DTech News

DIGITIMES

COMPUTEX 2026

Top News

COMPUTEX 2026: AI takes center stage as industry leaders gather in Taipei

Research & Analysis

Agentic AI sparks new revolution in automotive industry; smart cars enter the autonomous decision-making stage

Short-term energy solutions under AI power bottlenecks: AI data centers to accelerate deployment of energy storage systems

The AI Inference Revolution:

Silicon Motion's Wallace Kou on the New Memory Paradigm and Global Supply Risks

Wallace Kou points out that the prosperity driven by AI inference is accompanied by a severe supply-demand imbalance, which is expected to last until 2027-2028. Credit: Silicon Motion

The global semiconductor landscape is undergoing a fundamental shift, moving from a focus on raw training power to the practical complexities of large-scale deployment. In an in-depth interview, Wallace Kou, President and CEO of Silicon Motion, detailed how the generative AI has evolved beyond its initial stages.

While the market's early gaze was fixed almost exclusively on NVIDIA's GPUs, the High Bandwidth Memory (HBM), and the CoWoS advanced packaging technology, Kou argues that the industry is now entering the "Inference" era that is turning previous under-estimation about storage's importance on their head.

The Shift from Training to Inference

The turning point for this realization occurred during the NVIDIA GTC conference in March 2026. CEO Jensen Huang unveiled the Vera Rubin architecture, a move that signaled a massive spike in demand for NAND flash memory. During the initial AI boom, the industry was preoccupied with training massive models, a process that relies heavily on the lightning-fast throughput of HBM. However, as these models move into the inference phase—where they are actually used by end-users to generate content or solve problems—the access to context, historical data, and massive datasets storage become the primary bottleneck.

Kou notes a dramatic shift in market sentiment. Only two years ago, storage was often an afterthought in the AI conversation; today, it is a critical scarcity. "There is currently not a single global cloud service provider or major smartphone manufacturer whose demand for DRAM and NAND is being fully satisfied," Kou observed. This supply-demand gap has triggered a financial windfall for storage module manufacturers and memory giants, with some stock prices skyrocketing up to tenfold as the market reacts to persistent shortages and rising prices.

Technical Paradigm Shift: CMX and the Infrastructure of Thought

At the heart of this transition is a new architecture introduced by NVIDIA: the CMX Context Memory Storage platform. This architecture is designed specifically to handle the "KV Cache" (Key-Value Cache), which allows AI models to remember the context of a conversation or a complex task during the inference process.

The hardware requirements for the CMX architecture are staggering in their scale and technical demands. Each individual Rubin GPU requires 16TB of dedicated storage to function effectively within this framework. At a system-level scale, a single NV72 Vera-Rubin setup can demand more than 1 Petabyte, or 1,000 Terabytes, of total storage capacity. Beyond mere capacity, the CMX architecture facilitates direct GPU access to storage, a feature that bypasses traditional latency bottlenecks and ensures that AI inference remains fluid and responsive.

While this creates a massive commercial opportunity for the storage industry, it also places an unprecedented strain on NAND production. Kou emphasizes that this is not just a cloud-based phenomenon. The explosion of Edge AI—AI processed locally on devices—is further complicating the supply chain. For instance, driven by major players like Meta, the market for smart glasses is expected to reach 60 million units this year. These wearable devices require high-performance embedded storage, creating a secondary front in the war for NAND capacity.

Silicon Motion's Role: Solving the QoS Bottleneck

As the world's leading NAND controller maker, Silicon Motion sits at the intersection of these competing demands. The primary technical challenge in modern AI environments is maintaining Quality of Service (QoS). In a multi-tenant cloud environment, where multiple GPUs are accessing shared storage simultaneously for different inference tasks, data transfer speeds can often fluctuate or drop.

To solve this, Silicon Motion has deployed its proprietary PerformaShape® technology. This technology ensures that even under heavy, concurrent workloads, the transmission speed remains stable. By stabilizing these data flows, Silicon Motion has positioned itself as an "indispensable stabilizer" in the AI ecosystem.

Beyond data path optimization, Silicon Motion is also extending its role into system-level infrastructure by providing enterprise-grade boot drives for leading AI GPU, TPU, and DPU platforms, ensuring system reliability and fast initialization at scale.

The Crisis of Imbalance: Kou's "Capacity Persuasion" Efforts

Despite the record-breaking revenues, Kou is deeply concerned about the "shadows" lurking behind this prosperity. The current memory market is suffering from a dangerous imbalance. To maximize profits and satisfy the insatiable hunger of AI cloud giants, major manufacturers like Samsung, SK Hynix, and Micron are funneling the majority of their capital expenditure (CAPEX) into HBM and DDR5 production.

This strategic pivot has effectively "squeezed" the production capacity available for standard NAND flash. Kou warns that this "AI squeezing effect" could lead to a collapse in traditional sectors. Over the past eight months, Kou has embarked on a global mission, meeting with leaders at Samsung, SK Hynix, Kioxia, SanDisk, YMTC, and Micron. His message is one of "capacity persuasion": he is urging these giants to reserve a portion of their production lines for the automotive, PC, and smartphone industries.

"If these foundational industries break because they cannot find parts, Edge AI will have no 'soil' to grow in," Kou warned. He believes that a total focus on the high-margin AI server market could eventually backfire, destroying the broader technology ecosystem that supports AI development.

A Stabilizing Strategy: From Cloud to Edge

Silicon Motion is positioning itself as the “transition enabler” for an industry in flux amid an expected 2–3 year supply shortage. As NAND manufacturers concentrate their internal resources on AI-driven initiatives, they are increasingly outsourcing non-core and mainstream projects, such as PCIe Gen5 controllers and embedded solutions. In this shift, Silicon Motion has emerged as a preferred partner to fill the resulting gap.

At the same time, as rising prices weigh on demand in the PC and smartphone markets, the company is helping customers pivot toward automotive and AIoT applications, including rapidly growing segments such as smart glasses, which are seeing a surge in shipments this year.

One of the most critical areas is the automotive sector, where Silicon Motion has spent a decade building a presence. While memory giants might see automotive requirements as "niche" or low volume compared to AI servers, Kou views them as essential to global stability. When major OEMs consider abandoning these specialized demands due to capacity constraints, Silicon Motion steps in to ensure the global automotive supply chain does not grind to a halt.

"We are not just looking for a surge in revenue; we want to fulfill our responsibility to the industry," Kou said. By providing stable controllers and storage solutions for AIoT and automotive applications, Silicon Motion is effectively repairing the cracks in a fractured global supply chain.

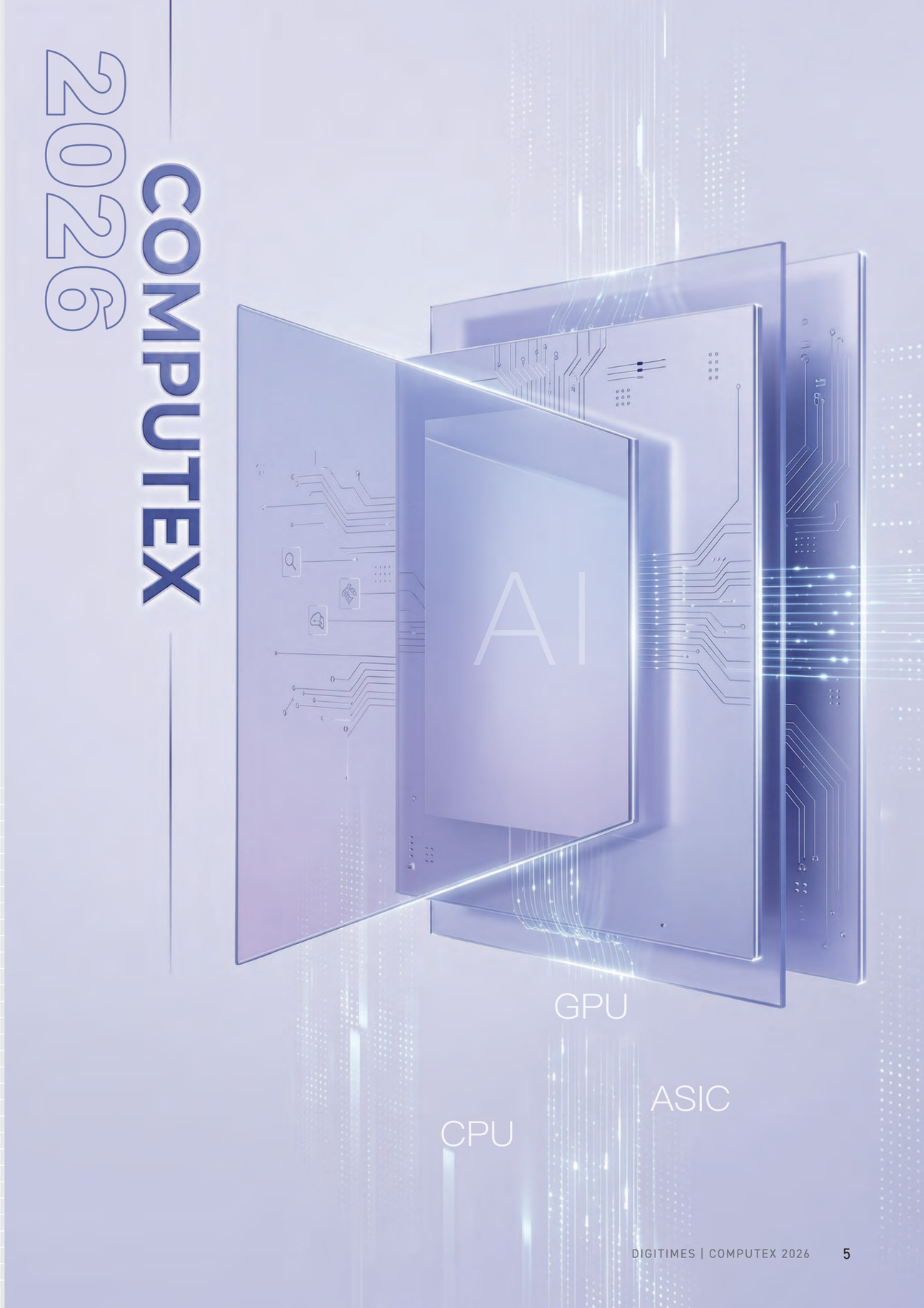
Future Outlook: 2027 and Beyond

The current supply-demand imbalance is not a temporary glitch but a structural reality that Kou expects to persist until at least late 2027 or 2028. Several factors make it nearly impossible to add capacity quickly, for example, land acquisition is increasingly difficult. The lead time for building specialized cleanrooms and procuring critical equipment now exceeds one year.

Kou predicts that while the DRAM shortage might begin to ease by the end of 2027, the relief for NAND will likely come even later. In this high-pressure environment, Silicon Motion's role as a key stabilizing force becomes increasingly important.

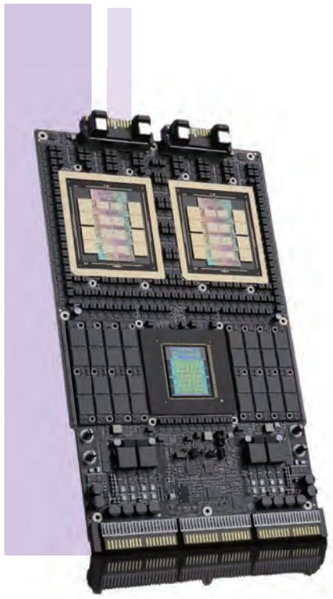
Particularly in emerging sectors such as smart IoT and automotive applications, Silicon Motion delivers reliable controller and storage solutions, filling the vacuum left by production shifts at major manufacturers or by projects lacking sufficient engineering support.

By helping global clients navigate the complexities of geopolitics and capacity wars, Silicon Motion aims to ensure that the AI revolution leads to a steady, sustainable future rather than a chaotic collapse of the broader tech industry.

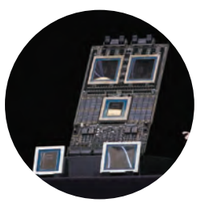


COMPUTEX 2026:
AI takes center stage as industry leaders gather in Taipei

Aaron Lee, Taipei; Elaine Chen, DIGITIMES Asia, Taipei

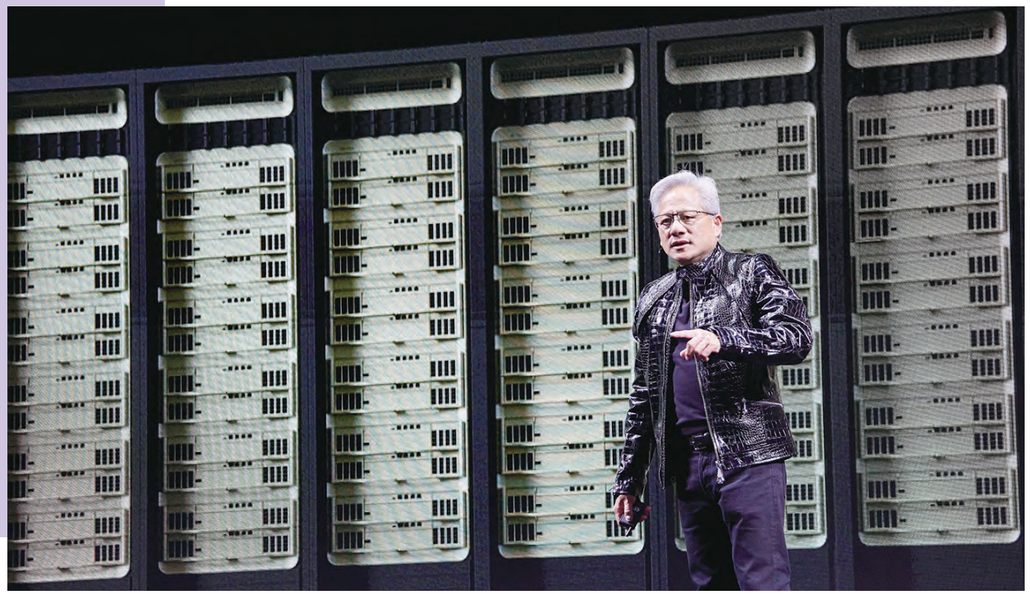


The NVIDIA Vera Rubin NVL72 platform integrates the Vera CPU, Rubin GPU, and NVLink 6 high-speed interconnect technology, targeting the large-scale training and inference demands of next-generation AI data centers. (NVIDIA)



COMPUTEX 2026 opened under the theme "AI Together," highlighting advances in AI computing, robotics, smart mobility and next-generation technologies. But this year's exhibition marks a notable shift in emphasis. Unlike the previous two years — when attention centered overwhelmingly on Nvidia and the computing infrastructure required for large-scale AI training — the conversation in 2026 has broadened considerably. GPUs remain at the center of the industry's ambitions, but custom AI chips, known as ASICs, along with a growing range of edge-AI applications, are now demanding equal attention. The change reflects a wider transition across the technology sector: from the race to build massive AI models toward the challenge of deploying artificial intelligence efficiently across devices, networks and real-world environments.

COMPUTEX 2026 is set to draw more than 1,500 exhibitors from 33 countries, filling 6,000 booths and underscoring the growing global stakes in artificial intelligence and next-generation computing. The event will feature keynote appearances by some of the industry's most influential executives, including Intel's chief executive, Lip-Bu Tan; Qualcomm's chief executive, Cristiano R. Amon; Marvell Technology's chief executive, Matt Murphy; and NXP Semiconductors's chief executive, Rafael Sotomayor. Jensen Huang, the chief executive of Nvidia, is expected to deliver a keynote on June 1 at the Taipei Music Center ahead of the exhibition. During the show, he is also likely to appear on the floor alongside supply-chain partners—appearances that often draw intense attention and help set the tone for industry conversations.



NVIDIA CEO Jensen Huang is scheduled to deliver a keynote at COMPUTEX 2026 preshow. Photo by Li Chien-liang (File photo)

A new AI architecture drives system-wide upgrades

A central theme this year will be Nvidia's next-generation AI architecture, Vera Rubin, which is expected to push AI servers into a new phase of design.

Taiwanese manufacturers are preparing to showcase technologies aligned with that shift, including full liquid-cooling systems, high-voltage power distribution, and rack-level power solutions—reflecting the rising energy demands of advanced AI workloads.

Companies such as Foxconn, Quanta Computer, Wistron, Wiwynn, Inventec, Compal Electronics, and Pegatron are expected to present a wide range of AI server and infrastructure solutions.

Foxconn, for instance, will highlight its vertically integrated AI server capabilities, from key components to full system design, along with emerging applications such as co-packaged optics and robotics. The company is also expected to demonstrate how AI is being deployed across smart manufacturing, electric vehicles, and urban infrastructure.

Quanta and its cloud subsidiary, QCT, will focus on systems built around Nvidia's Rubin architecture, including fully liquid-cooled designs and so-called "agentic AI" systems capable of autonomous task execution.

From training to inference, demand shifts reshape computing

Beyond hardware, a major shift is underway in how AI workloads are distributed.

As artificial intelligence moves from model training toward inference—the stage where trained models are deployed, demand is spreading beyond graphics processors to include central processing units.

That shift is benefiting companies like Intel and AMD. According to Intel, the ratio of GPUs to CPUs in AI workloads has narrowed from roughly 8-to-1

Intel CEO Lip-Bu Tan is scheduled to deliver a keynote at COMPUTEX on June 2. Photo by Li Chien-liang (File photo)

Google's 8th-generation TPU features a dual-chip architecture for the first time. (Google Cloud)

during the training phase to about 4-to-1 today, and could eventually approach parity.

The change is also lifting demand for general-purpose servers, with recent industry data showing stronger-than-expected shipments. Analysts expect global server shipments to surpass 5 million units in the second quarter of 2026 for the first time, reflecting both AI adoption and broader enterprise upgrades.

AI expands into the physical world

As AI applications broaden, they are increasingly moving beyond data centers into physical systems, robotics, autonomous vehicles, smart factories, and smart cities.

Taiwanese companies across industrial and embedded systems are expected to highlight these developments, reflecting a shift toward what some executives describe as "physical AI," where software intelligence is integrated into real-world machines and infrastructure.

Custom chips emerge as the fastest-growing segment

Another key focus at COMPUTEX will be the rapid rise of ASICs, servers, and custom-designed chips developed by cloud providers.

Led by systems built around Google's tensor processing units (TPU), ASIC-based servers are expected to be the fastest-growing segment of the AI infrastructure market in 2026.

Taiwanese companies are positioning themselves to capture that growth. MediaTek, in particular, has raised its outlook for ASIC revenue, doubling its 2026 target to US$2 billion and projecting that the broader market could reach US$70 billion to US$80 billion as early as 2027, ahead of previous estimates.

The company's chief executive, Rick Tsai, said demand for custom AI chips has accelerated further in recent months, with customers showing increasing urgency.

MediaTek has also brought on Douglas Yu, a former senior research executive at TSMC, as an adviser—an indication of its ambitions to strengthen capabilities in advanced packaging and system integration.

Mark Yee | DIGITIMES

Agentic AI sparks new revolution in automotive industry; smart cars enter the autonomous decision-making stage

The recent rapid development of agentic AI has become a key force driving the intelligent upgrade of the automotive industry. Compared to generative AI or a single AI agent, agentic AI centers on the collaborative operation of multiple AI agents, combining the ability to call external tools, and establishing a cross-module integration and autonomous decision-making architecture through shared memory and continuous learning. Agentic AI can not only accomplish complex goals but also adjust strategies in response to dynamic environments, bringing significant breakthroughs to the development of autonomous driving, smart cockpits, and smart manufacturing.

AI capabilities reshuffle; agentic AI leaps to the highest level of intelligent form

In the automotive industry, different AI types show clear stratification in core capabilities and application scopes. Generative AI mainly focuses on content creation, such as generating voice, images, or dialogue content in smart cockpits, which already has high commercial maturity. However, its operational nature is still a one-way mode from input to output, lacking task planning and autonomous decision-making capabilities.

An AI agent uses a large language model combined with external tools and limited memory to execute tasks in specific scenarios, such as assisting with navigation planning, restaurant reservations, or providing driving suggestions. AI agents already possess preliminary task processing capabilities, but their logic is still confined to a single agent architecture, making it difficult to handle cross-system and cross-module problems.

In contrast, agentic AI can form a system architecture with collective intelligence through collaboration and memory sharing among multiple agents, breaking down tasks, dividing roles, and adjusting strategies for complex goals, and making autonomous judgments and actions based on the context. In automotive scenarios, it can achieve cross-module tasks, such as combining autonomous driving decisions, cockpit interactions, and cloud services.

From a single model to multi-agent collaboration; AI decision-making systems comprehensively upgraded

The core operational mechanism of agentic AI lies in the collaboration among multiple AI agents, using a large language model as the reasoning hub to form a system architecture capable of autonomous learning and continuous optimization. Through such a design, the system can not only execute tasks but also observe the environment and accumulate experience during actions, further optimizing to enable overall capabilities to continuously grow over time.

In this architecture, the large language model plays the role of the brain, responsible for natural language understanding, logical reasoning, and task planning. When the system receives a goal, the large language model can parse complex instructions and break them down into multi-step actions, coordinating different agents to execute tasks, thereby improving overall decision-making efficiency and flexibility.

In summary, agentic AI is not a single large model, but is composed of multiple functionally specific AI agents. In addition, agentic AI has memory and tool usage capabilities. The system can store past decisions and experiences as a reference for subsequent tasks, reducing the probability of repeated errors, while also expanding its capability range through API or external tools.

Autonomous driving, smart cockpits, and manufacturing advance simultaneously; agentic AI accelerates broad market penetration

The practical application of agentic AI in the automotive industry mainly focuses on three major aspects: autonomous driving safety enhancement, smart cockpit experience upgrade, and smart manufacturing and supply chain optimization, becoming a key technology driving the automotive industry toward intelligence.

In the field of autonomous driving, agentic AI breaks through the limitations of traditional modular systems operating independently through real-time collaboration of multiple agents. Agents for perception, planning, control, and safety can operate synchronously and share information, enabling the system to respond in real-time to complex and unpredictable road conditions. For example, when facing scenarios such as pedestrians, bicycles, or temporary construction, multiple agents can quickly make collaborative decisions, plan the safest driving route, and continuously optimize decision-making capabilities through closed-loop learning.

In terms of smart cockpits, agentic AI shifts vehicles from passive response to active service. Agents for voice, navigation, entertainment, and vehicle status operate collaboratively, understanding user intent and context to provide a personalized experience. For example, when driving fatigue is detected, the system will not only issue a warning but also proactively play preferred music, adjust interior lighting, and search for the nearest rest stop in the navigation, forming a complete solution.

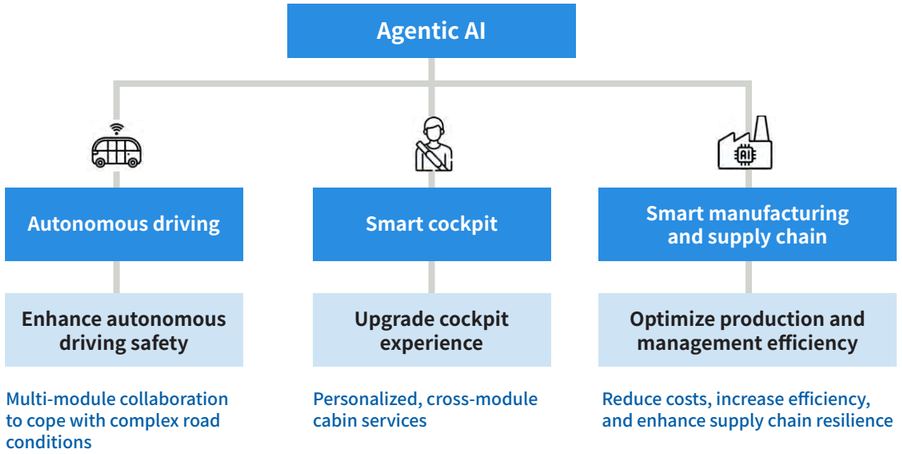
In terms of smart manufacturing and supply chains, agentic AI can achieve optimization of production processes and resource scheduling. Agents for scheduling, quality inspection, and equipment maintenance can make autonomous decisions based on real-time data, reducing downtime risks and improving production line efficiency. At the same time, integrating ERP and supply chain information helps optimize inventory and logistics scheduling, enhancing overall operational resilience.

Agentic AI, AI agent, and generative AI feature comparison

Item	Agentic AI	AI Agent	Generative AI
Core definition	Multiple AI agents demonstrate system-level intelligence through collaboration, shared memory, and decision-making	Uses LLM for reasoning and can call tools and memory modules to complete specific tasks	Primarily used for content creation, generating text, voice, or images based on input instructions
Autonomy	High Proactively executes tasks based on goals	Medium Can act autonomously within a limited scope	Low Relies on user input instructions
Memory capability	High	Medium	Limited
Example	In autonomous driving systems, agents for perception, planning, etc., collaborate	In-vehicle AI assistant handles driver needs, such as reservations and navigation	In smart cockpits, generates voice responses or entertainment content
Limitations	High collaboration complexity; risk of error propagation and control issues	A single agent has difficulty handling global tasks	Lacks actionability; unable to autonomously plan or operate
Development stage	Still in early development stages in the automotive industry	Gradually being introduced into industries like automotive	Already commercialized on a large scale

Source: Amazon, NVIDIA, compiled by DIGITIMES, April 2026

Three major application scenarios of agentic AI in automotive industry



Source: DIGITIMES, April 2026.

The arrival of the agentic AI era; the competitive logic of the automotive industry will completely change in the next 5 years.

In the short term (1-3 years), agentic AI applications will concentrate on single scenarios, such as smart cockpit voice assistants, autonomous driving navigation modules, and vehicle quality inspections, with clear functions and relatively low adoption thresholds, enabling faster commercialization.

Entering the medium term (3-5 years), multiple AI agents will gradually achieve cross-module collaboration. Different agents in systems such as smart cockpits, autonomous driving, and smart manufacturing will form more efficient integrated applications through data sharing and collaborative decision-making.

In the long term (over 5 years), it is estimated that agentic AI is expected to form a cross-industry and cross-automaker smart ecosystem, strengthening mutual collaboration among multiple AI agents in different fields, and driving the industry into an era of high automation and intelligent decision-making.

However, the adoption of agentic AI still faces multiple challenges. On the technical level, multi-agent collaboration increases system complexity, and a single agent error may affect the overall decision, so stability and reliability still need to be strengthened. On the application level, high development costs and un-unified communication standards limit the speed of adoption by businesses. The governance level involves issues such as liability attribution and data privacy; for example, the determination of liability for autonomous driving accidents and regulations on data usage both require synchronous improvement of laws and regulatory mechanisms.

Component Sourcing in 2026

Constraint, Volatility, and a New Procurement Reality

The next 12 months will be extremely challenging for purchasing departments. The global semiconductor supply chain remains constrained, with demand continuing to outpace supply. While some lead times have eased slightly, the market has not returned to balance. In key areas, memory demand shows no sign of slowing.

The AI boom is a major driver, pulling capacity toward high priority programs and consuming available supply faster than it can be replaced. However, AI is only part of the story. Capacity limits, allocation controls, uneven supplier recovery, and years of underinvestment continue to strain the ecosystem. As a result, the market is defined less by broad shortages and more by persistent, category-specific constraints.

For purchasing teams, the impact is immediate. Many are facing pricing increases of 25% to 100% or more while receiving only 40% to 50% of forecasted demand. This disconnect affects nearly every industry, including data center, industrial, automotive, medical, and consumer electronics.

The past year showed how costly misreading the market can be. Some OEMs delayed purchasing in 2025, expecting conditions to ease, only to return at higher prices with fewer supply options. What was once a pricing decision has become a continuity decision.

Access to supply, speed, and flexibility now matter more than theoretical cost savings. At NewPower Worldwide, we see customers prioritizing continuity and execution as constrained supply and uncertain lead times persist. How organizations respond will determine whether procurement becomes a source of risk or resilience.
www.newpowerww.com

NEWPOWER

EMPOWERING SUPPLY CHAIN
DATA CENTER • AUTOMOTIVE • $3B REVENUE
DRAM • ROBOTICS • SRAM • COST SAVINGS
GLOBAL FOOTPRINT • BOM ANALYSIS
CLOUD COMPUTING • NEWPOINTS®
$500M COMMITTED FACILITY
ARTIFICIAL INTELLIGENCE
PASSIVES • ACTIVES • HPC
CUSTOM SOLUTIONS
PROCESSORS
MEMORY
GPU
EMPOWER™
ALL-IN-ONE PLATFORM

INFO@NEWPOWERWW.COM / 800-985-9279 / WWW.NEWPOWERWW.COM

Adam White will deliver a keynote address at COMPUTEX 2026 on June 4 titled “Infineon Powering AI from Grid to Core to Physical AI.” The session will explore emerging trends and strategic opportunities in power infrastructure of AI era.

AI Data Center enters gigawatt scale; Power architecture emerges as industry's defining competitive edge

Infineon fuels power delivery innovation through Its “From Grid to Core” strategy

Infineon reveals the roadmap for the evolution of power architecture "from Grid to Core" in future AI data centers

As AI data center rapidly scales toward the gigawatt (GW) level, energy management is evolving from a supporting function into a defining strategic pillar. Against this backdrop, Infineon Technologies is redefining power infrastructure of AI era through its comprehensive “From Grid to Core” strategy, integrating energy efficiency, power density, and system resilience from the electrical grid to the processor core.

Scaling AI: Strategic Power Solutions Matter

The evolution of artificial intelligence (AI) has accelerated far beyond the trajectory once predicted by Moore's Law. As AI models continue to expand in parameter scale and real-time inference becomes increasingly critical, demand for computing power is rising at an unprecedented pace.

Today, the power consumption of a single GPU is rapidly approaching the kilowatt level, and the power density of a server rack has increased significantly from less than 60kW in the past to exceeding 100kW, now moving toward a new threshold in the megawatt range.

This is not just numerical growth; it represents a fundamental shift in power architecture. As AI computing clusters expand rapidly, traditional 48V busbars and AC power distribution architectures are quickly approaching their physical limits in terms of power loss, thermal management, and spatial constraints.

Adam White, President of Infineon's Power & Sensor Systems Division, emphasized that future competition in AI infrastructure will no longer be limited to chip performance. Instead, it will be a cross-disciplinary integration battle encompassing power electronics, materials technology, and system architecture.

From Grid to Core: Rethinking the power delivery architecture

At this critical turning point for the industry, Infineon's competitive advantage lies not in a single breakthrough product, but in its ability to orchestrate and optimize the entire power delivery chain at the system level.

“From Grid to Core” is more than a product strategy—it is a multi-phase architectural framework designed to reshape the future AI energy chain. Developed through early collaboration with global hyperscalers and ecosystem partners, the strategy enables Infineon to address evolving AI power demands across every phase of infrastructure, from utility grids to processor-level power management.

Power grid: Enabling a sustainable, high-efficiency power with HVDC

At the front end of the data center, power infrastructure is transitioning from traditional mechanical systems to highly integrated solid-state solutions. Future AI facilities are expected to increasingly adopt decentralized DC microgrids, enabling greater efficiency, flexibility, and resilience in energy management.

Adam White (right), Division President Power and Sensor Systems, and Peter Wawer (left), Division President Green Industrial Power of Infineon Technologies

By leveraging silicon carbide (SiC) technology in solid-state transformers (SSTs), system weight can be dramatically reduced—from nearly 20 tons to approximately 500 kilograms—while simultaneously improving overall energy efficiency by more than 1%. Beyond optimizing space utilization and operational costs, this advancement signals a broader industry migration from electromechanical infrastructure toward semiconductor-driven power systems.

As SSTs and related technologies become integrated into AI power infrastructure, a multi-billion-dollar semiconductor opportunity is emerging across next-generation energy systems.

At the same time, the power grid is evolving beyond its traditional role as just an energy source. Through digitally controlled power systems with real-time monitoring and remote management capabilities, combined with solid-state circuit breakers (SSCBs) featuring microsecond-level response times, the grid is becoming an intelligent energy platform capable of continuous optimization and predictive management.

Server rack: Reshaping power density and maximizing efficiency

As data centers advance toward GW-scale deployments, power distribution architecture is undergoing a fundamental redesign. Infineon is driving the industry's transition from traditional 48V systems to ±400V and 800V high-voltage DC

architectures. Through the design of three-phase power sidecars, Infineon is restructuring power supply and computing systems to establish a more efficient and flexible power distribution model.

At the same time, the power architecture of AI data centers is following a clear evolutionary path: moving from integrated server rack designs to high-voltage DC and sidecar power supply configurations, and ultimately advancing to gigawatt-scale infrastructures that incorporate DC microgrids.

Meanwhile, The high-frequency characteristics of gallium nitride (GaN) components enable intermediate bus converters (IBCs) to achieve over 98% conversion efficiency and exceptionally high power density in an extremely compact form factor, significantly reducing power transmission losses and freeing up more space for AI computing resources.

Processor core: Power density and new architectures for next-gen AI compute

At the processor core—the final stage of power delivery—the challenge shifts toward managing extreme current density and ultra-fast transient response.

To support next-generation GPUs requiring massive current delivery and rapid load transitions, Infineon has introduced a digital multiphase PWM controller alongside the industry's first TLVR four-phase power module. These technologies are engineered to provide highly stable, efficient, and responsive power delivery for AI processors operating under increasingly demanding workloads.

In response to the next generation of GPUs demanding ultra-high current and rapid load changes, Infineon has introduced digital multiphase PWM controllers and the industry's first TLVR quad-phase module. By leveraging high-precision telemetry and digital control technologies, power systems have transformed from energy suppliers into intelligent platforms capable of real-time monitoring, prediction, and optimization.

From AI data center to physical AI

If data centers form the foundation of AI computing power, the physical world will be where AI's true value is ultimately realized. As AI increasingly expands into humanoid robots, autonomous systems, and intelligent manufacturing equipment, demands for energy efficiency, real-time responsiveness, and system reliability will become even more critical.

Infineon is extending its long-established expertise in power management beyond data centers into the emerging era of Physical AI. By integrating sensing technologies, actuate, security and connectivity solutions, and high-efficiency power modules, the company is enabling a comprehensive functional blocks that empower humanoid robots to perceive, think, act and connect, safely and secured in a real-world environment.

Push the boundaries of power technology in AI era

"We Power AI" is not just a slogan for Infineon—it is a concrete commitment to the future of the industry. From gigawatt-scale data centers to physical AI, Infineon continues to push the boundaries of power technology, ensuring that every watt of energy is transformed into the greatest possible value for AI.

As the industry advances toward the next generation of computing, energy management will become the decisive key to truly unlocking the full potential of AI. In this wave of transformation, Infineon is joining forces with ecosystem partners to stand at the forefront of defining the future.

At COMPUTEX 2026, Fibocom outlined its perspective on the deployment path of edge AI: as AI moves from the cloud to end devices, the key to large-scale adoption lies not only in model performance, but also in the synergy between stable connectivity and local computing power.

Based on this insight, Fibocom continues to strengthen its "Connectivity + AI" capability. On one hand, with wireless modules at its core, Fibocom has built a connectivity product portfolio covering FWA, AI CPE, and next-generation Wi-Fi 8, providing a high-speed, low-latency, and highly reliable connectivity foundation. On the other hand, by leveraging its full-stack AI solutions, Fibocom enhances on-device AI capabilities, enabling terminals with local perception and processing power to improve the stability and responsiveness of AI applications in complex environments.

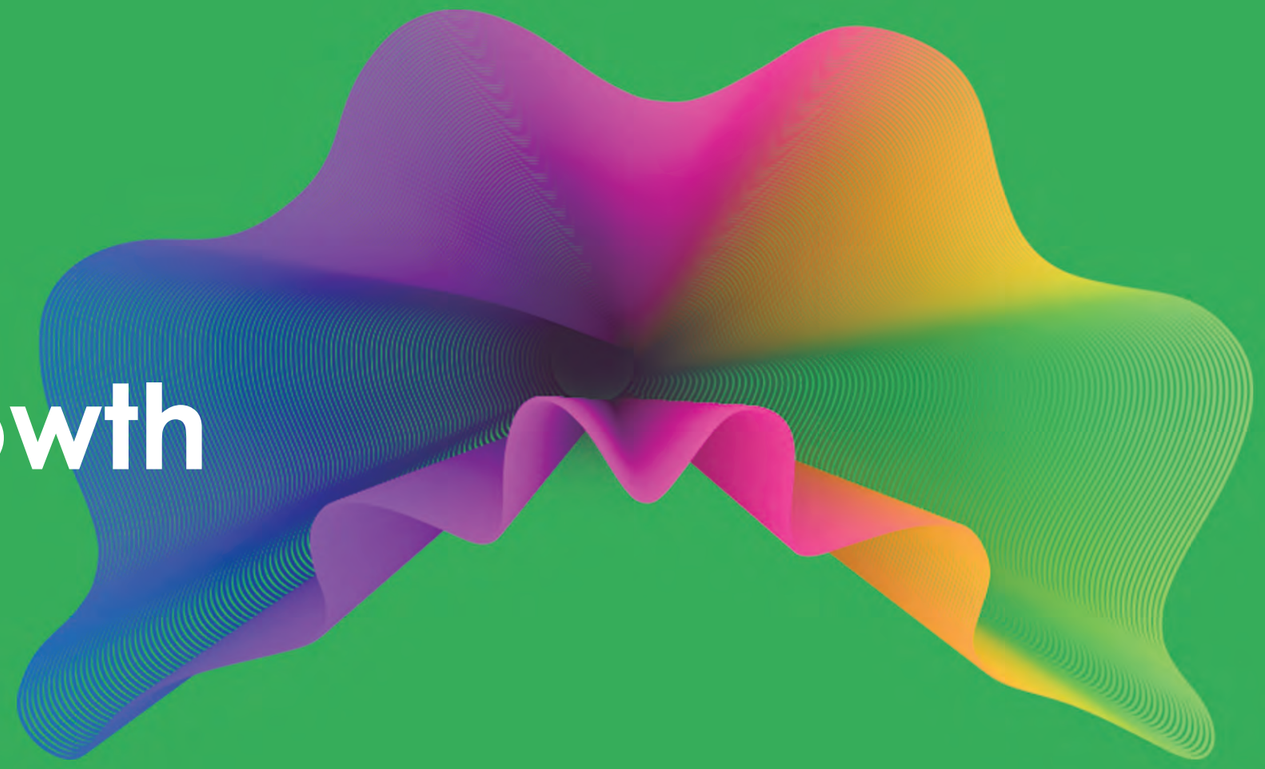
For intelligent edge scenarios, Fibocom launched the Fibocom ClawBox, a high-efficiency AI computing terminal natively compatible with intelligent agents such as OpenClaw and Hermes Agent. Built on a high-performance heterogeneous computing architecture, the product integrates CPU, GPU, and NPU, delivering up to 18 TOPS @ INT8 hybrid precision computing power. With typical power consumption at around 5W, it enables high-density, multi-task parallel inference. It supports multi-stream video analytics, complex visual structuring, voice, and multimodal model inference, while offering low latency, high reliability, and localized deployment capabilities. Combined with Fibocom's Skill Hub, it is widely applicable in security, transportation, robotics, and other intelligent edge scenarios, supporting the large-scale deployment of smart industry applications.

At COMPUTEX 2026, Fibocom will showcase a range of innovations at Booth K1022, including AI companion solutions, robotic lawn mower solutions, mobile robotics solutions, and more, demonstrating how "Connectivity + AI" is creating value in high value scenarios. Industry partners are warmly welcome to visit and connect with us.

Apacer

Storage, Empowering AI Growth

Hall 1, Booth I0108 | June 2–5, 2026



Gaming Beyond Limits

Creative Pro Studio

Smart Factory in Action

Smart Living

Sustainable Display

Industrial Growth

More Info

Sabrina Yu | DIGITIMES

Short-term energy solutions to AI power bottlenecks: Accelerating BESS Deployment in AI data centers

Artificial intelligence data center (AIDC) is entering an era of GW-level power competition, with energy becoming a key variable limiting the expansion of computing power. Battery energy storage systems (BESS), due to their rapid deployment and high power output characteristics, serve as mature solutions capable of providing MW-level dispatchable capacity for data centers in the short term. The AIDC energy storage market has already seen the emergence of three types of competing suppliers, among which energy storage system integrators are advancing the fastest in AIDC deployments. It is expected that AIDCs built in 2026 will accelerate the adoption of BESS.

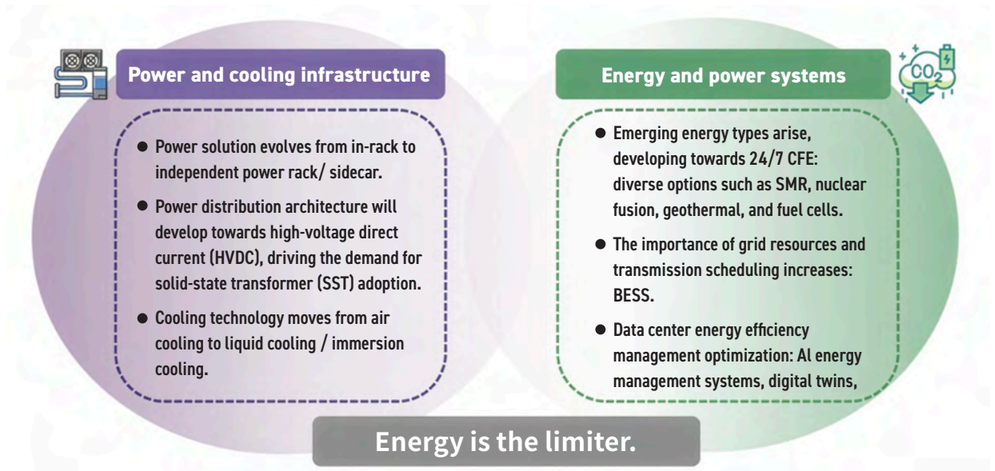
Rising AIDC power consumption makes energy a key limiting factor for computing power expansion

Looking at the evolution of AIDC power demand, AI chip power previously fell below 1kW. Today, taking Nvidia GB200 and GB300 chips as examples, the power consumption of a single GPU has surpassed 1kW, and the total energy consumption of a GB300 NVL72 rack reaches 120-140kW. DIGITIMES estimates that by 2030, a single AI chip will exceed 2kW, AI server racks will reach the MW level, and AIDC will enter the GW scale. Global data center power consumption is projected to surpass 1,000 TWh by 2030.

Currently, Meta, OpenAI, and xAI have all planned GW-level AIDCs, with Meta expecting to launch its operation as early as 2026. Among them, the Stargate project, co-developed by OpenAI and Oracle, aims for a data center capacity of 10GW within four years, making it the most ambitious in scale. However, OpenAI recently shifted from building its own data centers to prioritizing renting servers from major cloud providers, reflecting that the competition in AI infrastructure has pivoted from raw computing power to construction speed and energy acquisition capabilities. Additionally, it is worth noting that xAI claimed to have launched a GW-level AI training cluster in January 2026, reaching the GW scale faster than Meta. The key lies in xAI's energy strategy, which involves the massive deployment of gas turbines and energy storage systems.

The critical issue currently being discussed regarding AIDC is that energy has become the core variable limiting computing power expansion. The AIDC value chain is facing a dual-axis transformation with the first being power and cooling infrastructure (internal server architecture), including power solutions evolving towards independent power racks/sidecars, power distribution architectures shifting towards high-voltage direct current (HVDC), and cooling technologies moving towards liquid cooling to enhance the energy efficiency of unit computing power. The second is energy and power systems (external energy architecture), notably, the rise of emerging 24/7 carbon free energy (CFE) sources—including fuel cells, BESS, nuclear power, and geothermal energy—along with the transition of BESS into critical core nodes for grid resource allocation and power scheduling, are highly worthy of attention.

Dual-axis transformation of the AIDC value chain



Source: DIGITIMES, April 2026

*Note: CFE stands for carbon free energy; SMR stands for small modular reactor.

BESS serves as short-term AIDC energy solution; AIDC energy storage competition landscape takes shape

Global AIDC development is confronting a structural gap where "the speed of computing power construction exceeds the speed of power infrastructure deployment." BESS, due to their fast deployment and high power output, can provide MW-level scheduling capacity in the short term, making them a crucial energy configuration to meet the high load capacity of data centers.

It is worth noting that in the 800 VDC data center architecture promoted by Nvidia, energy storage will no longer serve merely as a backup system. Instead, it will be positioned as a critical active component in the power architecture, serving as the core of system-level power dispatch. This includes BESS located at the front of the medium-voltage grid, focusing on grid interconnection and stable campus-level power supply, as well as BESS located behind the DC power distribution, which will focus on regulating the real-time power of AI loads.

DIGITIMES observes that the competitive landscape of the AIDC energy storage system supply chain has already formed. Three types of enterprises are entering the AIDC energy storage market from the angles of "cell supply," "system integration," and "power architecture dominance." Future competition will shift from single products to overall power architecture and grid participation capabilities.

Representative players of cell suppliers include CATL and Hithium, mostly Chinese companies. In developing AIDC cells, they are all focusing on high cycle life, long-duration energy storage, and large capacity, pursuing advantages in cost and energy density. These companies not only supply BESS system integrators but also compete with them in the AIDC energy storage market.

Representative players of BESS system integrators include US-based Tesla and Fluence, as well as Chinese firm Sungrow. Most of these companies possess energy management system (EMS) capabilities, and some have developed their own power conversion systems (PCS). This type of company is currently making the fastest deployment progress in the AIDC energy storage market.

Power infrastructure providers include major industry partners named by Nvidia in its push for 800 VDC data center power systems, such as Vertiv, Schneider Electric, and Eaton. These companies master the overall data center power architecture and integrate energy storage batteries into their designs. It is expected that they will play an increasingly critical role in future AIDC energy storage systems.

Three major types of suppliers in AIDC energy storage market

Source: DIGITIMES, April 2026

AIDCs expected to accelerate BESS deployment in 2026

North America is the market with the most pronounced demand for AIDC BESS. The AIDCs under Elon Musk's companies have become testing grounds for BESS, and Fluence is currently negotiating AIDC energy storage system orders exceeding 30 GWh. It is anticipated that AI computing power construction in 2026 will drive a wave of demand for BESS installations.

As of early February 2026, approximately 150 Megapack energy storage systems had been deployed at Tesla's Cortex 2 supercomputing cluster site, which is expected to go online in mid-2026. The xAI Colossus data center has also successively deployed 156 and 168 Megapacks. Assuming a capacity of 3.9MWh per unit, the total energy storage capacity for the two phases reaches approximately 1.2-1.3GWh, marking the largest AIDC BESS deployment to date. Most of the Tesla Megapacks currently deployed in AIDCs are the existing second-generation 3.9MWh products, while the latest 5MWh Megapack 3 is expected to be gradually applied in subsequent projects.

During its fourth-quarter-2025 earnings call, Fluence mentioned that potential AIDC energy storage orders exceeded 30GWh, indicating that AIDCs have become a significant source of demand in the energy storage market. Most client requirements are for 6-hour energy storage to secure faster grid connection progress, lower the risk of AIDC downtime, and smooth out power fluctuations. Currently, the energy storage projects deployed by Fluence predominantly utilize Gridstack Pro (including Pro 5000) energy storage batteries, and it is highly likely that the AIDC projects under negotiation will also adopt the Gridstack Pro 5000 system.

Global AIDC energy storage demand is projected to reach 200-300GWh by 2030, with the US market accounting for the largest share. DIGITIMES believes that from a regional perspective, the US has become the market with the most prominent demand for AIDC BESS, driven by the pressure of GW-level AIDC construction and grid bottlenecks. As energy storage cells gradually standardize, the competitive focus for AIDC energy storage systems is shifting from batteries to PCS and system integration capabilities. With data centers expected to progressively adopt 800 VDC architecture in the future, the core of these systems is anticipated to transition to solid state transformer (SST).

JMicron

JMicron Technology Corp., a global leader in high-speed Bridge IC solutions, today announced it will unveil its flagship innovations at COMPUTEX 2026. Building on years of proven expertise in high-speed interface technologies and strong market recognition, JMicron will present a portfolio centered on extreme performance, intelligent backup, and hyperscale expansion, demonstrating its continued momentum in advancing the digital storage ecosystem.

In response to the stringent demands for performance and data protection in big data and professional storage markets, JMicron introduces three key controllers: JMS591, JMS591U, and JMB595.

JMS591 is purpose-built for multi-bay HDD array applications, supporting USB 20G and eSATA interfaces. In RAID 0 mode, it delivers read/write performance exceeding 2000 MB/s, providing a powerful foundation for NAS and DAS systems.

JMS591U targets enterprise and professional users, integrating JMicron’ s exclusive Offline Clone & Erase technology. It enables 1-to-4 high-speed duplication and DoD-compliant data sanitization, offering a one-touch solution for data migration and security compliance.

JMB595 is a high-performance PCIe Gen4 x4 to 16-port SATA III expansion controller. Through cascading architecture, it can scale to connect up to 240 storage devices, ideally suited for hyperscale data centers and surveillance storage infrastructures.

“JMicron has long been committed to transforming complex high-speed transmission technologies into intuitive and powerful hardware solutions,” said Ming-Cheng Lin, VP of Sales and Marketing at JMicron. “The products showcased at COMPUTEX not only push storage performance beyond the 2000 MB/s milestone, but also precisely address customers’ critical needs in large-capacity storage and security compliance through integrated RAID engines and offline cloning technologies. We are not merely a chip supplier—we are a strategic partner helping customers build a solid foundation in the data-driven era.”

JMicron cordially invites global partners and media representatives to visit its booth at COMPUTEX 2026. Through face-to-face engagement and forward-looking technology demonstrations, JMicron looks forward to exploring the future of storage innovation together and ushering in a new era of high-efficiency data transmission and intelligent data management.

TREND MICRO™

01

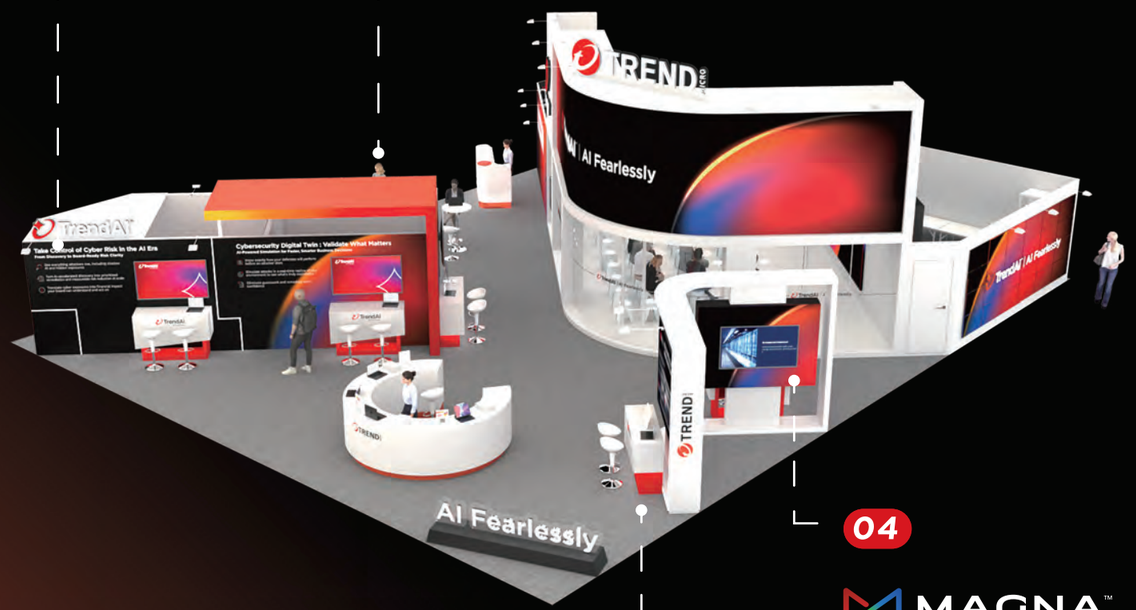
TrendAI™

Lead the future of AI, AI fearlessly

02

TrendLife™

Enabling Peace of Mind Across Your Family in the AI Era



04

MAGNA™

Powering the Sovereign Integrated AI Value Chain

03

VicOne

Secure Physical AI for Real-World Safety

Reserve First!

TrendAI™

AI Fearlessly

AI threats move fast.
We move first.

Welcome to visit TrendAI Booth

L0417a

LIVA

COMPACT EDGE AI FOR EVERY INDUSTRY.

From healthcare systems to smart retail and edge AI, LIVA Mini PC delivers reliable, compact computing for diverse real-world deployments.

LIVA MINI PC SERIES LINEUP

Q Series — Ultra Compact

Z Series — High Performance

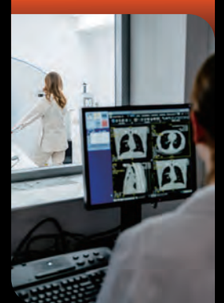
ONE Series — Flexible Platform

P Series — Expandable Performance

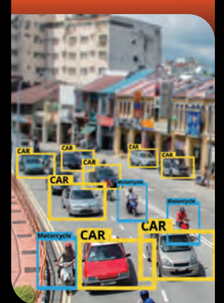
M Series — Aluminum Thermal

Versatile Mini PC Solutions for Diverse Deployments

Medical Center



Edge Computing



Retail kiosk

Office desktop

Digital Signage

ELITEGROUP

Elitegroup Computer Systems Co., Ltd.
9F., No. 22-1, Sec. 3, Zhongshan N. Rd., Zhongshan Dist, Taipei City 104, Taiwan (R.O.C.)
Tel: +886-2-2162-1177 Fax: +886-2-2658-8851
http://www.ecs.com.tw/

COMPUTEX TAIPEI
6 / 2 ~ 6 / 5

Invitation

BIOSTAR Booth: P0914
1F, TaiNEX 2

Exhibition Highlights

JMicron Technology Corporation
Booth No. Courtyard by Marriott Taipei #SEA

TEL 886-3-5797389 www.jmicron.com

A Premier Multi-Bay Storage Solution
Empower Extreme Performance: Pioneering the Future of Multi-bay Infrastructure

KIOXIA Taiwan Corporation
Booth No. X0001

TEL 886-2-5089909 www.kioxia.com

KIOXIA will showcase its latest innovative NAND flash and SSD solutions for the AI era at COMPUTEX 2026 (Booth X0001, 3F, TaiNEX 2). Additionally, don't miss the forum, "Exploring the Critical Role of SSDs in AI," on June 4, 14:30-14:55 at TaiNEX 2, Room 701. Join KIOXIA to discover the future of memory technology and make real innovations with us!

Apacer Technology Inc.
Booth No. I0108

TEL 886-2-22678000 www.apacer.com

For Edge AI, we deliver thermal, efficient, reliable storage; plus AOI, ePaper, gaming DRAM, USB4 SSD.

Silicon Motion, Inc.
Booth No. G0001

TEL 886-2-22196688 www.siliconmotion.com

Silicon Motion will showcase its latest Edge SSD, portable SSD, embedded eMMC/UFS controllers, along with Ferri solutions for automotive. The company will also highlight AI infrastructure applications, including enterprise SSD controllers and enterprise boot drives for the AI era.

Solid State Storage Technology Corporation
Booth No. R0924a

TEL 886-2-87982886 www.ssstc.com

SSSTC showcases its full industrial and enterprise SSD lineup at TaiNEX 2, 4F. A key highlight is the “Immersion Cooling SSD Solution” for AI data centers, addressing HPC thermal challenges while improving PUE. SSSTC SSDs are widely deployed across industrial, enterprise and data center, cloud, and client PC markets.

Infineon Technologies Taiwan Co., Ltd.

www.infineon.com

Infineon President Adam White will deliver a speech at the COMPUTEX Forum on June 4th titled " Infineon powering AI from grid to core to physical AI." Join us to explore new trends in the energy landscape of the AI era.

Carl Zeiss Co., Ltd.
Booth No. J1109

TEL 886-2-2182-2985 www.zeiss.com

ZEISS’ cross-business-unit Chip-to-Rack quality solutions span critical AI hardware applications from chips to rack-level integration. Supporting workflows from failure analysis to automated inspection, ZEISS helps improve AI hardware yield, reliability, and quality across the full value chain.

Trend Micro
Booth No. L0417a

https://go.trendmicro.com/tw/business/COMPUTEX2026_event/index.html

Trend Micro arrives at COMPUTEX 2026 under one rallying call: Adopt AI Fearlessly. The booth brings together four forces spanning every layer of AI adoption.
TrendAI™, the global AI security leader and enterprise business unit of Trend Micro, empowers organizations to securely drive forward at the speed of AI, through……

xMEMS Labs, Inc
Booth No. J1 229A

xmems.com

µCooling™: Solves thermal bottlenecks, boosts thermal budget for edge AI & fan-less datacenter racks.

Fusion Worldwide
Booth No. I-1330

www.fusionww.com

Global electronic components distributor. Focused on semiconductors, CPU, storage, memory, and supply chain risk management.

AGI Technology Co., Ltd.
Booth No. L1117a

TEL 886-27937256 agi-gear.com

We officially entered the industrial market by launching industrial-grade application solutions and DDR5 RDIMM memory, ushering in a new era of AI computing. We also partnered with StorArt for co-development and collaborated with the popular anime Dan Da Dan to create innovative crossover initiatives…….

DIGITIMES