Performance-Optimized Rack Infrastructure for Scalable AI

Abstract

Artificial intelligence (AI), especially Generative AI (GenAI), is transforming how work is done throughout the enterprise—across all industries and organizations worldwide. AI technology enables new dynamics in the workplace by enhancing business productivity, streamlining workflows, driving innovation, and accelerating time-to-market for CSPs and business end customers.

The ability to provide Scalable AI capabilities will position CSPs and MSPs for the rapid growth of AI and GenAI for their enterprise customers. It will determine which service providers can keep up with the intense competition to deliver scalable AI services to end customers – and which ones cannot.

Supermicro's rack-scale systems, customized to meet each CSP customer's specific requirements, deliver Scalable AI services as demand grows. Supermicro experts install the rack units on-site within days of the initial order. They are designed to expand by adding rack units as the enterprise's demand for AI services grows. As suppliers of scalable AI services, CSPs benefit by reducing their development and power/cooling costs and speeding the deployment of revenue-paying scalable AI services.

Executive Summary

Scalable AI systems are paying real-world business dividends. CSPs and MSPs know that AI systems with GenAI and ChatGPT create dynamic content that makes their customers' businesses run faster and better. That's why they want to scale up AI quickly. However, CSPs face setup costs for configuring and deploying AI while controlling their data center costs for space, power, and cooling.

Building dense clusters of compute, storage and networking requires an integrated system design that supports balanced performance for scalable AI workloads. All of this must be done with energy efficiency to support rapid buildouts for AI modeling, training, and inference. The business results are clear, but the question for CSPs and MSPs is how to get there quickly without driving up their own configuring, installation, and maintenance costs for scalable AI systems.

Based on a CSP customer's orders and technical specifications, Supermicro delivers pre-integrated, pre-tested rack solutions for scalable AI, ready for installation and use. The servers, storage, networking, and switches are all included—per the customer's exact specifications—and are added to industry-standard racks in Supermicro's integration centers. All hardware components are burned in and stress-tested by Supermicro experts before being brought online at the CSP data center site.

The Drivers of AI Adoption

Enterprise customers increasingly view AI as a "must-have" technology that allows companies to identify new business opportunities and generate more revenue and profits by finding the best ways to grow their businesses. Customers are accelerating business growth by entering new market spaces and generating more revenue and profits. Figure 1 shows the transformative benefits of AI.

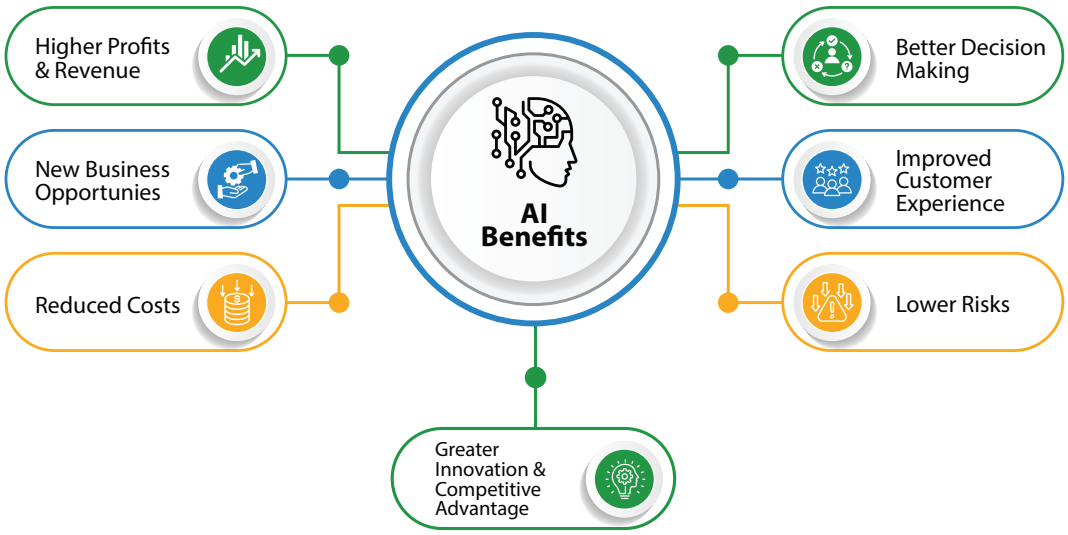

Figure 1: The Benefits of AI

Demand for scalable AI is rising worldwide—and it's been growing rapidly ever since the release of GenAI software tools for developers in Fall 2023. The industry is seeing double-digit growth in many AI sectors, including the exploding use of AI/ML models, large language models (LLMs), and AI inference engines.

However, to access that opportunity, CSPs and MSPs realize that their AI systems need (Figure 2) more powerful, compact, and energy-efficient hardware and software. They need to be able to support the latest technology in GPUs, storage, networking, and software, but they may not have the time or in-house skillsets to spend months re-architecting their scalable AI systems.

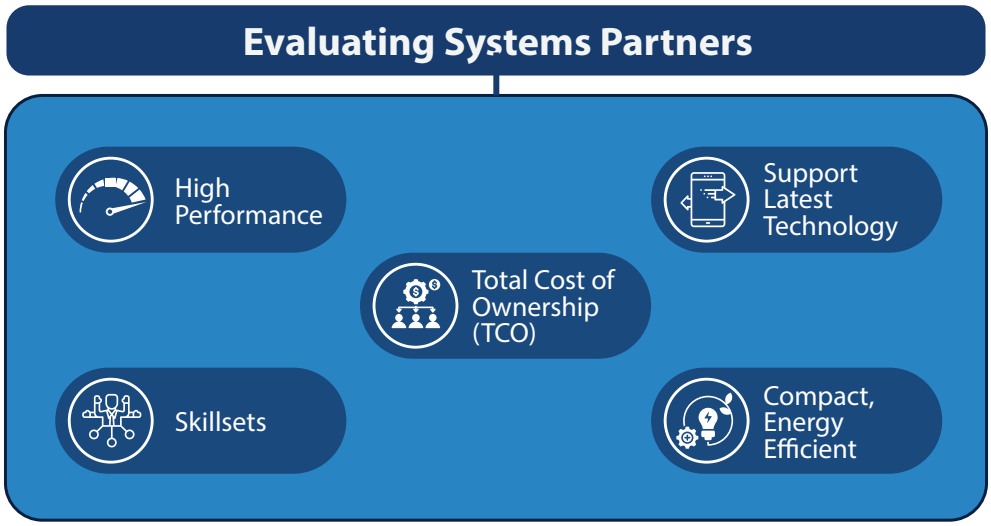

Figure 2: What CSPs need from AI Systems providers

Supermicro's pre-integrated racks are customized for Scalable AI customer requirements, supporting LLMs for data modeling at CSP data center sites. They bring a customer's scalable AI solutions together very quickly – within days of the CSP customer's initial order. These rack units stay on the technology curve as new GPUs, memory, storage, and networking become available in the marketplace.

By contrast, most AI inference workloads run locally – in distributed Edge sites close to where the data is generated –-- to find patterns in enterprise data that disclose important business trends. Examples for inferencing at local enterprise sites include retail stores, banks, factories and oil/gas exploration sites. Inferencing allows CSP customers to gain actionable insights, adjusting their business models to generate more revenue and profits.

"CSPs looking to expand their AI services may not have all the IT skill sets they need to build, deploy and scalable AI systems. They need to reduce the time to deployment and their time-to-revenue for new technology investments. They must supply AI-enabled hardware/software infrastructure quickly to "ride the AI wave" to grow their AI-services revenue."

Delivering Scalable AI

Here are the key factors (Figure 3) for delivering Scalable AI:

- **Rapid Infrastructure deployments to expand scalable AI services to business customers.** CSPs and MSPs must provide reliable, efficient AI services to their business customers as they scale up their AI capabilities. CSPs and MSPs must configure and install AI infrastructure quickly – and at affordable price points. Using pre-tested rack-ready units – which can be quickly added to adjacent racks and linked together –reduces preparation time and speeds time-to-revenue for CSPs and MSPs that deliver AI services to business customers.
- **Pre-built, pre-tested infrastructure for Scalable AI.** Pre-tested infrastructure is installed without special, one-off configurations. It offers CSPs and MSPs scalability, cost efficiency, flexibility, global reach, security, disaster recovery/business continuity, and technological advancements.

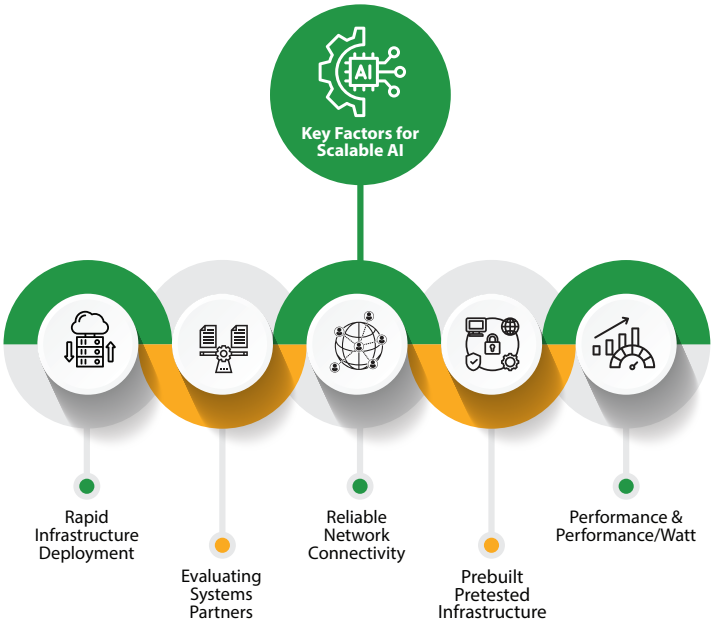

Figure 3: Key factors for Scalable AI

- **Performance and Performance/Watt.** Scalable AI often requires dense clusters of systems with closely packed GPUs and related components. Supermicro is known for its inside-the-rack liquid-cooling systems, which are offered as an option for CSP customers who want to lower their data center PUE. This allows CSPs to support more AI performance while reducing the power/energy envelope in the data center. Energy efficiency is a top priority for budgetary and ESG/environmental reasons. Performance-per-watt is another key indicator of energy efficiency for CSP-delivered AI services.
- **Reliable network connectivity.** Delayed data transmission is problematic for Scalable AI deployments. Reliable network connectivity in scalable AI infrastructure is critical for high-speed communications between multiple rack-unit enclosures, reaching the systems populating the rack. Systems providers such as Supermicro must have deep partnerships with networking providers.
- **Evaluating systems partners.** Scalable AI often results in high costs for compute resources, storage, data transfer, and third-party services. Vendors that provide easy-to-install systems reduce the "friction" that could otherwise be associated with installing Scalable AI. More efficient buildouts of AI systems lead to more efficient provisioning of CSP services – an important ingredient of successful cloud-based AI deployments.

Two Ways of Working with AI: LLMs and Inference Engines

Business customers can choose how they will be consuming AI.

Depending on the customer's AI workload and applications, large language models (LLMs) and inference engines – reflecting both forms of AI for enterprises – can be supported at customer sites. LLMs and inference engines are both used by enterprises – but for different reasons.

- **Large Language Models (LLMs).** LLMs are based on extensive AI modeling and training data. Many of these large LLMs were originally developed by hyperscalers (e.g., AWS, Azure, or Google Cloud), while others originated at tech companies or were developed with open-source software.
- **RAG (Retrieval Augmented Generation).** Now, a way to customize large LLMs with an enterprise's proprietary data exists. Using RAG (Retrieval Augmented Generation) tunes the general-purpose LLMs, making them more responsive to customers in specific industry segments. For example, RAG can be used for proprietary data associated with modeling the motions of self-driving cars or a factory's manufacturing processes. This data is well-described and well-documented – and, therefore, known to the enterprise customer. RAG technology has a major benefit for enterprise end-user inquiries: matching the LLM resources to specific and well-described data resources that enterprise customers use to support GenAI general inquiries. RAG offers an essential safeguard against incorrect answers based on large volumes of data from large general-purpose LLMs. RAG provides specific contexts for enterprise user queries that run against general-purpose LLMs. This provides authoritative knowledge resources based on a customer's proprietary data and will give better responses to some of the LLM queries made by enterprise customers.

- **Inferencing.** By contrast, inferencing is often used in "close to the customer" sites, including edge sites, which require fewer CPUs and GPUs and less storage per site (e.g., modeling the motions of self-driving cars or manufacturing processes in a factory). Inference engines are often hosted on smaller systems located at the customer sites.
- **Enterprises use both types of AI.** This paper discusses both types of AI services: AI/ML modeling/training with LLMs running in the data center and AI inferencing running on Edge servers. A CSP provides both to its enterprise customers, depending on their application needs and data resources.

> "Service providers must adapt to change in infrastructure and be able to upgrade quickly, as processor and storage devices are being continually upgraded with new technology. The ability to adapt to those changes while controlling power/cooling costs will be a differentiator for CSPs and MSPs in this market space."

Looking ahead, the entire service-provider space is being reinvented to support next-wave AI systems for the data center, the cloud, and the Edge. Service providers must adapt to changes in infrastructure and be able to upgrade quickly, as processor and storage devices are continually upgraded with new technology. The ability to adapt to those changes while controlling power/cooling costs will be a differentiator for CSPs and MSPs in this market space.

CSPs and MSPs should have their best practices (Figure 4) to overcome these build-out and deployment challenges – reducing the cost of doing business with end-customers leveraging AI and GenAI. These include:

- **Planning for growth in AI Services.** Power/cooling efficiency and reduced Opex (operational expenses) are competitive advantages in an already-crowded business environment for cloud service providers – when labor, data-center real estate, and energy resources are key factors in competing in the CSP world.

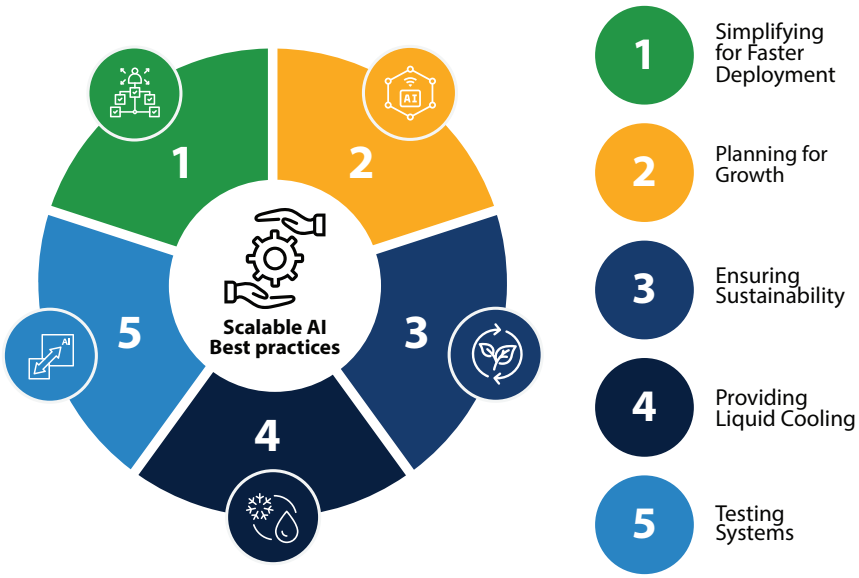

Figure 4: Critical Best Practices for Deploying Scalable AI

- **Simplifying Scalable AI Systems for Faster Deployments.** CSPs delivering Scalable AI will look for ease of ordering, configuration, setup, and fast installation and deployment. Reducing the need for complex one-off configurations, using simplified infrastructure for Scalable AI speeds up order fulfillment, speeds systems deployment, and reduces power/cooling costs by as much as 50%, compared with less optimized AI systems.
- **Testing Scalable AI systems for performance, reliability, capacity, security, and power/cooling.** Testing and Proofs of Concept (PoCs) for new deployments will be critical to installing and maintaining scalable infrastructure for fast-growing AI and GenAI services.
- **Providing liquid cooling for data center racks.** Liquid cooling based on water or chemical solvents keeps data center installations from exceeding temperature limits to comply with environmental regulations set by countries and regions worldwide. Other in-rack cooling options include cold plate technology—on top of processors—and rear-door heat exchangers (RDHx) that guide heat away from the system.

- **Ensuring Sustainability is one of the most important factors for cloud service providers.** Supporting environmental limits imposed by the need for compliance with environmental legislation is vital to AI deployments in many countries and regions worldwide. Exceeding these compliance limits can also be costly. European (E.E.U.) regulations for environmental protection have been a leading example of "how to" reduce environmental damage from carbon emissions—a model for protecting localities from the damage of climate change that is being replicated in other geographic regions worldwide.[1]

CSPs must partner with the right infrastructure solutions provider, such as Supermicro, to grow their CSP business by adopting these best practices.

Key Considerations for CSPs and MSPs

CSPs should consider partnering with infrastructure providers, such as Supermicro, that can help them grow their CSP business by adopting these best practices for Scalable AI (Figure 5):

- **Selecting a partner** that can provide a total infrastructure solution including systems, servers, edge, networks, storage, software, and services) with comprehensive global support.
- **Adopting a "rack-ready" approach to Scalable AI** supports buildouts, allowing CSPs and MSPs to grow AI services quickly and easily. It provides energy-efficient, cost-effective solutions that reduce the cost of expanding a CSP's AI services business by lowering onsite operational costs.
- **Acquiring rack-ready systems** for scalable AI reduces configuration times for CSPs and MSPs and speeds the delivery of AI-enabled cloud services to business customers. Rapid buildouts and rack-level expandability that support the latest processor roadmaps and technology updates allow CSPs and MSPs to grow AI services quickly and easily.
- **Deploying integrated systems reduces configuration, deployment, and maintenance costs.** Acquiring AI-ready systems by the rack—with the servers, storage, switches, and software already integrated—reduces CSPs' overhead costs for AI infrastructure and helps them gain traction for Scalable AI cloud services. Supermicro delivers this important differentiator for business customers, leveraging their firm's corporate data to become more competitive and agile in their industry.

Figure 5: Key Considerations to Deploy Best Practices for Scalable AI

Benefits of Supermicro Systems

Supermicro provides the entire end-to-end infrastructure optimized for CSPs and MSPs, providing scalable AI services to business customers. Rapid deployment of well-integrated, rack-ready systems in the service provider's infrastructure accelerates services delivery to business customers.

Here are the key things about Supermicro's approach to rack-scale, rack-ready systems that CSPs and MSPs can deploy in the data center.

These rack-ready systems are designed – and tested – to support scalable AI services for business end-customers. Responding to the exact requirements of the CSP customer's order, Supermicro's technical experts work directly on rack design, rack installation, and rack-testing,

Supermicro has four rack-integration centers worldwide: in the United States (San Jose, CA), Malaysia, Taiwan, and the Netherlands. These locations have high-skilled rack-integration expertise, allowing Supermicro to ship AI systems quickly to local customers across regions and stay up-to-date with the latest technology as it comes to market.

Importantly, Supermicro has optimized its global supply chain for configuring and delivering rack-ready systems and enclosures – with final delivery at the customer site within weeks of customer order. As a result, customers will see several benefits (Figure 6):

- **Rapid deployment of well-integrated, rack-ready systems.** Delivering these rack-ready systems as pre-integrated units is a touchstone of Supermicro's offers for CSPs and MSPs. This accelerates the delivery of AI-enabled services to end customers. Supermicro's speed and accuracy of Rack-scale configurations reduce the time and cost of system setup for CSPs.
- **Pre-testing systems for Scalable AI.** Before they are delivered to CSP customers, pre-testing systems reduces time-to-market for new services and increases CSP revenue from the fast-growing AI opportunities emerging in the marketplace.
- **Benefiting from technological advances.** Supermicro's AI solutions offer a wide range of system choices, including racks, servers, storage, software, and services that support AI – reaching from the Edge to the Cloud.

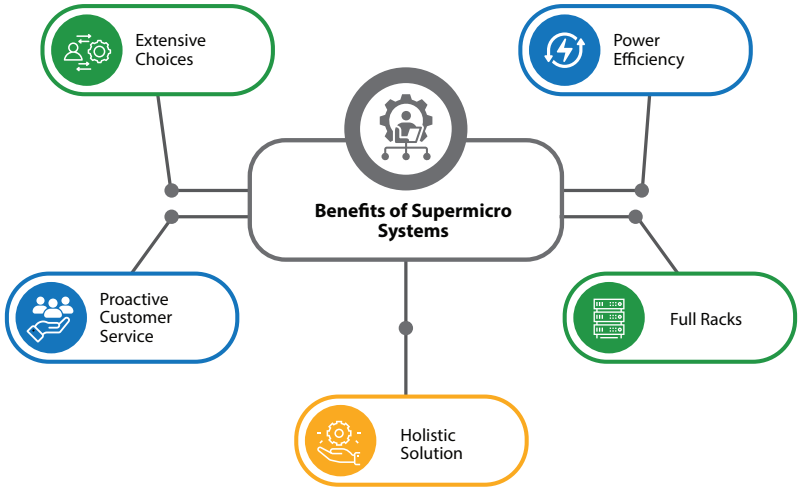

Figure 6: Benefits of Supermicro Systems for Scalable AI

Working with Supermicro reduces the costs of configuration and setup – while speeding the delivery of new and expanded services to enterprise business customers. Taking a rack-scale approach to Scalable AI reduces ongoing operational costs for CSP businesses that adopt capital-cost-effective Supermicro rack-scale systems using AMD processors.

Benefits of Supermicro Systems

The AI and GenAI computing models are hybrid, with training done in the data center on racks and inference done at the edge. AMD processors and accelerators power Supermicro systems to enable Scalable AI from the Edge to the Cloud (Figure 7).

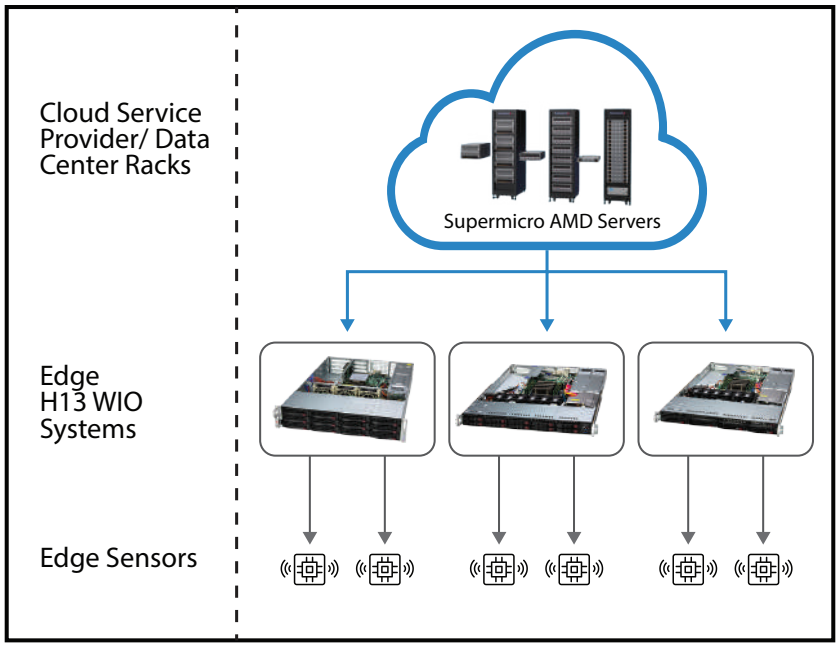

Figure 7: AMD-Powered Supermicro Systems for Scalable AI from the Edge to the Cloud

AMD EPYC™ PROCESSORS

If organizations aren't careful, massive amounts of data can result in high energy use. Efficiency is key for data-intensive applications, and AMD EPYC™ processors power the most energy-efficient servers available. In addition to delivering overall better performance per watt, AMD EPYC™ processors can closely match CPU resources with application requirements, creating even greater efficiency.

For example, some analytic applications do not scale well to high core counts.

Using high-frequency AMD EPYC™ processors can increase per-core performance to speed up these applications without the burden of carrying additional cores not essential to the mission. Some technical computing applications operate best when processors are equipped with large L3 caches. AMD EPYC™ processors with AMD 3D V-Cache™ technology free CPUs to process data with fewer cache misses and therefore unimpeded performance.

Whether an enterprise needs as few as 8 or as many as 128 cores, or specialized processors, AMD EPYC™ processors offer the freedom to choose. All core features—including memory capacity, I/O bandwidth, and security features—are consistent within each processor family.

In addition, there are multiple AMD options for scalability, including AMD EPYC™ 9004 and AMD EPYC™ 8004 processors, AMD Instinct™ MI300X AI accelerators, and AMD Instinct™ MI300A CPU/GPU integrated processors. Supermicro offers options for Scalable AI with 8U, 4U, and 2U modules[2] in rack-optimized enclosures.

- **8U systems for large AI clusters.** For example, customers can use an 8U, 8 GPU AMD-based AI system with AMD Instinct™ MI300X accelerators for the largest AI workloads, reducing lead time for deploying powerful AI systems.
- **2U and 4U systems.** Another way to deploy scalable AI is to select 2U 4-way systems supporting the AMD Instinct™ MI300A APUs or 4U liquid-cooled systems with AMD instinct™ MI300A APUs. These options can be combined inside rack-optimized, rack-ready enclosures for rapidly deploying scalable AI systems.
- **AMD Instinct™ MI300A Accelerators.** These 2U and 4U systems can be used singly or in combination in a rack enclosure. A single system can have up to 4 AMD Instinct™ MI300A accelerators in a single rack.
- **Direct access to high-speed networking cards**. Rack units can have up to 8 high-speed 400G networking cards in a cluster, providing maximum networking bandwidth for each CPU processor and GPU in a scalable AI rack unit.
- **Air-cooling and liquid-cooling.** As in all AI clusters, the density of the CPU and GPU processors within a given AI enclosure generates heat. Customers can use air-cooled or liquid-cooled system configurations that draw the heat away from the central processing units and GPUs, supporting large AI deployments while reducing cooling costs.

Figure 8: Supermicro Systems and Storage optimized for Scalable AI

Supermicro provides Supermicro systems and Supermicro storage optimized for Scalable AI (Figure 8), including the following models:

- ASG-2115S-NE332R **Petascale Storage Server** offers super-fast data processing, extensive storage with 32 hot-swap E3.S-drives, expandable memory for demanding tasks, and redundant power supplies for uninterrupted operation.
- ASG-2015S-E1CR24H delivers flexible storage with 24 hot-swap drive bays, fast and reliable performance, easy upgrades with multiple expansion slots, and convenient remote control and monitoring.
- AS -1115HS-TNR saves space with a 1U form factor, ensures fast computing, offers scalable storage, and simplifies control with advanced remote management.
- AS -8125GS-TNMR2 saves space with a 1U form factor, ensures fast computing, offers scalable storage, and simplifies control with advanced remote management.

The Supermicro Advantage for Scalable AI

Supermicro's approach to building efficient rack-scale, ready-to-go systems—including careful system testing and vendor service – supports growth in the important enterprise sector.

The company's Scalable AI philosophy and rack-optimized systems benefit CSP and MSP customers that deliver leading-edge scalable AI services to enterprise business customers. Specifically:

- Supermicro offers a wide range of system choices, including racks, servers, storage, software, and services that support scalable AI – reaching from the Edge to the Cloud.
- Working with Supermicro reduces the costs of configuration and setup – while speeding the delivery of new and expanded services to enterprise business customers.
- Taking a rack-integrated approach to Scalable AI reduces ongoing operational costs for CSP businesses that adopt capital-cost-effective Supermicro rack-scale systems using high-performance AMD EPYC processors.

Call to Action

For more information about Scalable AI, and how to get started with rack-enabled Scalable AI, please visit our website (supermicro.com/aplus) or contact Supermicro AI experts directly to schedule a meeting.

---

[1]For more details, see: "Top 10 Best Practices for a Green Data Center" paper, which is already posted on the CSP website (https://www.supermicro.com/en/white-paper/datacenter-report).]

[2]https://www.supermicro.com/en/accelerators/amd