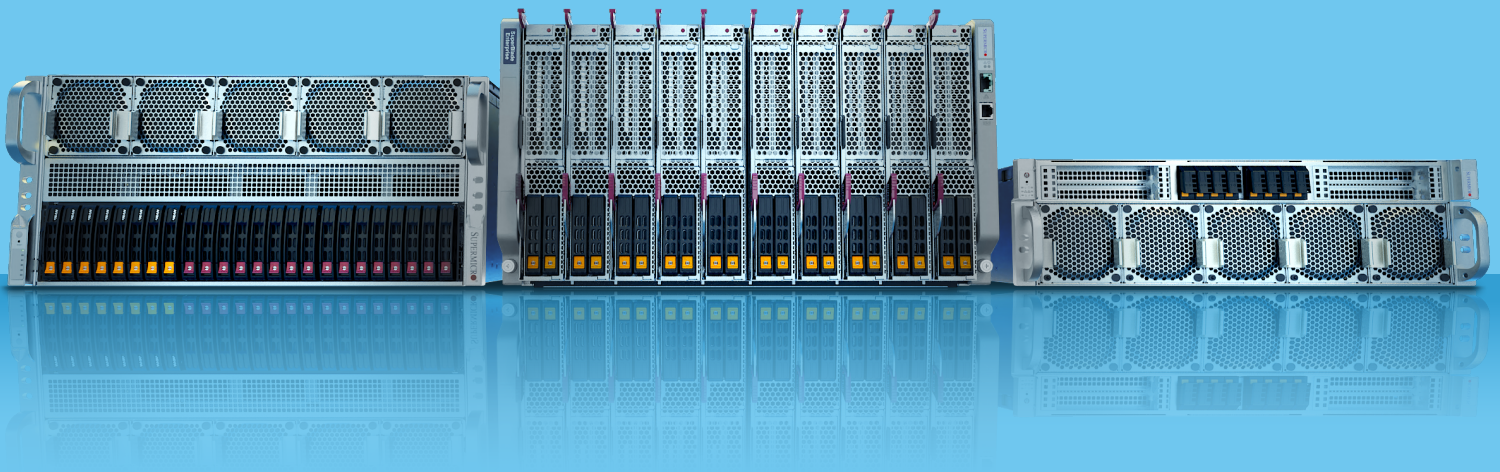# Accelerate Enterprise AI

Selecting Optimal GPU System Configurations to Accelerate Enterprise AI and Visual Workloads

EXECUTIVE SUMMARY

The rapid evolution of AI is enabling enterprises across a wide range of industries to enhance business operations and realize revenue from their data. This white paper highlights Supermicro's industry-leading portfolio of GPU systems and how they can be combined with a range of NVIDIA PCIe GPUs for workload-specific optimization, maximizing ROI and simplifying the adoption of enterprise AI.

Table of Contents

AI is changing what's possible across every industry

The rapid evolution and accelerating adoption of AI is transforming how organizations do business. Thanks to the proliferation of powerful pre-trained models, AI-assisted applications, and full-stack solutions, virtually every organization across every industry is exploring what's possible through greater access to advanced AI capabilities.

Hardware and software innovations are dramatically increasing compute power, energy efficiency, and scalability. These advancements are lowering technical barriers and accelerating the adoption of AI, enabling businesses to streamline processes and create value from their data more effectively than ever.

AI adoption is on the rise as barriers fall

As these technology advancements that power AI become more accessible, organizations are rethinking how they leverage them to deploy AI. Over the next three years, 92% of companies plan to increase their AI investments. However, despite this surge in spending, only 1% of leaders describe their organizations as "mature" in AI deployment—where AI is fully embedded in workflows and consistently delivering significant business outcomes.[1]

In practice, AI strategies vary widely: some organizations train their own models, which demand specialized, large-scale infrastructure, while others rely on fine-tuning pre-trained models, leveraging industry-standard hardware that can be easily integrated into existing facilities. Some organizations face restrictions that limit the use of public, proprietary models or cloud services, requiring them to deploy AI infrastructure on-premises. Others may simply seek to leverage API access to integrate AI into their workflows without needing their own infrastructure. Embracing pre-built, domain-specific models, enterprises are looking for solutions that reduce the complexity of implementation and are more cost-effective, lowering the barriers of AI adoption.

---

[1] Superagency in the workplace: Empowering people to unlock AI's full potential – McKinsey, Jan 2025

Meanwhile, demand is growing for localized, GPU-accelerated infrastructure to support enterprise AI—particularly in industries where responsible AI, privacy, governance, and compliance are critical. Together, these trends highlight a clear enterprise focus: making AI integration scalable, accessible, and low-friction—minimizing complexity, reducing deployment barriers, and managing cost.

Enterprise AI demands a modular approach

As AI continues to develop and its use becomes more widespread in enterprise settings, many businesses are exploring modular, high-performance infrastructure built on accelerated systems that leverage the latest-generation PCIe-based GPUs. PCIe (Peripheral Component Interconnect Express) provides a flexible, accessible, and economical path to AI—supporting enterprise workloads without the need for complex, highly-specialized components—while also allowing organizations to adapt and expand compute environments as needs evolve.

The standardized nature of the PCIe interface allows support for a range of deployment models, from GPU-dense on-premises clusters to distributed edge nodes and even all-in-one development platforms, helping IT teams build solutions that suit the performance needs, power usage, and even the available physical space within their organization. By combining the broad range of flexible Supermicro systems with powerful NVIDIA GPUs, enterprises can customize the configuration of AI infrastructure according to their deployment and workload requirements, regardless of industry, scale, or application.

Enterprise AI in Focus

When discussing enterprise AI, it is important to note the distinction between the broader adoption of artificial intelligence technologies—such as machine learning, natural language processing, and computer vision—and the integration of pre-built AI tools into the core systems, processes, and workflows of a business. Unlike consumer AI, which serves relatively generalized tasks at the individual level, enterprise AI is designed to address specific workloads at scale within an organization, supporting everything from operations and customer service to supply chain management and decision-making.

It is also important to consider the scope of most enterprise AI workloads, which, while substantial, are typically focused on leveraging existing pre-trained models to solve domain-specific problems—unlike large-scale AI research and development that involves training foundation or frontier models with massive computational and data resources. There is a substantial difference in infrastructure needs between an organization focused on developing the next major LLM, the IT team at a retail or banking institution building an interactive chatbot to streamline their customer service, or a design agency leveraging generative AI to help with the generation of graphics or images.

Given their scale and complexity, training large language models requires vast compute power; GPU clusters made up of specialized and therefore costly hardware designed for maximum data processing throughput with enormous training datasets. The physical infrastructure alone is on another level: GPU clusters for foundation model training can occupy entire data center floors and require significant power and cooling infrastructure, while enterprise teams typically work within more space-constrained environments, using limited on-premises setups to fine-tune or deploy existing models for targeted business needs.

August 2025

As enterprises move beyond foundational model training to real-world deployment, the integration of agentic AI and intelligent agents is reshaping how organizations operate day to day. These agents, often built on pre-trained models and embedded into existing workflows, can autonomously complete tasks, interact with users, and adapt based on context—extending the impact of AI beyond analytics into action. Whether used to streamline operations, personalize customer experiences, or support predictive decision-making, agentic AI is helping businesses translate raw data into outcomes faster and more effectively, accelerating the broader shift toward AI-driven transformation.

For many industries, enterprise AI is becoming increasingly vital as it enables businesses to quickly gain valuable insights from vast amounts of data, uncover patterns—such as customer behavior trends, equipment failure signals, or fraud indicators—and automate complex tasks. It enables individuals and teams across the organization—from the C-suite to the facility management team—to analyze data or gather insights in real time, empowering leaders to make faster, more informed decisions. It can improve operational efficiency through the automation of repetitive or manual processes, reducing costs, improving accuracy, and increasing productivity. With the growing use of intelligent agents and personalized, predictive capabilities, businesses can deepen customer engagement, streamline internal workflows, enabling companies to identify and explore new business opportunities, optimize product development, and stay competitive. Across all these scenarios, AI is becoming a core driver of enterprise transformation. It's embedding into every facet of the business, turning every company into an AI company, and powering long-term strategic goals.

New implementations beget new challenges

While these advancements hold immense potential, organizations face significant challenges in effectively implementing new technology for enterprise AI, especially if there may be a secondary requirement for other workloads such as high-performance computing (HPC) or graphics & media. One of the primary hurdles is ensuring the proper infrastructure is implemented and strategically located to integrate these workloads into business processes, an undertaking significant enough to require careful planning in order to maximize return on investment. Planning that must account for the distinct demands of enterprise AI workloads, including the need for efficient systems and pipelines that support data preprocessing and annotation—crucial steps for producing high-quality inputs to fine-tune pre-trained models. Just as important is considering how AI integrates with the existing IT infrastructure. In most enterprise cases, AI is not a standalone environment, but an important element added to established systems, requiring thoughtful alignment with current compute, storage, and networking resources to ensure scalability and efficiency without disrupting operations.

Given these significant hurdles, organizations need a flexible, scalable foundation for AI—one built on the right hardware, appropriately sized and tailored to the environment to support performance, scalability, and seamless integration into enterprise AI workloads. That foundation must also be cost-effective and adaptable, creating the need for a flexible range of AI systems and GPUs that can efficiently meet evolving enterprise demands.

Supermicro Systems with NVIDIA GPUs; Flexibility and Performance for Enterprise AI at Scale

Addressing enterprise AI challenges

Addressing enterprise AI challenges requires more than just powerful infrastructure—it demands alignment between hardware capabilities and the specific operational, environmental, and performance needs of the organization. Efficient, cost-effective AI solutions need scalable performance, flexible system architectures, modular designs, and optimized power and thermal management.

Let's explore how Supermicro systems are able to meet these needs to deliver the optimal balance of performance and efficiency for faster time-to-results.

Scalable Performance

Supermicro designs systems that scale compute power to meet growing enterprise AI and data demands, leveraging the PCIe interconnect for high-bandwidth GPU-to-CPU and GPU-to-GPU connectivity. For maximum processing density in traditional data center environments or even cluster-scale deployments, multi-GPU-optimized systems such as the 5U 8-GPU or 10-GPU SYS-522GA-NRT, 4U NVIDIA MGX™ RTX PRO Server, or the multi-node SuperBlade® platform that can support up to 120 GPUs per rack, along with high-speed networking for scale-out connectivity. In situations where balanced compute and GPU acceleration is required, a range of standard rackmounts such as the 2U MGX systems or Hyper family provide flagship performance and the flexibility to add up to 4 GPUs. At the network edge, where efficiency and thermal performance is paramount, compact and short-depth systems can accommodate 1 or 2 GPUs for edge inferencing. This scalability helps businesses stay competitive by delivering the acceleration needed for their specific AI workloads while also fitting into existing infrastructure with minimal modification.

Flexible System Footprints

Supermicro offers AI infrastructure in a variety of physical sizes suited to different environments. Traditional rackmount systems are available in 2U, 4U, and 5U form factors to easily integrate into existing IT equipment rooms, while for AI inferencing at the edge such as in retail or smart city deployments, systems like the Compact Box PC SYS-E403-14B-FRN2T can bring GPU acceleration closer to the source of the request for a better user experience. For environments where traditional compute infrastructure facilities are not available or practical, workstations like the SYS-532A-W offer data center performance in a portable, under-desk form factor ideal for offices, laboratories, or educational institutions.
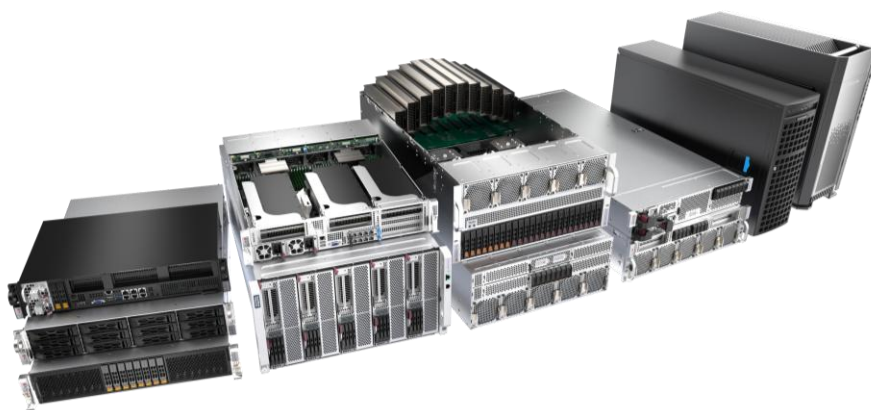
Modular Design

Supermicro's modular Building Block Solutions® architecture provides a high level of flexibility at the system level, allowing for workload-specific customization of system configurations including GPUs, CPUs, RAMs, PCIe switches, storage, networking, cooling, power supplies, and more, according to customers' specific requirements. Not only does this provide the flexibility for enterprises to build systems exactly as they require, but it also means that they are not incurring additional costs for components which they do not require.

Power and Thermal Optimization

Supermicro systems are engineered for optimal thermal and power efficiency, ensuring that high-performance components such as PCIe GPUs are able to perform to their maximum potential. By carefully designing systems to optimize airflow, component placement, and power delivery, Supermicro systems are able to support a wide range of deployment scenarios— from dense edge systems like the 3U SYS-322GA-NR or 2U SYS-212GB-NR that can enable powerful GPU acceleration in edge data center or cabinet environments, to 4U and 5U multi-GPU servers like the SYS-522GA-NR that maximize GPU density per rack without significant modification to existing air-cooled data centers. This attention to system-level cooling and power efficiency helps ensure reliable operation across diverse performance tiers and environments, from small-scale edge deployments to large-scale, high-density AI inference clusters.

PCIe, a Versatile Foundation for Enterprise Systems

Ultimately, Supermicro systems offer a versatile and scalable foundation for enterprise AI, supporting everything from AI inference and model fine-tuning to edge AI, and AI-assisted graphics, media, and transcoding. The ability to design and customize Supermicro PCIe GPU systems to adapt to varying power, thermal, and physical constraints, combined with modular expansion capabilities, make them the ideal building blocks for enterprise AI infrastructure at any scale. As enterprises continue to scale their AI workloads, these systems are poised to meet these evolving demands, offering the performance, flexibility, and reliability needed for success.

Enhancing Enterprise AI Workloads with PCIe GPUs

PCIe is Ideal for Balanced AI Performance

Enterprise AI workloads range from fine-tuning pre-trained models for specific business needs to low-latency inference at edge locations, as well as AI-assisted graphics rendering and video transcoding. PCIe GPUs provide a balanced foundation for these varied demands—delivering sufficient performance for most enterprise use cases without the cost or complexity of larger-scale

systems. Their availability, flexibility, and easy integration into existing infrastructure make them a practical choice for organizations looking to roll out AI initiatives efficiently and reliably.

NVIDIA, a longstanding leader in enterprise AI, has developed a robust ecosystem of hardware and software designed to accelerate and simplify the development, deployment, and scaling of AI applications. By leveraging Supermicro NVIDIA-Certified™ Systems that feature high-performance GPUs including the new RTX PRO 6000 Blackwell Server Edition as well as H200 NVL, L40S, and L4, organizations are able to speed their deployment of AI training, inference, and other data-intensive workloads in both research and production settings.

Introducing Enterprise AI Factory Validated Design

NVIDIA's Enterprise AI Factory validated design provides a full-stack blueprint for building on-premises AI infrastructure. Validated designs are based around NVIDIA-Certified Systems supporting NVIDIA RTX PRO 6000 Blackwell GPUs, ensuring compatibility with NVIDIA Spectrum-X Ethernet networking, NVIDIA BlueField® DPUs, NVIDIA-Certified storage, and NVIDIA AI Enterprise software—enabling enterprises to speed their deployment of complete data center infrastructure in order to begin generating revenue from their AI factories faster. It offers guidance for scaling an enterprise AI factory with RTX PRO Servers, including deployment best practices, to help organizations meet their growing AI business needs efficiently and reliably. With support for NVIDIA NIM microservices and NeMo frameworks, organizations can also accelerate the deployment of digital AI agents and multimodal inference pipelines.

For more information, be sure to read NVIDIA's full announcement here.

Supermicro: Accelerating AI Factory Deployments

To reduce the complexity of building complete Enterprise AI Factories, Supermicro delivers infrastructure that can be easily integrated into NVIDIA's Enterprise AI Factory validated designs. Supermicro's proven ability to be first-to-market when integrating the latest GPU advancements means organizations will be able to take advantage of the powerful new NVIDIA RTX PRO 6000 Blackwell GPUs sooner, so that they can begin generating revenue from their AI factories faster. Supermicro's Data Center Building Block Solutions® (DCBBS) is the ideal platform for collaboration with NVIDIA on Enterprise AI Factory validated designs based on the Blackwell architecture, with a comprehensive approach that includes data center design consultation, solution validation, and professional onsite deployment services, reducing typical deployment timelines from 12-18 months to as little as three months.

Supermicro now offers over 20 systems optimized for the NVIDIA RTX PRO 6000 Blackwell GPUs, which will provide significant performance increases over systems with previous-generation GPUs. These include 5U and 4U rackmounts with up to 8 GPUs per system, MGX-based systems for the data center and the edge—including a new RTX PRO Server design—, high-density SuperBlade multi-node platforms, and tower and rackmount workstations. Supermicro is also collaborating with NVIDIA on NVIDIA-Certified Systems that will serve as foundational building blocks for Enterprise AI Factory deployments.

For more information, be sure to read Supermicro's full announcement here.

Bringing AI to the Edge

Supermicro also offers a number of architectures designed to bring the power of **RTX PRO 6000 Blackwell GPUs** closer to the edge, enhancing the performance of edge AI inference workloads. These compact, efficient systems support up to four GPUs in 2U or 8 GPUs in 3U and are ideal for AI inference, industrial automation, and retail analytics in decentralized environments. The significant performance and memory capacity increases of the RTX PRO 6000 Blackwell architecture compared to previous

generations enable each GPU to process more data, reducing the number of servers required to support advanced AI inference and in turn helping lower the cost and complexity of deploying AI at the edge.

Versatility Across Industries

With configurations ranging from dense rackmount servers to compact edge systems, PCIe-based architectures can be tailored to industry-specific demands. This versatility makes them well-suited for AI applications in sectors like healthcare, finance, manufacturing, industrial automation, retail, and digital content creation—including simulation, rendering, and generative AI.

Let's take a closer look at the NVIDIA GPUs and Supermicro systems that best support enterprise AI:

Workload	RTX PRO 6000 Blackwell (Model size: up to 70B) / L40S (Model size: 13-70B)	H200 (Model size: 70B-175B)	L4 (Model size: Up to 7B)
AI	Text to Image/Video AI Agentic and Generative AI Fine Tune Training/Inference	AI Training LLM Inference & Retrieval-Augmented Generation	Edge AI Inference + Video and AI
HPC	Scientific Computing and Data Analytics	HPC & Data Analytics	-
Media & Graphics	Omniverse AI-driven Rendering and Graphics	-	Mobile Cloud Gaming Virtual Workstation

NVIDIA PCIe GPUs: Powering the AI Factory for Enterprise Transformation

NVIDIA RTX PRO 6000 Blackwell

The NVIDIA RTX PRO™ 6000 Blackwell Server Edition is a universal GPU optimized for AI and visual computing. Built on the Blackwell architecture, it delivers unmatched compute power and memory capacity across multimodal AI inference and physical AI, scientific computing, graphics, and video applications. Featuring 96GB of GDDR7 memory and an optimized combination of NVIDIA Blackwell CUDA cores, fourth–generation RT cores, and fifth-generation Tensor Cores, the NVIDIA RTX PRO 6000 Blackwell Server Edition GPU also introduces native support for FP4 and INT4 data formats—enabling up to 2× faster inference performance by reducing precision while maximizing throughput. Allowing this powerful GPU to handle a broad spectrum of AI, HPC, and graphics workloads in parallel, making it ideal for mixed-workload and AI-assisted applications. In addition, support for multiple concurrent instances via Multi-Instance GPU (MIG) makes it ideal for shared or virtualized deployments.

NVIDIA H200 NVL

The NVIDIA H200 NVL GPU, based on the Hopper architecture, accelerates high-performance AI and HPC workloads—from training and low-latency inference to advanced scientific simulations. In a 2-way configuration, the NVLink Bridge enables up to 900GB/s of GPU-to-GPU communication. With 141GB of HBM3e memory and 4.8TB/s bandwidth, the H200 NVL is ideal for handling large datasets and powering advanced RAG pipelines, all while supporting NVIDIA AI Enterprise for deployment in data-rich, AI-enabled business environments.
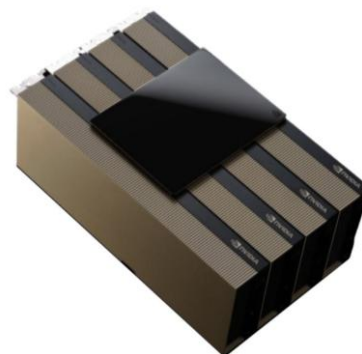
NVIDIA L40S

The NVIDIA L40S GPU, built on the Ada Lovelace architecture, is a versatile accelerator for AI and graphics workloads featuring both fourth-generation Tensor Cores and third-generation RT cores. With high FP8/FP16 throughput and strong media processing, it's well suited for inference, training, graphics, and video workloads. Fully compatible with NVIDIA AI Enterprise, it simplifies AI integration into everyday business applications.

NVIDIA L4

The NVIDIA L4 GPU, built on Ada Lovelace, is a compact, energy-efficient PCIe accelerator ideal for edge deployments and low-power environments. Its 72W single-slot form factor fits easily into workstations and small form-factor servers, ideal for space and thermally-constrained environments such as the edge. With 4th-gen Tensor Cores, 3rd-gen RT Cores, FP8 precision support, and AV1-capable media engines, the L4 delivers efficient inference and strong performance across AI, graphics, and video workloads at low thermal load. With 24GB of GDDR6 memory, the NVIDIA L4 supports a broad range of AI applications including recommendations, voice-based AI avatar assistants, chatbots, visual search, and contact center automation. Fully compatible with NVIDIA AI Enterprise, it enables scalable AI in space- and power-constrained scenarios.
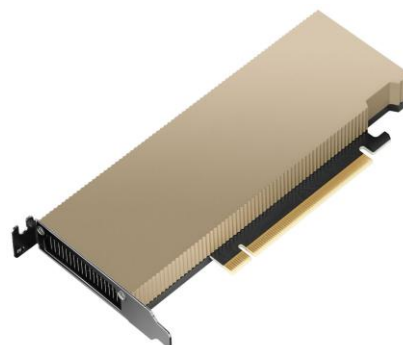
HPC, Graphics, Media—and Enterprise AI Synergy

PCIe-based, GPU-accelerated infrastructure isn't just for AI —it also benefits organizations deploying AI-enhanced tools in media, design, research & development, customer engagement, and operations. The NVIDIA RTX PRO 6000 Blackwell Server Edition offers excellent support for graphic rendering & streaming, cloud gaming, and visual content development, while the NVIDIA H200 delivers unmatched parallelized computing performance for HPC workloads acceleration such as for engineering simulation, scientific research, and big data analytics. For environments where efficiency is paramount, the NVIDIA L4 is perfect for transcoding, virtual workstations, and mobile cloud gaming.

For organizations that frequently handle multiple and/or mixed workloads, the universal nature of GPUs such as the RTX PRO 6000 Blackwell are the obvious choice, offering rounded performance for AI and non-AI workloads. As traditional HPC and

graphics workloads increasingly leverage AI and parallel computing to accelerate time-to-result and efficiency, the ability for a single GPU to handle multi-modal tasks provides significant benefits, both in terms of cost and physical infrastructure footprint.

Next, let's explore how Supermicro's Enterprise AI infrastructure is optimized to support these GPUs and accelerate enterprise AI workloads.

Enterprise AI: Flexible, Scalable Solutions for Enterprise Workloads

Supermicro delivers enterprise-ready infrastructure designed to accelerate time-to-result and time-to-revenue across AI, graphics, and HPC deployments. With industry-leading time-to-market, a global manufacturing presence, and first-in-market support and validation for new NVIDIA GPUs, Supermicro helps organizations deploy enterprise AI infrastructure faster—, speeding the adoption of AI and delivering faster return on investment.

Accelerated Deployment, Flexibility, and Availability

Supermicro's agile R&D ability and close collaboration with partners such as NVIDIA enable it to bring new platforms to market in a matter of weeks rather than months—adapting quickly to emerging hardware and unique customer needs. Supermicro systems are often among the first validated to support next-generation GPUs, providing customers with critical competitive advantages in a rapidly evolving AI landscape. Its NVIDIA-Certified and Qualified systems are guaranteed to work out of the box with the NVIDIA AI Enterprise software stack, helping reduce integration time and accelerate deployment of complete AI solutions. Additionally, deploying Supermicro systems with PCIe GPUs offers clear availability advantages, with shorter lead times compared to more specialized GPU infrastructure. Available certified systems include:

NVIDIA-Certified™ Systems	
RTX PRO 6000 Blackwell Server Edition	**AS -5126GS-TNRT2***, **SYS-522GA-NRT***, **SYS-422GL-NR*****,** **AS -2115HV-TNRT** (Max-Q)* *NVIDIA-Qualified systems. NVIDIA-Certified currently in progress
H200 NVL	**SYS-522GA-NRT**, **SYS-521GE-TNRT**, **AS -5126GS-TNRT2**, **AS -5126GS-TNRT**
L40S	**AS -2115HV-TNRT**, **SYS-521GE-TNRT**, **SYS-741GE-TNRT**, **SYS-421GE-TNRT**, **SYS-421GE-TNRT3**, **SYS-621C-TN12R**, **SYS-221GE-NR**
L4	**SYS-621H-TN12R**, **SYS-741P-TR**

Supermicro also offers a broad range of NVIDIA-Certified systems supporting other GPUs, as well as NVIDIA-Qualified systems. Find out more by visiting Supermicro's NVIDIA-Certified landing page.

Global Delivery & Government Readiness

With manufacturing operations in the U.S., Taiwan, Malaysia, and the Netherlands, Supermicro boasts a global capacity of more than 5,000 racks per month, enabling delivery of complete solutions to customers around the world with shorter lead times. A large US-based manufacturing capacity and Supermicro's Made In The USA (MITU) program combine to produce servers manufactured at the Supermicro San Jose, California factory, which meet the high standards many government agencies require. Supermicro's MITU program gives customers the confidence that the designated servers with the "Made In The USA"

label have been manufactured in the USA. As a result, the highest quality standards are monitored, controlled, and measured. In addition, Supermicro assembly lines are ISO 9001 certified.

Supermicro Systems: Built for AI Workload-Optimized Performance

Large-scale AI Inference, HPC, and Media - Scalable 4U/5U Systems

Supermicro's 4U and 5U GPU systems are engineered for high-throughput, multi-GPU performance across demanding AI, HPC, and media workloads. The 5U SYS-522GA-NRT and AS -5126GS-TNRT2 dual-processor systems are thermally-optimized to support up to 10 double-width GPUs, as well as high-speed networking via NVIDIA BlueField®-3 DPUs and ConnectX®-7 NICs— ideal for AI inference, fine-tuning, simulation, and cloud gaming. Similarly, the 4U SYS-422GL-NR is based on the modular NVIDIA RTX PRO Server design which is optimized for enterprise AI factories, capable of supporting up to 8 double-width GPUs in 4U to accelerate a wide range of enterprise workloads—from agentic AI and LLM inference to industrial AI and digital twins. In the future, this 4U design will also support the new NVIDIA PCIe switch board with ConnextX-8 to accelerate workloads that require significant GPU-GPU connectivity across systems. These systems use standard rackmount chassis with onboard storage and redundant power supplies, enabling seamless integration into existing data center environments and scalable deployment of GPU-accelerated infrastructure.

*Learn more about the 5U SYS-522GA-NRT and AS -5126GS-TNRT2*

Learn more about the 4U SYS-422GL-NR

AI Training & Large-Scale Inference – High-Density Systems

Supermicro's SuperBlade® multi-node platform is optimized for GPU-intensive workloads that demand both high density and scalability—such as professional rendering, industrial simulation, and digital twin applications. Available in 6U and 8U enclosures, supporting up to 20 hot-swappable GPU nodes. SuperBlade can be configured with NVIDIA RTX™ PRO 6000 Blackwell GPUs, enabling up to 120 GPUs per rack to deliver exceptional throughout and efficiency. By integrating shared, redundant components—including advanced cooling, networking, power supplies, and centralized chassis management—SuperBlade dramatically reduces data center footprint while maximizing rack-level GPU utilization. This system design fits seamlessly into existing environments, making it ideal for enterprises looking to scale real-time ray tracing, immersive content creation, CAE/CFD workloads, and AI training pipelines. Designed specifically for workloads that benefit from RTX-class compute, SuperBlade empowers organizations to accelerate design cycle, enhance visual fidelity, and streamline production across large-scale AI-powered pipelines.

Learn more about the SuperBlade SBI-411E-1G, SBI-612BA-1NE34, SBI-612B-1NE34, and SBI-612B-1C2N

August 2025
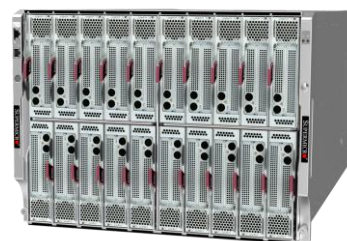
Visualization & Creative Workloads – A+ Workstations and SuperWorkstations

Supermicro's professional workstation portfolio includes both rackmount and tower systems designed for high-performance visualization, virtualization, rendering, and AI-assisted creative workflows. The 2U rackmount AS -2115HV-TNRT can support up to four dual-width GPUs—ideal for centralized content creation, simulation, and enterprise AI development in secure, managed environments. For desktop deployments, the SYS-532AW-C and AS -531AW-TC mid-towers are optimized for RTX PRO 6000 Blackwell Workstation Edition and Max-Q GPUs, delivering powerful performance with thermal optimization for workstation form factors.

Learn more about the [2U rackmount AS -2115HV-TNRT](#)

Learn more about the [SYS-532AW-C mid-tower](#) *and* [AS -531AW-TC mid-tower](#)

Edge AI - IoT SuperServers

Supermicro's edge-optimized and compact form-factor servers, including the SYS-E403-14B-FRN2T Box PC and the high-density SYS-212GB-NR, are purpose-built for deployment in remote or space-constrained environments. These systems are engineered with features like front-accessible power and I/O, and support for low-power GPUs to ensure reliable operation in telecom, retail, manufacturing, and other settings with distributed infrastructure. The SYS-E403-14B-FRN2T offers a fan-based, front-access design in a compact form factor that is ideal for retail AI applications, while the MGX-based SYS-212GB-NR supports up to four NVIDIA RTX PRO 6000 Blackwell GPUs—delivering powerful AI inference in a cost-effective single-processor 2U form factor. Together, they represent Supermicro's ability to deliver scalable, GPU-accelerated computing with integrated networking, storage, and compute in edge-friendly footprints.

Learn more about the [SYS-E403-14B-FRN2T](#)

Learn more about the [SYS-212GB-NR](#)

Conclusion

AI is reshaping what's possible across every industry, and its increasing accessibility is accelerating adoption within the enterprise space. Unlike consumer AI, enterprise AI must operate at scale—deeply embedded into business systems to enhance operations, decision-making, and efficiency. These implementations require a modular, thoughtful approach, and while they don't demand the scale of massive model training farms, they still present real infrastructure challenges. That's where scalable, high-performance solutions become essential.

The combination of Supermicro enterprise AI systems and NVIDIA PCIe GPUs provides unrivalled choice to meet the demands of modern enterprise AI —delivering the flexibility, performance, and scalability needed to integrate AI seamlessly into your operations. **As your organization advances its AI strategy, explore the latest innovations at supermicro.com/pcie-GPU — your future-ready destination for enterprise AI,** with detailed product information, resources and industry expert analysis to help you make an informed decision about enterprise AI infrastructure. Supermicro's enterprise AI platforms featuring NVIDIA PCIe GPUs are designed to accelerate deployment, optimize performance, and position your business for sustained AI-driven innovation.

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs and enables us to build and deliver application-optimized solutions based upon your requirements. See www.supermicro.com.

NVIDIA

NVIDIA accelerated computing platforms power the new era of computing, performing exponentially more work in less time with much lower energy consumption than traditional CPU-based computing. Accelerated computing revolutionizes energy efficiency across industries by harnessing NVIDIA GPUs, CPUs, and networking, all optimized through NVIDIA enterprise software solutions. More information at https://nvidianews.nvidia.com.

August 2025