows large logic dies, HBM, and other components to be integrated within a tightly coupled package.

As AI systems have grown larger, CoWoS has evolved from a specialized packaging technology into a critical production platform supporting larger interposers, more HBM stacks, greater chiplet integration, and increasingly complex heterogeneous designs.

Taiwan's advanced-packaging strength extends beyond TSMC. ASE Technology Holding, the world's largest outsourced semiconductor assembly and test (OSAT) provider with roughly 40% market share, offers extensive assembly, testing, system-in-package, and advanced 2.5D and 3D integration capabilities. Taiwan also has a broad network of substrate, wafer-bumping, testing, printed-circuit-board, and materials suppliers.

For AI chip suppliers, advanced-packaging capacity can directly affect time-to-market and shipment volumes, making it a strategic manufacturing resource rather than a routine back-end process. Taiwan's ability to combine leading-edge fabrication with advanced packaging, assembly, and test, strengthens its position as AI architectures become increasingly modular and heterogeneous.

From Semiconductors to AI Systems

Taiwan's capabilities extend beyond semiconductors into the engineering, integration, and volume production of AI servers and rack-scale systems. Taiwanese ODMs—including Foxconn, Quanta, Wistron, Wiwynn, Inventec, and MiTAC—play leading roles in producing server boards, systems, and increasingly rack-scale AI infrastructure. Taiwanese ODMs account for approximately 90% of global server-motherboard production and approximately 80% of server-system production.

Their importance is increasing as AI architectures move toward rack-scale systems. NVIDIA's GB200 NVL72, for example, integrates 72 GPUs and 36 CPUs within a single liquid-cooled rack.

As AI systems combine large numbers of accelerators with high-speed networking, sophisticated power delivery, and liquid cooling, semiconductor, packaging, and system-engineering decisions become increasingly interdependent. Taiwan's ability to connect these disciplines and rapidly move complex designs into volume production extends its advantage from semiconductor manufacturing to the industrialization of complete AI systems.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Regional Dynamics: A More Distributed Asian Supply Chain

Taiwan's centrality does not mean that the AI semiconductor supply chain is contained within the island. The supply chain is distributed across Asia, with Singapore, Malaysia, Vietnam, and China contributing different but increasingly important capabilities. This diversification expands capacity and resilience without necessarily displacing Taiwan's role at the center of advanced semiconductor manufacturing and integration.

Singapore has Southeast Asia's most developed end-to-end semiconductor ecosystem, spanning wafer fabrication, memory production, assembly and testing, advanced packaging, equipment manufacturing, regional headquarters, and research institutions. Singapore accounts for approximately 10% of global semiconductor output, 5% of global wafer-fabrication capacity, and 20% of semiconductor equipment production. Recent investments include additional UMC foundry capacity, Micron memory and HBM packaging operations, and facilities supporting semiconductor equipment and advanced-packaging research. Its combination of advanced manufacturing, engineering talent, infrastructure, research capability, intellectual property (IP) protection, and logistics makes Singapore particularly suited to capital-intensive and highly automated activities.

Malaysia is a major center for semiconductor assembly, testing, and packaging, accounting for approximately 13% of such activities, with major clusters in Penang and Kulim. It hosts operations belonging to global semiconductor manufacturers, OSAT providers, equipment suppliers, and electronics manufacturers. Malaysia also has an established front-end wafer fabrication base, with facilities operated by several domestic and multinational semiconductor manufacturers. Since the early 2020s, momentum in front-end manufacturing has accelerated, particularly in power and compound semiconductors. Malaysia is also moving toward advanced packaging, semiconductor equipment, automation, and higher-value functions such as IC design. Investments by Intel, ASE, Infineon, Micron, and others are expanding its role beyond traditional labor-intensive back-end manufacturing.

Vietnam is at an earlier stage of semiconductor development, but is rapidly becoming an important provider of packaging, testing, and electronics manufacturing capacity. Amkor, for example, has committed $1.6bn to the first two phases of its Bac Ninh facility, designed for advanced system-in-package and memory products and potentially supporting 200,000 square meters of cleanroom space. Vietnam's large workforce, expanding electronics industry, and proximity to established Asian supply chains make it attractive for incremental capacity and diversification. Its near-term role is likely to remain concentrated in assembly, testing, electronics manufacturing, and supporting activities, while providing a foundation for more advanced capabilities over time.

China occupies a different position, combining an enormous electronics-manufacturing scale with a rapidly expanding domestic semiconductor industry, particularly in mature and selected advanced processes. Semiconductor-equipment spending in China reached $49.3bn in 2025, the highest of any region, reflecting continued investment in domestic manufacturing capacity. At the same time, trade restrictions and geopolitical considerations are encouraging multinational and Taiwanese manufacturers to diversify portions of production elsewhere in Asia.

The result is not the replacement of Taiwan by another manufacturing center, but a more distributed Asian production network. Singapore contributes high-value manufacturing and equipment capabilities; Malaysia provides scale in packaging and testing; Vietnam offers an emerging base for expansion; and China retains enormous semiconductor and electronics-manufacturing capacity. Taiwan remains the principal center connecting advanced manufacturing, integration, and the industrialization of AI infrastructure.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

The Emerging AI Semiconductor Platform

AI is changing not only the types of semiconductors being used but also how they must work together. AI workloads place unusually high demands on computation, memory bandwidth, data movement, and power, increasingly making system performance dependent on coordination among multiple technologies rather than the capabilities of any single device.

The prioritization of system-level performance over individual component specifications is giving rise to the AI semiconductor platform, in which components and subsystems are optimized as parts of an integrated system.

Building Blocks of the AI Platform

AI platforms combine several closely connected semiconductor technologies.

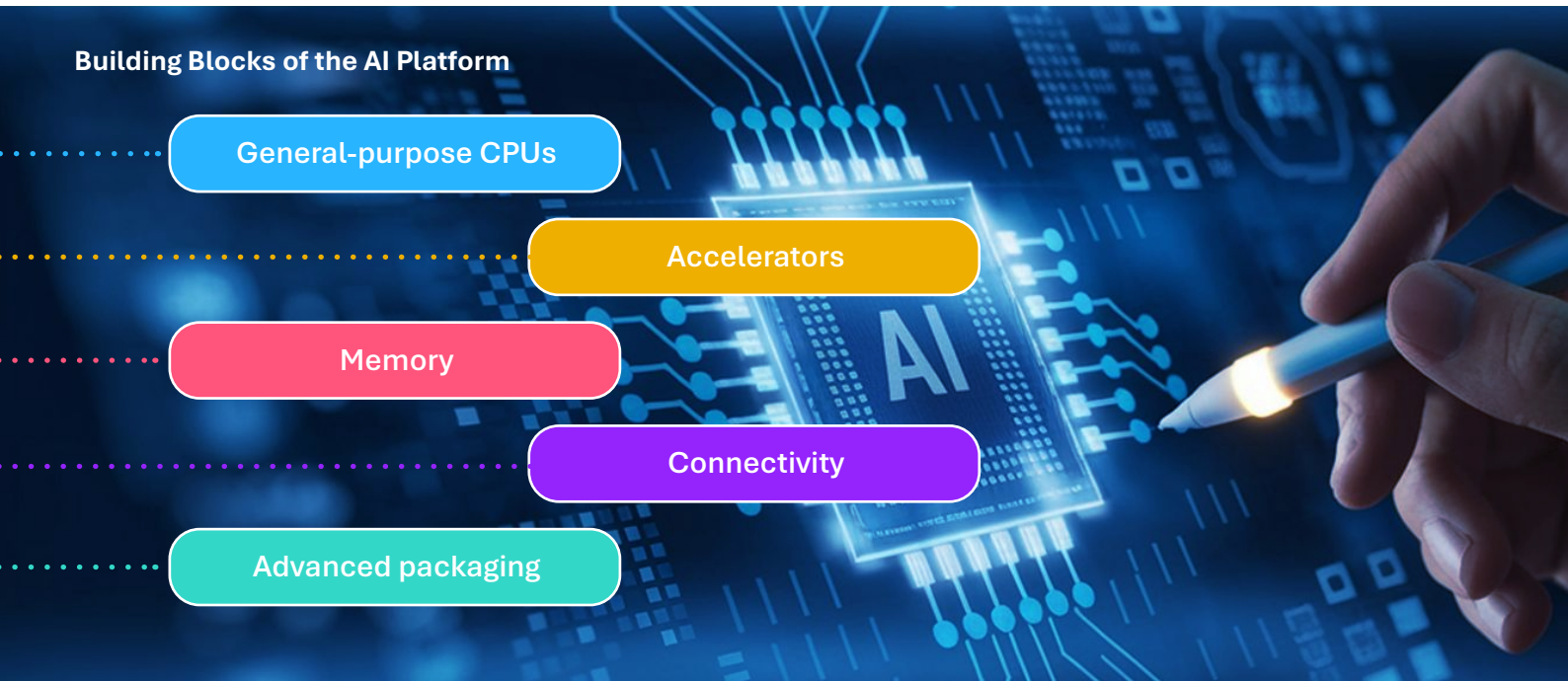
General-purpose CPUs continue to perform essential control and orchestration functions, including operating systems, data preparation, storage access, networking, security, and portions of AI applications not well suited to massively parallel acceleration. Their role is increasingly to coordinate specialized computing resources rather than serve as the sole center of computation.

Accelerators, including GPUs and purpose-built AI devices, perform the computationally intensive matrix and vector operations required for AI training and inference. Different architectures are emerging for large-scale training, cloud inference, enterprise systems, and power-constrained edge applications.

Memory has become a critical determinant of performance because large models require both substantial capacity and high bandwidth. HBM, placed close to accelerators through advanced packaging, supplies data at much higher rates than conventional memory configurations. More broadly, AI systems increasingly rely on memory hierarchies designed to keep frequently used data close to the computing resources that need it.

Connectivity links computing resources within packages, servers, racks, and large clusters. As workloads are distributed across more accelerators, the ability to move data efficiently among them becomes part of the computing architecture itself rather than simply a communications function.

Advanced packaging provides the physical foundation for integrating these technologies. Chiplets, high-density interconnects, and technologies such as TSMC's CoWoS allow logic and HBM to be placed much closer together while supporting larger and more complex semiconductor designs.



Integration and the Evolution Toward Platforms

The effectiveness of these building blocks depends on how well they are integrated. At the silicon and package levels, compute, memory, I/O, power delivery, and thermal requirements must increasingly be designed together. At the server and rack levels, accelerators must operate efficiently with networking, storage, power, cooling, and software. Improvements in one area can otherwise simply shift the bottleneck elsewhere in the system.

Software provides another essential layer of integration. Compilers, libraries, frameworks, schedulers, and management tools must translate AI models into workloads that use the underlying hardware efficiently. Hardware and software, therefore, increasingly evolve together as parts of the same platform.

This represents a significant evolution from earlier accelerated-computing systems. Initial implementations commonly paired conventional servers with discrete accelerator cards connected through standardized interfaces. Subsequent systems introduced dedicated accelerator boards, high-speed interconnects, locally attached HBM, advanced packaging, and chiplets. Leading AI systems increasingly extend platform design to the rack level, coordinating compute, networking, power, cooling, and software as a complete system.

The direction is therefore clear: AI infrastructure is moving from collections of individually optimized semiconductor components toward **integrated, modular platforms designed around the requirements of the workload as a whole.**

Redefining Performance for the AI Era

Peak floating-point operations per second (FLOPS) and trillions of operations per second (TOPS) remain widely used measures of computational capability, but they are increasingly poor proxies for the value of an AI platform. AI performance is an end-to-end system outcome shaped by compute, memory and data movement, connectivity, software, and workload characteristics. The limiting factor can change not only between training and inference, but also between phases of the same workload. AI performance is shifting from a measure of theoretical computation to a measure of useful work completed under real operating constraints. Greater peak compute does not necessarily mean faster model training, higher serving capacity, lower energy consumption, or lower cost.

Performance Measures Become Workload-Specific

Training performance is increasingly measured by time to reach a defined model-quality target. MLPerf Training adopts this approach because hardware throughput alone does not determine training time; scaling efficiency, software, data pipelines, numerical precision, and model convergence also affect the result.

Inference creates a different optimization problem. Commercial performance depends on the interaction between throughput and responsiveness—particularly token-generation rates and latency for GenAI—under defined workload and service requirements. Larger batches may improve hardware utilization and aggregate throughput while simultaneously increasing individual response times.

GenAI further illustrates why headline compute specifications can mislead. Prompt processing can be relatively compute-intensive, while token-by-token generation can become increasingly constrained by memory bandwidth, capacity, and data movement.

Measures of real-world AI platform performance, therefore, include:

- **Throughput:** queries, samples, or tokens processed over time
- **Latency:** response time under defined workload conditions
- **Time to first token/token generation rate:** responsiveness of GenAI
- **Serving capacity:** workload sustained within defined latency targets

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

- **Energy per output:** joules per inference or token
- **Cost per output:** cost per query or token

The competitive question increasingly becomes not which platform provides the most theoretical compute, but which architecture delivers the required workload performance with the least resources.

Power and Economics Reshape the Definition of Performance

Performance per watt has moved from a secondary engineering measure to a strategic metric of AI infrastructure productivity. In power-constrained data centers, additional computational capability has limited value if its power and cooling requirements prevent equivalent capacity from being deployed elsewhere. Therefore, the more meaningful measure is useful workload output per unit of energy—not theoretical TOPS per watt.

The same logic applies to the total cost of ownership (TCO). Acquisition price and benchmark leadership matter less than the cost of producing the required AI outcome. Higher utilization, lower energy consumption, stronger scaling, reduced supporting infrastructure, or better software can allow a nominally slower device to deliver superior economics.

For training, the relevant measure can be expressed as the cost of reaching the required model quality. For inference, it increasingly becomes cost per query or token at the required quality and latency. Semiconductor performance and semiconductor economics are therefore becoming progressively harder to separate.

Research into data center TCO for LLMs further reinforces this point: workload-specific performance can matter more than theoretical peak throughput. Performance on the operations dominating autoregressive decoding, for example, may have a greater effect on inference economics than headline FLOPS.

Why Traditional Benchmarks Miss AI Semiconductor Value

AI platforms are increasingly optimized around the characteristics of specific workloads, whereas traditional benchmarks are designed to compare general-purpose computational performance. Performance is therefore more meaningfully measured by the rate and efficiency with which the platform completes useful work than by its theoretical computational capability.

AI workloads vary substantially in their demands on compute, memory, communication, latency, and power. In addition, software determines how much a platform's theoretical capability can actually be realized. Compilers, numerical precision, parallelization, scheduling, and model optimization can materially affect performance on identical hardware.

Therefore, even standardized AI benchmarks provide only part of the picture. **MLPerf** itself avoids creating a single composite training score because weighting dissimilar workloads would implicitly favor particular applications.



© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Production environments add further variation through changing workloads, request patterns, system utilization, and service-level constraints.

Therefore, while benchmarks remain useful reference points, they cannot by themselves measure how effectively a platform performs useful AI work under real operating conditions.

From FLOPS to a Performance Scorecard

No single metric will replace FLOPS because there is no single optimization objective for AI. The appropriate measure depends on the workload and its binding constraint: time-to-train when development speed matters; throughput and latency for serving; energy per output when power is constrained; and cost per output when economics dominate.

The emerging performance model is therefore a workload-specific scorecard rather than a universal ranking. Competitive advantage will increasingly belong to platforms that optimize the combination of performance, energy, cost, and quality that matters for the target workload—not those that simply maximize theoretical compute.

From Hyperscale Data Centers to Distributed Intelligence

The first phase of modern AI was concentrated in hyperscale data centers, where enormous pools of accelerators, memory, networking, storage, and power could be assembled economically. Centralized hyperscale infrastructure will remain essential. AI deployment, however, is entering a second phase, in which inference increasingly takes place across enterprise servers, telecommunications networks, factories, vehicles, personal computers, smartphones, cameras, and other connected devices.

The result is a more **distributed AI architecture** spanning hyperscale data centers, enterprise infrastructure, and edge systems in which cloud, edge, and endpoint systems perform different parts of the AI workload. Large data centers will continue to train foundation models and serve the most computationally demanding requests, while smaller models and latency-sensitive functions increasingly execute nearer to where data is created.

This transformation expands the semiconductor opportunity while changing its composition. The technology required to train a frontier model differs measurably from that needed to recognize a voice command on a battery-powered device or make a safety-critical decision in a vehicle.

AI Requirements Diverge Across Deployment Environments

The distributed AI architecture can be understood as a hierarchy. Central AI factories and hyperscale clouds provide the scale needed to train and run large AI models and the centralized control needed to coordinate distributed AI. Regional clouds and enterprise data centers provide intermediate capacity closer to users and corporate data. Local edge servers aggregate and process data within factories, stores, hospitals, telecommunications networks, and other operating environments. Endpoint systems interact directly with the physical world.

Several operating models will coexist. Some applications will execute primarily on-device and call the cloud only when additional capability is required. Others will perform sensing and preprocessing locally while sending compact representations or selected data to larger models elsewhere. More advanced implementations may divide workloads dynamically across tiers according to latency, energy, privacy, connectivity, model availability, and cost.

The location of AI execution, therefore, becomes an architectural decision rather than a fixed property of the application.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.



Implications for Semiconductor Design

Distributed AI creates several important design consequences.

First, **heterogeneous compute becomes increasingly important**. No single architecture efficiently handles application logic, neural inference, signal processing, graphics, communications, security, and real-time control. Systems will therefore combine general-purpose cores with GPUs, NPUs, DSPs, MCUs, and other specialized accelerators in configurations appropriate to the workload.

Second, **memory architecture becomes deployment-specific**. Hyperscale systems require very high memory bandwidth; enterprise systems emphasize capacity and predictable economics; consumer devices favor power-efficient shared memory; embedded systems may depend heavily on on-chip SRAM and embedded nonvolatile memory. The optimum architecture depends on model size, data movement, latency, and power.

Third, **connectivity becomes part of the compute architecture.** Within data centers, this includes accelerator fabrics and high-speed networking; at the edge, it can include 5G, Wi-Fi, automotive Ethernet, time-sensitive networking, and industrial communications. Distributed AI performance depends in part on how effectively computation and communication are partitioned.

Fourth, **security requirements extend to every tier.** More devices will contain proprietary models, sensitive data, and network credentials, increasing the importance of secure boot, roots of trust, isolated execution environments, authentication, protected memory, and secure updates.

Finally, **software portability becomes a competitive advantage.** Hardware suppliers increasingly need common development paths for model conversion, quantization, compilation, runtime scheduling, deployment, and monitoring across multiple hardware classes.



Implications for Manufacturing and the Supply Chain

Distributed AI also broadens semiconductor demand across the manufacturing spectrum.

At the highest end, foundation model training will continue to consume leading-edge wafers, HBM, advanced substrates, interposers, chiplets, and complex packaging. Bottlenecks can therefore emerge not only in wafer capacity, but also in memory, packaging, substrates, testing, or thermal integration.

Edge and endpoint AI create a different mix. Premium smartphones, PCs, and automotive systems may use advanced logic nodes, but distributed systems also depend extensively on specialty and mature-node technologies for sensing, analog interfaces, communications, embedded memory, control, and power management.

Distributed AI will therefore reinforce a **two-track manufacturing model:**

- Leading-edge logic, HBM, and advanced packaging will support the most computationally intensive systems.
- Specialty and mature-node technologies will connect those systems to communications, sensors, power supplies, actuators, and the physical world.

Heterogeneous integration can bridge these two tracks. Chiplet architectures allow functions to be manufactured on processes suited to their individual requirements and then combined within a package. Standards such as UCIe and foundry integration platforms are intended to make such architectures more modular and interoperable.

The manufacturing economics of distributed AI will nevertheless be complex. Edge markets encompass numerous industries, performance levels, qualification requirements, power envelopes, and product lifecycles. Suppliers will therefore need reusable IP, configurable architectures, modular designs, and common software environments to serve this diversity economically.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

A Broader Semiconductor Opportunity

The spread of distributed intelligence does not diminish the importance of hyperscale data centers. Broader AI adoption may, in fact, increase centralized demand as more endpoints call upon cloud models, generate data, and require fleet-wide coordination. What changes is the architecture surrounding the data center. AI becomes a continuum extending from the largest computing clusters to the smallest intelligent devices, with each tier creating a different semiconductor value proposition.

Distributed intelligence, therefore, represents more than the geographic dispersion of inference. It is the diversification of AI semiconductor design itself.

Building the AI Semiconductor Ecosystem

AI is changing not only semiconductor technology but also the structure of the industry that supports it. As AI performance becomes increasingly dependent on system-level optimization, the traditional semiconductor value chain is evolving from a sequence of supplier relationships toward a network of technology partnerships organized around complete platforms.

AI Demand is Reshaping Industry Structure

Cloud providers are increasingly becoming semiconductor architects. For instance, Google has developed successive generations of TPUs, AWS has its Trainium family, and Microsoft has its Maia accelerators. These programs allow hyperscalers to optimize silicon around their own workloads, infrastructure, and economics rather than relying exclusively on merchant devices. The distinction between hardware customers and semiconductor developers is therefore blurring.

Google describes Ironwood as part of its broader AI Hypercomputer architecture, while Microsoft says Maia 200 was developed alongside networking and system software before final silicon was available. In both cases, system requirements increasingly influence semiconductor architecture from the workload downward.

Semiconductor companies are moving in the opposite direction—from devices toward complete computing platforms. NVIDIA, for example, increasingly develops silicon, networking, rack-scale systems, and software as an integrated architecture rather than as independent products.

Foundries are also extending their roles into advanced packaging, 3D integration, design enablement, and capacity coordination. TSMC has said that capacity planning increasingly involves not only direct customers but, in some cases, their customers' customers.

The semiconductor industry is therefore not becoming less specialized. Rather, its specialization is becoming **more tightly orchestrated.**

From Supply Relationships to Co-Development

Three forms of collaboration are becoming particularly important.

Workload-to-silicon co-design brings cloud providers, AI developers, and semiconductor companies closer together. Information about model behavior, memory demand, utilization, and inference economics can shape chip architecture, while knowledge of future hardware capabilities can influence software and model development.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Silicon-to-package co-design reflects the increasing importance of HBM, chiplets, and advanced packaging. Memory bandwidth, interconnect density, power delivery, and thermal performance increasingly influence architecture itself. TSMC's 3DFabric Alliance, spanning design tools, IP, memory, packaging, substrates, and test, illustrates the breadth of coordination required.

Memory suppliers are undergoing a similar transition. SK Hynix has argued that AI is pushing memory companies beyond the role of component supplier toward technology partner, requiring a greater understanding of customers' accelerators, systems, power envelopes, and thermal requirements.

The third form is **rack-scale system integration.** At AI infrastructure densities, accelerators must be engineered together with networking, power conversion, cooling, mechanical design, firmware, and system management. This elevates ODMs and system integrators from contract manufacturers toward partners responsible for turning semiconductor platforms into manufacturable infrastructure.

Broadcom's relationships with Google and Meta illustrate this shift toward co-development. In 2026, Broadcom and Google entered a long-term agreement to develop and supply custom TPUs for future Google generations, while Broadcom and Meta expanded their partnership across multiple generations of MTIA accelerators, encompassing chip design, packaging, and networking. Together, these relationships show how hyperscalers and semiconductor suppliers are increasingly shaping silicon and system architectures jointly rather than operating through conventional customer-supplier relationships.

Partnerships Become a Scaling Requirement

Lastly, these changes make partnerships important not just for sharing development costs, but for preventing system-level bottlenecks and accelerating deployment. The competitive advantage, therefore, shifts partly from owning the best individual technology to **orchestrating the ecosystem needed to deploy it.** Chip development increasingly extends beyond the chip itself to the package, system, rack, network, software environment, and manufacturing base. The ability to coordinate these layers is becoming a prerequisite for scaling AI infrastructure.



© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Strategic Priorities for Industry Leaders

The AI semiconductor transition is changing the basis of competition. Technology leadership remains essential, but long-term value creation will depend on deciding where to differentiate, where to partner, where to invest ahead of emerging constraints, and how to convert innovation into scalable production.

Differentiate Where It Matters

For semiconductor companies, the first strategic decision is where proprietary technology can create a durable advantage. Not every layer of an AI platform needs to be owned. Differentiation may come from compute architecture, memory, connectivity, packaging, software, or specialized workload expertise; complementary capabilities can be supplied through partners and industry standards.

OEMs face the corresponding decision at the system level: how to combine available technologies around the requirements of target workloads rather than headline component specifications.

Enterprise buyers should likewise evaluate infrastructure according to application performance, utilization, energy consumption, scalability, software support, and TCO rather than peak compute alone.

The objective is not maximum vertical integration, but control of the capabilities that contribute most directly to customer value while avoiding unnecessary duplication elsewhere.

Invest Ahead of the Next Constraint

As AI architectures become more interdependent, the location of the next bottleneck becomes strategically important. Capital allocation should therefore anticipate where capacity, bandwidth, power, packaging, or manufacturing limitations could constrain the roadmap—not simply follow the fastest-growing technology category.

Modular architectures can provide additional flexibility. Chiplets, reusable IP, and standardized interfaces can allow selected parts of a platform to evolve without requiring the entire system to be redesigned. UCIe, for example, is intended to support interoperable die-to-die connectivity and multi-vendor chiplet ecosystems.

Investment discipline nevertheless remains essential. Semiconductor capacity is capital-intensive and slow to expand, while technology and demand can change rapidly. Leaders must balance the cost of insufficient capacity against the risk of committing too aggressively to technologies whose economics may shift.

Treat Manufacturing and Partnership Decisions as Product Decisions

Partnership and manufacturing strategies should be established at the beginning of the product roadmap, not after the architecture has been defined.

Companies should evaluate partners according to their contribution to differentiation, roadmap execution, capacity assurance, qualification, manufacturability, and time-to-market. Product architecture should likewise reflect realistic manufacturing constraints, including capacity availability, yield, packaging readiness, and supply-chain qualifications.

Technical capability alone, however, is not sufficient. Quality of execution, clear communication, transparent progress reporting, and timely resolution of issues are also important to successful execution and can be critical factors in building client confidence and long-term partnerships.

The competitive advantage is not simply the ability to innovate quickly, but the ability to repeatedly translate innovation into high-yield, high-volume production.

Build Regional Optionality

For Asia-Pacific companies, manufacturing strategy must also account for a more geographically distributed semiconductor industry. The objective should not be to reproduce every capability in every

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

market. Semiconductor ecosystems derive much of their strength from concentrations of specialized suppliers, infrastructure, engineering talent, and manufacturing expertise that are difficult to replicate quickly.

A more practical strategy is to preserve access to established centers of specialization while developing geographic and sourcing alternatives for strategically important capabilities.

Geopolitical and regulatory considerations must also become part of technology planning. Export controls governing advanced computing products and semiconductor manufacturing technologies continue to affect product configuration, customer access, and supply-chain decisions. Therefore, companies need sufficient architectural, sourcing, and manufacturing flexibility to respond as market-access conditions change.

For Asia-Pacific players, resilience should mean optionality rather than duplication.

From Technology Leadership to Execution Leadership

The AI era is broadening the definition of semiconductor leadership. Superior technology remains necessary, but it increasingly must be accompanied by disciplined capital allocation, ecosystem coordination, manufacturing scalability, and supply-chain flexibility.

The leaders will be those that manage these decisions as a single strategy—differentiating selectively, investing ahead of constraints, partnering deliberately, and designing scale into innovation from the beginning.

“The leaders will be those that manage these decisions as a single strategy—differentiating selectively, investing ahead of constraints, partnering deliberately, and designing scale into innovation from the beginning.”



© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

From AI Capability to Business Value

Advances in AI semiconductors do not automatically translate into business value. The value of an AI platform ultimately depends on how effectively its computational capability is converted into a useful outcome at an acceptable cost, latency, energy, and reliability. This makes workload fit increasingly important: the most powerful platform is not necessarily the most valuable if its architecture, software, or deployment model is poorly matched to the application.

Sustained value, therefore, depends on more than semiconductor performance alone. Utilization, energy efficiency, software maturity, system integration, and the ability to scale economically all influence the return generated from AI infrastructure investments. For technology providers and enterprise users alike, the objective is increasingly to align AI capabilities with the workloads and operating constraints that produce measurable business outcomes.

Conclusions

AI is broadening both the semiconductor opportunity and the definition of industry leadership. Success will increasingly depend on optimizing complete platforms, matching architectures to increasingly diverse workloads, coordinating deeper technology ecosystems, and scaling innovation efficiently.

Asia Pacific—and Taiwan in particular—will remain central to this transition because of the region's concentration of semiconductor manufacturing, packaging, systems integration, and increasingly distributed production capabilities.

The companies best positioned for the next phase of AI will therefore be those that connect technology leadership with ecosystem coordination, manufacturing scale, strategic flexibility, and measurable business value.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Appendix

Methodology

This whitepaper draws on Omdia's semiconductor market research and industry expertise, supplemented by company disclosures, technical publications, industry reports, academic research, and other publicly available sources. Findings were developed through comparative analysis of market, technology, and ecosystem trends, with key data points and conclusions cross-checked across multiple sources where possible. The analysis emphasizes structural developments in AI semiconductor platforms, manufacturing, performance, deployment, and regional ecosystems rather than short-term market fluctuations.

Further Reading

The following selected sources provide additional perspectives on the technologies, market developments, and strategic considerations shaping the AI semiconductor ecosystem.

Regional Semiconductor Landscape and Manufacturing

1. Taiwan Semiconductor Manufacturing Company (TSMC), *2025 Annual Report*.

   Provides a broad perspective on semiconductor technology, capacity investment, advanced manufacturing, advanced packaging, geographic expansion, and the evolving foundry ecosystem.

2. Singapore Economic Development Board, "Singapore's Semiconductor Industry."

   Provides an overview of Singapore's semiconductor ecosystem and its role in wafer fabrication, packaging and testing, equipment manufacturing, R&D, and heterogeneous integration within the Asia-Pacific semiconductor supply chain.

3. SEMI, "SEMI Projects Double-Digit Growth in Global 300mm Fab Equipment Spending for 2026 and 2027," April 1, 2026.

   Provides industry perspective on semiconductor manufacturing investment and capacity expansion driven by AI, advanced nodes, memory demand, supply-chain restructuring, and regional manufacturing initiatives.

AI Semiconductor Platforms and System Architecture

4. NVIDIA, "NVIDIA Blackwell Architecture."

   Illustrates the evolution from individual AI accelerators toward tightly integrated AI computing platforms encompassing compute, memory, interconnect, networking, and system-level design.

5. TSMC, "CoWoS Advanced Packaging Technology."

   Explains the role of advanced packaging and heterogeneous integration in bringing together high-performance logic and high-bandwidth memory for AI and high-performance computing systems.

6. Google, "Ironwood: The First Google TPU for the Age of Inference."

   Provides an example of workload-driven custom silicon and system co-design as AI infrastructure increasingly shifts toward large-scale inference.

7. Norman P. Jouppi et al., "TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings," ISCA 2023.

   Examines Google's TPU v4 system architecture and provides a technical example of how accelerators, interconnects, networking, and system-level design are co-optimized for large-scale machine-learning workloads.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

8. UCIe Consortium, "UCIe Specifications."

   Describes the open die-to-die interconnect standard supporting interoperable chiplets and illustrates the growing importance of standards and ecosystem collaboration as semiconductor systems become more heterogeneous.

9. Compute Express Link Consortium, "About CXL."

   Explains the open interconnect standard enabling coherent connectivity among processors, accelerators, memory, and other system components, illustrating the growing importance of system-level integration in AI computing platforms.

Redefining AI Performance

10. Amir Gholami et al., "AI and Memory Wall," 2024.

    Examines why improvements in peak compute alone do not determine AI performance and how memory capacity, bandwidth, and data movement increasingly constrain modern AI workloads.

11. MLCommons, "MLPerf Benchmarks."

    Provides industry-standard approaches to evaluating real-world AI training and inference performance, including workload completion time, throughput, latency, model quality, and energy efficiency.

12. Yuhui Zhang et al., "Metron: Holistic Performance Evaluation Framework for LLM Inference Systems," July 2024.

    Examines the limitations of individual inference metrics and presents a more holistic approach to evaluating LLM serving performance across latency, throughput, and workload characteristics.

Distributed AI and Infrastructure

13. Qualcomm, *The Future of AI Is Hybrid*.

    Explores the emerging distribution of AI workloads between cloud infrastructure and edge devices and the implications of latency, energy consumption, privacy, cost, and personalization for AI deployment.

14. Open Compute Project, "Open Data Centers for AI" and "Rack & Power Project."

    Provides broader infrastructure context for AI systems, including rack-scale architecture, power delivery, cooling, and the integration requirements created by increasingly dense AI computing platforms.

Ecosystem and Industry Strategy

15. TSMC, "3DFabric Alliance."

    Describes the collaborative ecosystem spanning EDA, IP, design services, memory, substrates, packaging, and testing required to support increasingly complex heterogeneous integration and advanced semiconductor platforms.

16. Jeroen Kusters et al., "2026 Global Semiconductor Industry Outlook," *Deloitte Insights*, February 5, 2026.

    Provides broader industry context on AI-driven semiconductor demand, investment priorities, supply-chain restructuring, capacity decisions, and strategic considerations facing semiconductor industry leaders.

Author

Evan Wang, Consulting Director, Semiconductors

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.

Omdia consulting

Omdia is a market-leading data, research, and consulting business focused on helping digital service providers, technology companies, and enterprise decision makers thrive in the connected digital economy. Through our global base of analysts, we offer expert analysis and strategic insight across the IT, telecoms, and media industries.

We create business advantage for our customers by providing actionable insight to support business planning, product development, and go-to-market initiatives.

Our unique combination of authoritative data, market analysis, and vertical industry expertise is designed to empower decision-making, helping our clients profit from new technologies and capitalize on evolving business models.

Omdia is part of Informa TechTarget, a B2B information services business serving the technology, media, and telecoms sector. The Informa group is listed on the London Stock Exchange.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help your company identify future trends and opportunities.

Get in touch

www.omdia.com
askananalyst@omdia.com

Copyright notice and disclaimer

The Omdia research, data, and information referenced herein (the "Omdia Materials") are the copyrighted property of TechTarget, Inc. and its subsidiaries or affiliates (together "Informa TechTarget") or its third-party data providers and represent data, research, opinions, or viewpoints published by Informa TechTarget and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document.

The information and opinions expressed in the Omdia Materials are subject to change without notice, and Informa TechTarget does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an "as-is" and "as-available" basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa TechTarget and its affiliates, officers, directors, employees, agents, and third-party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa TechTarget will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.