WHITE PAPER

# SELECTING THE RIGHT AI INFRASTRUCTURE IMPERATIVE TO INNOVATION AND GROWTHS

*A Comprehensive Guide to Supermicro Systems with NVIDIA GH200 Grace Hopper™, NVIDIA H200 Tensor Core GPU and NVIDIA H100 Tensor Core GPU*

## Table of Contents

## Introduction

The integration of AI is no longer just a competitive advantage—it's a necessity for businesses looking to streamline operations, unlock new growth opportunities, and personalize customer experiences at scale.

According to IDC, worldwide total IT spending for AI technologies is projected to approach $750 billion by 2028 at CAGR of 32.8%[1], a testament to the growing reliance on AI technologies for competitive advantage. The need for robust, GPU-enabled infrastructure has become paramount, with businesses striving to manage vast datasets, process complex algorithms, and deliver real-time insights.

Certain industries are massively accelerating AI adoption at scale, leveraging its power to transform areas of fraud detection, energy efficiency, and scientific breakthroughs. Given the nature of some of the industries in focus, they require immense computational capabilities to process large datasets, run complex simulations, and derive insights in real time. The demand for high-performance AI solutions has never been greater, and ensuring organizations have the right infrastructure is key to fuel innovation and solve critical global challenges.

This whitepaper aims to guide organizations in selecting the most suitable Supermicro systems powered by NVIDIA GPUs, helping them balance performance, scalability, and cost to meet their unique needs and market challenges.

[1] Worldwide Artificial Intelligence IT Spending Forecast 2024-2028, IDC, Oct. 2024

## Accelerating AI adoption with NVIDIA and Supermicro: Best-in-class technology for optimal performance

While the adoption of AI is accelerating, selecting the right AI infrastructure is easier said than done. With a variety of systems and architectures available, organizations must identify solutions that align with their unique demands and budget constraints. To meet current and future demands, organizations must balance performance, scalability, and cost when selecting their AI infrastructure, and aligning the right infrastructure to your key workloads is critical. Understanding how these core workloads and infrastructure choices align with critical applications and selecting the right systems for the job can enhance performance, efficiency, and the long-term success of these solutions.

This whitepaper will focus on Supermicro systems purpose built for supporting AI adoption and expansion. Each system we discuss is powered by a combination of NVIDIA's top-of-the line NVIDIA GH200 Grace Hopper Superchip, the NVIDIA H200 Tensor Core GPU, or the NVIDIA H100 Tensor Core GPU, all of which are purpose-built for AI infrastructure and workloads.

### When to choose Supermicro systems featuring NVIDIA's GH200 Grace Hopper Superchip

For workloads that require high compute power, extensive memory capacity, and seamless data sharing—leveraging a unified CPU+GPU memory model architecture can enable faster data access and improved efficiency for these data-intensive tasks. Additionally, the NVIDIA GH200 delivers outstanding performance in LLM, RAGs, GNNs, HPC, making it ideal for scenarios that require real-time insights and large-scale data processing.

Integrated with Supermicro's infrastructure, the NVIDIA GH200 Grace Hopper Superchip offers transformative performance for AI and HPC workloads. By combining an ARM-based* CPU with GPU-accelerated computing and high-bandwidth memory, it provides AI-driven applications and similar workloads the power and efficiency they need to perform at scale.

### When to choose Supermicro systems featuring NVIDIA's H200 and H100 Tensor Core GPUs

In environments where robust computing power is essential for managing vast datasets, executing complex algorithms, and enabling real-time data processing, hardware designed for parallel processing, high GPU density, and advanced tensor core acceleration plays a critical role. These features are vital for reducing training times and enhancing model accuracy in tasks such as deep learning model training and high-performance simulations.

The **NVIDIA H100 Tensor Core GPU** enables an order-of-magnitude leap for large-scale AI and HPC with extraordinary performance, scalability, and security for every data center and includes the NVIDIA AI Enterprise software suite to streamline AI development and deployment. With NVIDIA fourth generation NVLink™, H100 accelerates exascale workloads with a dedicated Transformer Engine for trillion parameter language models. For small jobs, H100 can be partitioned down to right-sized Multi-Instance GPU (MIG) partitions. With Hopper Confidential Computing, this scalable compute power can secure sensitive applications on shared data center infrastructure.

The **NVIDIA H200 Tensor Core GPU** builds upon the groundbreaking capabilities of the H100, offering everything the H100 delivers and more. With features such as Multi-Instance GPU (MIG), NVLink®, and the Transformer Engine, the H200 ensures the same unparalleled versatility for AI and HPC workloads while pushing boundaries further. The biggest leap forward lies in the HBM3e memory—faster and larger than its predecessor—fueling the acceleration of generative AI, Large Language Models (LLMs), and scientific computing for HPC workloads. NVIDIA HGX™ H200 delivers record breaking performance—an eight-way HGX H200 delivers over 32 petaflops of FP8 deep learning compute and 1.1 terabytes (TB) of aggregate high-bandwidth memory, setting a new benchmark for generative AI and HPC applications.

Combined with Supermicro's highly flexible building block architecture for AI infrastructure, these NVIDIA accelerated computing platforms deliver exceptional performance for AI and high-performance computing workloads. With advanced GPU-GPU interconnectivity, high GPU compute density per system and rack, high scalability, and optimized NVIDIA AI Enterprise software libraries, frameworks and toolsets, these systems are designed to accelerate deep learning model training, large-scale simulations, and data analysis, all within scalable, energy-efficient systems that can be tailored to specific workloads. Supermicro offers system options for both the NVIDIA HGX and NVIDIA PCIe (Peripheral Component Interconnect Express) GPUs, allowing organizations to easily adopt and scale based on existing infrastructure.

By identifying these workloads and the systems optimized for them, organizations can make more informed decisions on the right systems with the right GPUs for the job, ultimately reducing resource expenditure and accelerating the deployment of the right AI solutions.

**The following sections will take a more detailed look into the Supermicro systems leveraging each of these NVIDIA platforms, technical details and workloads suited for each system.**

> *"The performance on the Grace CPU is better than many of the other processors out right now and competes very well for general processing and really takes off with AI workloads."*
>
> **– Christopher M. Sullivan | Director - Research and Academic Computing,**
> **College of Earth, Ocean, and Atmospheric Sciences, Oregon State University**

** NVIDIA Grace systems are Arm system ready SR Certified. The GH200 combines NVIDIA's Hopper GPU with an Arm-based Grace CPU, creating a unified architecture that is highly efficient for large-scale AI workloads and high-performance computing.*

## NVIDIA GH200 Grace Hopper Superchip

### About

The NVIDIA GH200 Grace Hopper Superchip's breakthrough design forms a high-bandwidth connection between the NVIDIA Grace™ CPU and Hopper™ GPU to enable the era of accelerated computing and generative AI. GH200 delivers up to 10X higher performance compared to the NVIDIA A100 Tensor Core GPU for applications running terabytes of data, helping scientists and researchers reach unprecedented solutions for the world's most complex problems. It is available in two memory configurations, 96GB of HBM3 or 144GB of HBM3e combined with 480GB of integrated LPDDR5X , allowing customers to choose the capacity that best fits their workload needs, whether for AI model inference, data analytics, or advanced simulations.

**How it works:** The heart of the NVIDIA GH200 Grace Hopper Superchip is the NVIDIA® NVLink®-C2C interconnect. C2C is a memory-coherent, high-bandwidth, and low-latency interconnect for superchips. NVLink-C2C enables applications to oversubscribe the GPU's memory and directly utilize the Grace CPU's memory at high bandwidth.

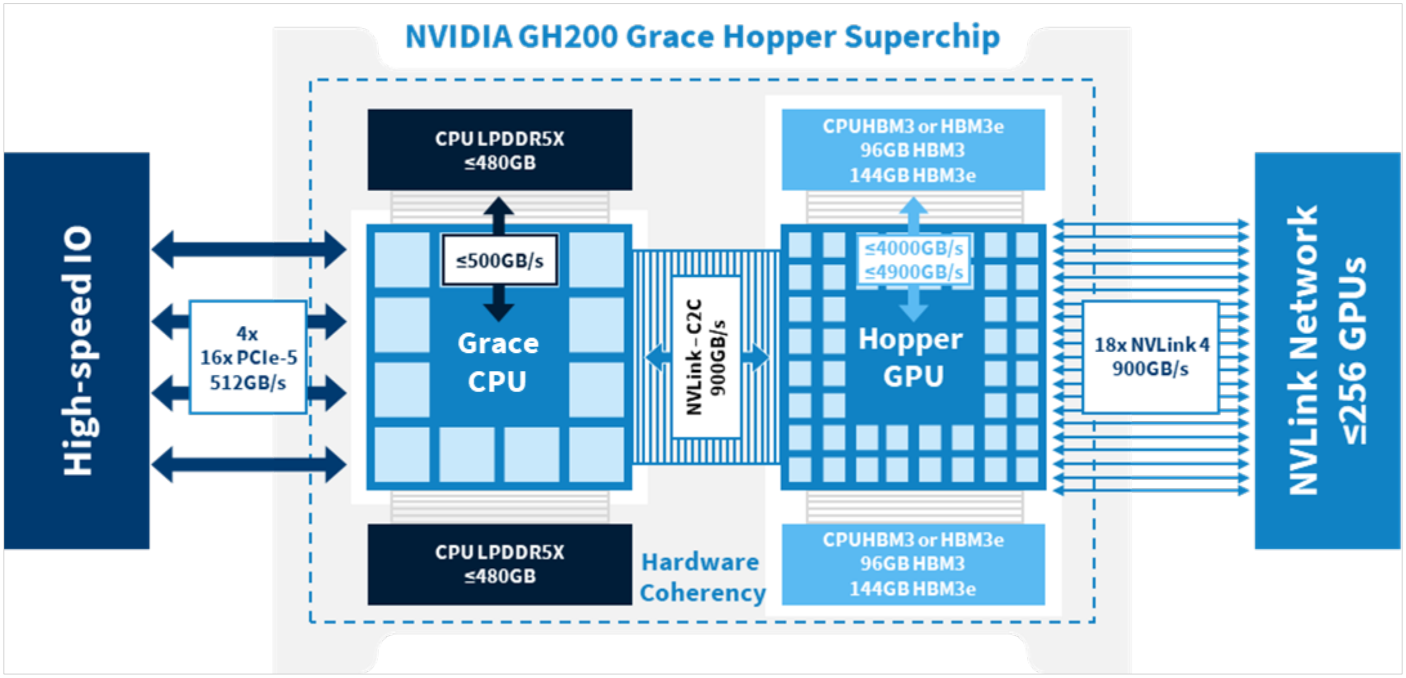
**Performance:** With 900 gigabytes per second (GB/s) of coherent interface, NVLink-C2C is 7X faster than PCIe 5.0. And with HBM3 and HBM3e GPU memory, it supercharges accelerated computing and generative AI. The NVIDIA GH200 runs all NVIDIA software stacks and platforms, including NVIDIA AI Enterprise, the HPC SDK, and Omniverse™. The NVIDIA GH200 is designed to accelerate applications with exceptionally large memory footprints, larger than the capacity of the HBM3 / HBM3e or LPDDR5X memory.

**Scalability:** The unified CPU+GPU architecture and physical design of the NVIDIA GH200 provide one of the more scalable solutions on the market. Its high bandwidth memory and advanced parallel processing capabilities efficiently manage complex tasks notably, real-time simulations and large-scale model inference while reducing latency. The NVIDIA GH200's streamlined design simplifies deployment and scaling, enabling organizations to effortlessly expand their infrastructure to meet growing demands. The shared memory capacity of up to 624GB (144GB HBM3e + 480GB LPDDR5X) significantly improves performance by removing bottlenecks between the CPU and GPU, allowing faster data throughput and enhanced efficiency.

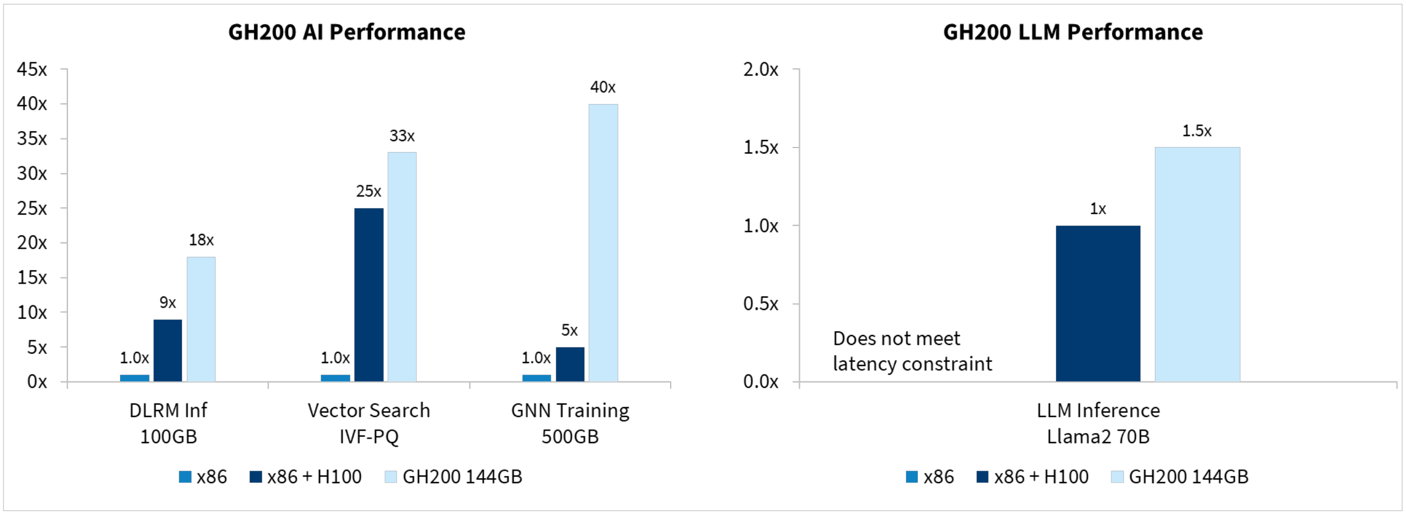
**Cost efficiency and energy optimization:** The shared memory architecture of the NVIDIA GH200 eliminates bottlenecks between the processor and GPU, which results in better performance for memory-intensive workloads. Additionally, the linear scaling of Thermal Design Power (TDP) and energy usage for the Grace CPU proves highly advantageous for large data center customers, delivering lower energy costs when not operating at full capacity. This flexibility allows organizations to optimize energy consumption and reduce operational expenses, especially in scenarios where workloads fluctuate.

> *"There was no lift for us to get Grace Hopper deployed. It fit into SLURM like it was candy. It deployed on the software stacks like it was nothing. It was less than three hours before it was deployed. It deploys like a dream come true. Furthermore, I can scale as I need. I need another one; buy another one, plug it in. I just keep adding; all I had to have in between is the NDR switch."*

– **Christopher M. Sullivan | Director - Research and Academic Computing, College of Earth, Ocean, and Atmospheric Sciences, Oregon State University**



Source: [NVIDIA GH200 Superchip Data Sheet]()



*Comparison: 2S Xeon Platinum 8480+, Xeon Platinum 8480+ and H100 Tensor Core GPU, and NVIDIA GH200 144GB: DLRM 100GB inference, vector search (batch size = 10,000 | queries = 10,000 over 85M vectors IVF-PQ) GNN (GraphSage OGB-100M papers dataset), Llama2 70B (batch size = 64 (GH200), 96 (2x H100s) | precision = FP8 | TensorRT-LLM - throughput per GPU).

Results subject to change.

Source: [NVIDIA GH200 Superchip Data Sheet]()

## Workloads powered by Supermicro systems with the NVIDIA GH200 Grace Hopper Superchip

In this section we will look at the workloads suited for compute-heavy applications and the ***Supermicro systems*** that enable those workloads.

The architecture for compute-heavy applications is optimized for demanding tasks, making it ideal for deep learning inference and recommendation engines that require rapid processing and vast data handling. With support for high-bandwidth data transmission and substantial memory capacity, this system is well-equipped to tackle applications that rely on extensive datasets and require real-time or near-real-time processing capabilities.

Supermicro systems with the NVIDIA GH200 harness the power of NVIDIA's advanced compute infrastructure to drive exceptional performance in data-intensive workloads. These systems are engineered to meet the demands of high-performance computing and AI-driven applications, offering seamless scalability and efficiency across diverse workloads and industries.

Here's a detailed look at the relevant workloads and Supermicro systems supporting them:

1. **Inference for Large Language Models (LLM):** The NVIDIA GH200 significantly enhances inference for large language models (LLMs) with high CPU-to-GPU bandwidth and HBM3/HBM3e memory. It efficiently meets the growing memory demands of LLMs, especially with larger batch sizes, by utilizing NVIDIA NVLink-C2C for fast access to LPDDR5X system memory. This reduces tensor offloading times and boosts performance, achieving 4.5x throughput over PCIe solutions at batch size 4. Overall, the NVIDIA GH200 enables double[2] the inference speed for GPT-3 compared to earlier architectures, making it ideal for next-gen LLM applications. The NVIDIA GH200 facilitates Retrieval-Augmented Generation (RAG), which combines LLMs with external knowledge sources to enhance response accuracy. Applications such as copilot utilize this capability to assist developers by generating context-aware code snippets. Furthermore, the NVIDIA GH200's strengths in fast and low-latency summarization and text generation through generative AI enable efficient synthesis of information, making it valuable for content creation and automated reporting.

   [2] A server cluster with 32 GPUs using NVIDIA GH200 NVL32 will deliver 2x faster GPT-3 model inference performance compared to a 32 GPU cluster of HGX H100 accelerated servers with 8-way NVLink (NVL8) and with Ethernet inter-node connections.

2. **Databases and big data analytics:** Databases manage vast amounts of data, often exceeding the memory capacity of traditional GPUs. The NVIDIA GH200 overcomes this limitation by enabling high-speed access to datasets stored in CPU memory, significantly improving performance and reducing bottlenecks caused by slow PCIe connections. With Grace Hopper's NVLink-C2C and Address Translation Service (ATS), GPUs can work directly with large datasets, enhancing the efficiency of database operations, analytics, and machine learning tasks. This integration allows for concurrent processing and easier access to memory, making it possible to handle large-scale databases seamlessly and effectively. Additionally, the NVIDIA GH200 is optimized for big data analytics frameworks like Apache Spark, enabling faster data processing, real-time analytics, and efficient in-memory computing. Industries such as healthcare, finance, telecommunications, and e-commerce can leverage these advancements for improved data management, big data analytics, and machine learning capabilities.

### a) Supermicro SuperServer ARS-221GL-NHIR enabling these workloads

The Next Gen 2U 1 Node System, model ARS-221GL-NHIR, is a high-performance Supermicro system designed specifically for handling inference workloads. **Equipped with NVIDIA GH200 NVL2 which fully connects two GH200 Superchips through NVIDIA NVLink**, this system maximizes computational power in a 2U chassis with air-cooling for demanding applications.

**ARS-221GL-NHIR**

**Key features:**

- High density 2U 1-node system with NVIDIA GH200 NVL2, which fully connects two NVIDIA GH200 Superchips with NVIDIA NVLINK.
- The system is equipped with up to 1248GB of onboard memory (two of 480GB LPDDR5X for the CPUs and 144GB HBM3e for the GPU-intensive workloads).
- The system offers front hot-swap E1.S NVMe drives and multiple PCIe 5.0 slots supporting NVIDIA BlueField®-3 SuperNIC and NVIDIA ConnectX-7 NIC for 400Gb/s high-speed networking.

- 2U form factor with air cooling for optimal thermal management and a robust power supply to support continuous intensive processing.
- For detailed technical specification, refer to the system page here.

> **The oil and gas industry,** in particular, can benefit from this system as they reduce the time needed for complex subsurface simulations and reservoir modeling, expediting exploration and production decisions. By accelerating data-intensive seismic analysis, these systems help lower operational costs.

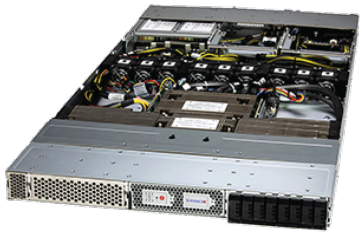
3. **Computer vision:** Computer vision enables machines to interpret and analyze visual data, driving innovation across industries. This data-intensive workload benefits significantly from the NVIDIA GH200 Grace Hopper Superchip, especially when integrated into Supermicro systems. These systems provide the computational power and efficiency required to process vast streams of visual data with high precision. The GH200's high memory bandwidth and AI-accelerated performance enable advanced computer vision tasks, including object detection, image segmentation, facial recognition, and behavior analysis, to be executed with unprecedented speed and accuracy. Whether analyzing video feeds for autonomous driving or optimizing traffic flow in smart cities, Supermicro systems with NVIDIA GH200 deliver the performance needed to unlock the full potential of computer vision.

### b) Supermicro system SuperServer ARS-111GL-NHR and SuperServer ARS-111GL-NHR-LCC enabling these workloads

The 1U Grace Hopper Superchip System, available in models ARS-111GL-NHR and ARS-111GL-NHR-LCC, integrates NVIDIA's GH200 Superchip with a powerful 72-core Grace CPU and NVIDIA H100 Tensor Core GPU within a compact 1U form factor.

**Key features:**

- High density 1U GPU system with one Integrated NVIDIA Hopper Tensor Core GPU and one 72-core NVIDIA Grace CPU
- 576GB of onboard memory (480GB LPDDR5X for the CPU and 96GB HBM3 for the GPU).
- The system offers front hot-swap E1.S NVMe drives and multiple PCIe 5.0 slots supporting NVIDIA BlueField®-3 SuperNIC and NVIDIA ConnectX-7 NIC for 400Gb/s high-speed networking.
- Available with both air-cooled and liquid-cooled configurations, this system offers robust scalability and flexibility for AI and HPC applications.
- For detailed technical specification, refer to the ARS-111GL-NHR system page here and ARS-111GL-NHR-LCC system page here.



**ARS-111GL-NHR**

**ARS-111GL-NHR-LCC**

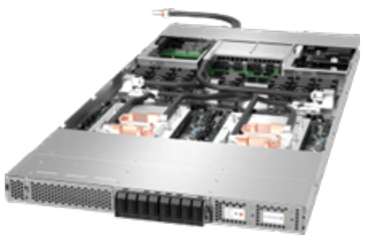
Applications such as autonomous vehicles, intelligent transportation systems, and smart manufacturing rely on real-time image and video processing to make critical decisions.

4. **Science and research:** AI is driving advancements in molecular dynamics simulations and complex scientific research, where speed and accuracy are critical. The NVIDIA GH200 Grace Hopper Superchip, with its high memory bandwidth and NVLink architecture, delivers unparalleled performance for workloads such as GROMACS, a leading software for molecular dynamics simulations. By efficiently handling tasks like Particle Mesh-Ewald (PME) calculations and Particle-Particle (PP) interactions, the GH200 accelerates simulation speeds while optimizing GPU utilization. Its integrated CPU-GPU design reduces latency and enhances communication, enabling faster, multi-node performance compared to traditional InfiniBand-based systems. For scientists running large-scale simulations, Supermicro systems powered by the GH200 provide a highly efficient and scalable solution, significantly reducing time-to-insight for complex molecular modeling and other research workloads.

5. **Linear solvers:** Building on advancements in science and research, linear solvers address a specific yet critical computational challenge: solving systems of linear equations that underpin applications in engineering, physics, and climate modeling. These workloads rely on multigrid iterative methods to improve efficiency by solving problems at varying resolutions and smoothing errors across grids. The NVIDIA GH200 Grace Hopper Superchip enhances linear solver performance through its high memory bandwidth and integrated CPU-GPU architecture, enabling seamless coordination for fast data transfer and processing. By optimizing coarse and fine grid levels, the GH200 reduces latency and accelerates complex simulations, such as finite element analysis and computational fluid dynamics (CFD). Supermicro systems powered by the GH200 deliver the scalability and speed required to tackle large-scale linear systems efficiently, driving breakthroughs in engineering and scientific innovation.

### c) Supermicro system SuperServer ARS-111GL-DNHR-LCC enabling these workloads

The 1U 2-Node NVIDIA GH200 Grace Hopper Superchip system, model ARS-111GL-DNHR-L CC, is a dual-node, Supermicro solution designed to maximize density and performance for demanding workloads through liquid-cooling. Each node integrates the NVIDIA GH200 Grace Hopper Superchip, which combines a 72-core NVIDIA Grace CPU with an NVIDIA Hopper Tensor Core GPU, offering substantial processing power.



ARS-111GL-DNHR-LCC

**Key features:**

- 1U 2-node system with NVIDIA GH200 Grace Hopper Superchip per node (liquid-cooled)
- 576GB of onboard memory per node (480GB LPDDR5X for the CPU and 96GB HBM3 for the GPU), it is designed to handle high-performance computing needs efficiently.
- The system offers front hot-swap E1.S NVMe drives and multiple PCIe 5.0 slots supporting NVIDIA BlueField®-3 SuperNIC and NVIDIA ConnectX-7 NIC for 400Gb/s high-speed networking.
- For detailed technical specification of this system, refer to the product page here.

Industries such as **finance** can leverage advanced processing for linear solvers. This Supermicro system can handle complex models, reduce latency, and improve performance in high-frequency trading, enabling faster, data-driven decisions.

> *"I have processed the entire plankton population, freshwater, and seawater, but it generates about 400 terabytes of output. Processing took about a month with Grace Hopper. In the past, we would have spent about 2 1/2 months processing it on our old machines. Now we do it in less than half the time or even faster."*
>
> **– Christopher M. Sullivan | Director - Research and Academic Computing, College of Earth, Ocean, and Atmospheric Sciences, Oregon State University**

## Summary of Supermicro systems with the NVIDIA GH200 Grace Hopper Superchip

| System Name | Form factor | Processor | GPU | System Cooling |
|---|---|---|---|---|
| ARS-221GL-NHIR | 2U Rackmount | CPU: 72-core NVIDIA Grace CPU on GH200 Grace Hopper™ Superchip | Max GPU Count: Up to 2 onboard GPUs<br>Supported GPU: NVIDIA Hopper Tensor Core GPU on GH200 Grace Hopper™ Superchip (Air-cooled)<br>CPU-GPU: Interconnect: NVLink®-C2C<br>GPU-GPU Interconnect: NVIDIA® NVLink® | Air |
| ARS-111GL-NHR | 1U Rackmount | CPU: 72-core NVIDIA Grace CPU on GH200 Grace Hopper™ Superchip | Max GPU Count: Up to 1 onboard GPU<br>Supported GPU: NVIDIA Hopper Tensor Core GPU on GH200 Grace Hopper™ Superchip (Air-cooled) | Air |
| ARS-111GL-NHR-LCC | 1U Rackmount | CPU: 72-core NVIDIA Grace CPU on GH200 Grace Hopper™ Superchip | Max GPU Count: Up to 1 onboard GPU<br>Supported GPU: NVIDIA Hopper Tensor Core GPU on GH200 Grace Hopper™ Superchip (Air-cooled) | Air and liquid |
| ARS-111GL-DNHR-LCC | 1U Rackmount | CPU: 72-core NVIDIA Grace CPU on GH200 Grace Hopper™ Superchip | GPU (per node)<br>Max GPU Count: Up to 1 onboard GPU<br>Supported GPU: NVIDIA Hopper Tensor Core GPU on GH200 Grace Hopper™ Superchip (liquid-cooled) | Air and liquid |

To conclude this section, it's worth highlighting that while the workloads and Supermicro systems discussed here provide a strong indication of how the Supermicro systems leveraging NVIDIA GH200 Superchips can be utilized, they are by no means the only configurations suited for these tasks. Supermicro's diverse portfolio enables the delivery of tailored solutions based on key factors such as workload size, cooling requirements, and cost.

## NVIDIA H200 and H100 Tensor Core GPUs

NVIDIA's Tensor Core GPUs are designed to drive innovation in AI and HPC by delivering exceptional performance, scalability, and efficiency for a variety of workloads. The NVIDIA H200 Tensor Core GPU builds on the capabilities of the NVIDIA H100 GPU, introducing HBM3e memory for faster access and larger capacity, enabling exceptional performance for generative AI, large language models, and HPC workloads. While both GPUs feature NVLink™, the Transformer Engine, and Multi-Instance GPU (MIG) to deliver robust scalability and efficiency, the H200 takes these capabilities further with its expanded memory pool, allowing it to tackle more demanding workloads with unprecedented efficiency. Together, these GPUs and platforms empower organizations to address a wide range of AI and HPC needs, from secure small-scale tasks to complex, memory-intensive applications.

**Choosing the right form factor:** While both the NVIDIA H100 and H200 deliver strong ROI and share many of the same benefits, the key distinction lies not just in choosing between the two GPUs but in selecting the right form factor to match specific workloads and infrastructure. Both the NVIDIA H100 and H200 are available in two primary form factors: **NVIDIA HGX** AI supercomputing platform which comes with 8 or 4 GPUs interconnected through high-speed NVIDIA NVLink and NVSwitches, and **PCIe (Peripheral Component Interconnect Express)** with optional **4-way or 2-way** NVLink Bridge.

**Form factor design differences: NVIDIA HGX™ H200 or NVIDIA HGX™ H100** is designed for AI training and inference, and high-performance computing (HPC) environments. It is typically found in large data centers for the most demanding AI and scientific computing workloads. Whereas the PCIe form available **as NVIDIA H200 NVL and NVIDIA H100 NVL** are designed to fit into standard PCIe slots in workstations and smaller servers. It's more accessible for a wide range of systems, including desktops, where fewer GPUs are needed, and is ideal for smaller-scale AI and inference deployment while providing versatility and maintaining performance.

**Performance:** The NVIDIA H200 is the first GPU to offer 141 gigabytes (GB) of HBM3e memory at 4.8 terabytes per second (TB/s) while the H100 provides 80GB of HBM3 at 3 terabytes per second (TB/s) of memory bandwidth per GPU. Both provide scalability with NVLink and NVSwitch to tackle AI and data analytics with high performance and scale to support massive datasets. With advanced tensor cores and the Transformer Engine, the H-series accelerates large-scale AI models and complex computations. The NVIDIA H200 provides additional memory bandwidth and efficiency, allowing it to manage larger datasets and more complex algorithms with fewer GPUs. Meanwhile, the NVIDIA H100 offers scalability, enabling organizations to achieve comparable performance by deploying more GPUs in parallel, making it a flexible choice for environments with expanding workloads.

### Workloads powered by NVIDIA HGX H200, HGX H100, H200 NVL, and H100 NVL on Supermicro Systems

This section provides an in-depth overview of workloads optimized for large-scale model training and AI inference, supported by Supermicro systems utilizing NVIDIA H200 and H100 Tensor Core GPUs. Supermicro systems harness the advanced processing capabilities of these GPUs to drive faster data processing, streamline deep learning workflows, and deliver enhanced precision for complex computational tasks.

Supermicro systems with NVIDIA HGX are purpose-built for deep learning, data analytics, high-performance computing (HPC), and large-scale generative AI, making them a valuable asset in data center infrastructure. Their features provide the necessary resources to support data-intensive and compute-heavy applications in AI and HPC environments. The Supermicro 5U PCIe GPU System with NVIDIA H100 NVL or H200 NVL are ideal for generative AI, HPC, and complex scientific workloads that require top-tier computational resources with flexible configurability.

1. **Large-scale AI model training:** As AI models grow in size and complexity, the need for high-performance infrastructure becomes critical to handle the vast amounts of data and computations required for training. 4th Generation NVIDIA NVLink provides up to 900 GB/s bandwidth between GPUs, with 80 GB HBM3 memory (for NVIDIA HGX H100) or 141GB HBM3e memory (NVIDIA HGX H200) per GPU and tensor cores optimized for massive AI models. Each NVIDIA HGX system includes NVLink interconnect for 8 GPUs per system and 8 NICs (BlueField-3 SuperNIC or ConnectX-7) for 1:1 GPU-to-NIC connections, enabling seamless scaling across systems and racks. Scaling across thousands of GPUs is further enhanced by NVIDIA

Quantum-2 InfiniBand and Spectrum™-X ethernet platforms, facilitating parallel processing for training models with hundreds of billions or trillions of parameters in expansive data center setups.

2. **AI inference**: Inference is critical for deploying AI applications in real-world scenarios, enabling models to make predictions and generate responses. AI inference powers tasks such as natural language processing, image recognition, recommendation systems, and real-time decision-making. NVIDIA H100 and H200 are optimized for processing inference tasks quickly and efficiently. HGX systems are ideal for high-performance inference at scale, offering seamless GPU-to-GPU communication for handling large and complex models. For flexible deployments in standard servers, PCIe GPUs like H100 NVL and H200 NVL provide scalable solutions, making them well-suited for cost-effective production environments or smaller-scale setups.

## Supermicro systems enabling these workloads

The systems described below are engineered for superior performance, offering an unparalleled performance-per-dollar ratio that meets the demands of customers seeking exceptional computational power.

### a) SuperServer SYS-821GE-TNHR (Intel CPU) and A+ Server AS -8125GS-TNHR (AMD CPU)

The Supermicro 8U NVIDIA HGX H200 8-GPU or H100 8-GPU system SYS-821GE-TNHR, featuring Intel® Xeon® 4th and 5th Gen Scalable processors, and the A+ Server AS-8125GS-TNHR featuring AMD's 4th Gen EPYC processors, represent high-performance computing solutions tailored for demanding workloads. The Intel-based system supports dual-socket CPUs with up to 64 cores and a TDP of up to 385W, while the AMD system offers up to 256 cores across dual processors, showcasing exceptional parallel processing capabilities. Both platforms feature advanced thermal designs and modular architectures for efficiency and flexibility.

**SYS-821GE-TNHR**

**Key features across both systems:**

- Both systems support NVIDIA HGX H200 8-GPU or H100 8-GPU configurations with NVLink bandwidth up to 900 GB/s for seamless GPU-to-GPU communication. They utilize NVSwitch for efficient communication between GPUs.
- These systems are equipped with a 1:1 networking configuration for each GPU, enabling NVIDIA GPUDirect RDMA and Storage. Leveraging NVIDIA Quantum-2 InfiniBand and NVIDIA Spectrum™-X Ethernet platform, these systems provide high-speed, low-latency connections and scaling.
- The SYS-821GE-TNHR features 12 front hot-swap NVMe and 3 SATA drive bays, expandable to 16 NVMe and 3 SATA bays, plus two M.2 NVMe slots, while the A+ Server AS-8125GS-TNHR provides 16 front hot-swap NVMe drive bays with additional options for SATA and M.2 storage
- The SYS-821GE-TNHR supports up to 8TB of DDR5 ECC DRAM memory via 32 DIMM slots with speeds up to 5600MT/s, while the AS-8125GS-TNHR accommodates up to 6TB of DDR5 ECC memory through 24 DIMM slots, running at speeds up to 4800MT/s.
- Equipped with 6x 3000W Titanium-level redundant power supply, both systems ensure high power efficiency and reliability under demanding workloads, providing stable performance even during peak utilization.
- Certified for the NVIDIA AI Enterprise Platform, these systems integrate NVIDIA NIM microservices, streamlining the deployment and scaling of AI models across diverse use cases.
- For detailed technical specification, refer to the product page of SuperServer SYS-821GE-TNHR and A+ Server AS -8125GS-TNHR

With these features, the system provides the necessary resources to support data-intensive and compute-heavy applications in AI and HPC environments. This system is purpose-built for deep learning, data analytics, high-performance computing (HPC), and large-scale AI deployment, making it a valuable asset in AI data center infrastructure.

### b) SuperServer SYS-421GE-TNHR2-LCC (Intel CPU) and A+ Server AS -4125GS-TNHR2-LCC (AMD CPU)

SYS-421GE-TNHR2-LCC

Supermicro's 4U HGX H200 or H100 8-GPU systems, including the Intel-based SuperServer SYS-421GE-TNHR2-LCC and the AMD-based A+ Server AS-4125GS-TNHR2-LCC, are cutting-edge, high-density computing solutions tailored for the most demanding AI and HPC applications. Both systems leverage advanced liquid-cooling technology, doubling computing density per rack while maintaining optimal thermal efficiency. The Intel-based model features dual-socket Intel® Xeon® Scalable processors, supporting up to 64 cores with a TDP of up to 385W, while the AMD-based system utilizes dual AMD EPYC™ 9004 Series processors, offering up to 192 cores and a TDP of up to 400W for unparalleled multi-threaded performance.

**Key features across both systems:**

- Doubling compute density through Supermicro's custom liquid-cooling solution with up to 40% reduction in electricity cost for data center
- Both systems support NVIDIA HGX H200 and H100 8-GPU configurations with NVLink bandwidth up to 900 GB/s for seamless GPU-to-GPU communication. They utilize NVSwitch for efficient communication between GPUs.
- These systems are equipped with a 1:1 networking configuration for each GPU, with AS -4125GS-TNHR2-LCC enabling NVIDIA GPUDirect RDMA and Storage. Leveraging NVIDIA Quantum-2 InfiniBand and NVIDIA Spectrum™-X Ethernet platform, these systems provide high-speed, low-latency connections.
- The Intel system provides 32 DIMM slots for ECC DDR5 memory at speeds of up to 5600 MT/s, while the AMD system supports 24 DIMM slots with DDR5 memory up to 4800 MT/s, ensuring excellent bandwidth for diverse workloads.
- Both systems feature 8 hot-swap bays for NVMe and SATA drives, M.2 slots for added flexibility, and up to 12 PCIe 5.0 slots for seamless integration of additional network or storage components.
- Equipped with a 4x 5250W (2+2) Redundant Titanium Level power supplies, both systems ensure stable, resilient operation, making them ideal for high-scale AI training, deep learning, and advanced analytics.
- Certified for the NVIDIA AI Enterprise Platform, these systems integrate NVIDIA NIM microservices, streamlining the deployment and scaling of AI models across diverse use cases.
- For detailed technical specification, refer to the product page of SuperServer SYS-421GE-TNHR2-LCC and A+ Server AS -4125GS-TNHR2-LCC

3. **Data analytics and Machine Learning (ML):** Data analytics and ML involve uncovering insights from massive datasets and building predictive models, which are critical for driving innovation across industries. Use cases include customer behavior analysis in retail, fraud detection in finance, predictive maintenance in manufacturing, and data science tasks like sentiment analysis, time-series forecasting, and clustering for market segmentation. These workloads demand robust infrastructure to process large volumes of data efficiently and enable complex computations. With technologies like NVSwitch and NVLink, NVIDIA H200 and H100 platforms ensure high-speed communication and seamless scaling across GPUs for data-intensive tasks. Whether deployed in HGX systems for high-performance environments or PCIe systems for broader compatibility and flexibility, Supermicro systems powered by NVIDIA Tensor Core GPUs excel in supporting advanced analytics, machine learning, and data science with efficiency and scalability

4. **Parallel HPC simulations:** High-Performance Computing (HPC) Simulations: HPC simulations involve modeling complex systems to solve challenging problems in areas like climate research, physics, engineering, and pharmaceuticals. Examples include simulating weather patterns, optimizing aerodynamic designs, and predicting molecular interactions in drug discovery. Supermicro systems powered by NVIDIA H200 and H100 GPU, whether in HGX configurations for ultra-high performance or PCIe setups for flexibility, empower organizations to meet their HPC needs with efficiency and precision.

### Supermicro systems enabling these workloads:

#### c) SuperServer SYS-521GE-TNRT (Intel CPU)

The Supermicro 5U PCIe GPU system, is a high-performance server designed for intensive AI, deep learning, and scientific visualization workloads The SYS-521GE-TNRT, features Intel® Xeon® 4th and 5th Gen Scalable processors with support for up to 64 cores and a TDP of up to 385W.

**SYS-521GE-TNRT**

**Key Features:**

- The system supports up to 10 double-width NVIDIA PCIe GPUs, including H100 NVL, H100, and L40S, with a PCIe 5.0 x16 dual-root CPU-GPU interconnect and optional NVIDIA NVLink Bridge for enhanced GPU-to-GPU communication.
- This system includes 13 PCIe Gen5 x16 slots and an AIOM/OCP 3.0 slot, with support for NVIDIA ConnectX-7 NIC and NVIDIA BlueField®-3 SuperNIC for advanced networking capabilities.
- It features 32 DIMM slots supporting up to 5600 MT/s ECC DDR5 memory and offers flexible storage configurations with up to 16 hot-swap 2.5" drive bays in its default configuration allowing for 24 hot-swap bays (16 NVMe and 8 SATA). It also includes two M.2 NVMe slots.
- The SYS-521GE-TNRT is equipped with 4x 2700W Redundant Titanium Level power supplies, ensuring energy efficiency and reliable operation for demanding workloads.
- Certified for the NVIDIA AI Enterprise Platform, this system integrates NVIDIA NIM microservices, streamlining the deployment and scaling of AI models across diverse use cases.
- For detailed technical specification, refer to the product page of SuperServer SYS-521GE-TNRT

#### d) SuperServer SYS-522GA-NRT (Intel CPU) and A+ Server AS -5126GS-TNRT (AMD CPU)

The SuperServer SYS-522GA-NRT and AS-5126GS-TNRT are high-performance 5U systems tailored for enterprise AI, HPC, and visualization workloads. The SYS-522GA-NRT features dual Intel® Xeon® processors, while the AS-5126GS-TNRT supports AMD EPYC™ 9004 series processors. Both systems incorporate advanced thermal designs to handle high-power GPUs, making them ideal for AI training, rendering, and demanding computational tasks.

**SYS-522GA-NRT**

**Key Features:**

- The SYS-522GA-NRT supports up to 10 double-width GPUs, including NVIDIA PCIe options like H100 NVL, H200 NVL (141GB), and L40S, with a PCIe 5.0 x16 dual-root interconnect for high-speed data transfer. The AS -5126GS-TNRT accommodates up to 8 double-width GPUs, compatible with the same NVIDIA PCIe GPUs and offers optional NVIDIA NVLink for GPU communication.
- Both systems support advanced networking options, such as NVIDIA ConnectX-7 NIC and NVIDIA BlueField-3 SuperNIC for enhanced connectivity and performance

- 16 front hot-swap NVMe bays and an additional 8 hot-swap NVMe bays, providing flexible and high-capacity storage options.
- With up to 13 PCIe 5.0 x16 slots, these systems are optimized for maximum expansion and performance.
- For detailed technical specification, refer to the product pages of SuperServer SYS-522GA-NRT and A+ Server AS -5126GS-TNRT

5. **Business Intelligence (BI) and data visualization:** In today's data-driven landscape, effective business intelligence and data visualization are essential for informed decision-making. The PCIe version supports accelerated data processing and real-time analytics using GPU-accelerated tools such as Tableau or Power BI. With high throughput and large memory bandwidth, it allows businesses to generate fast, insightful data visualizations and reports, improving decision-making processes.
6. **CNN/RNN vision task:** Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are widely used for various vision tasks, such as image recognition, object detection, and video analysis. CNNs are highly effective in processing grid-like data structures, making them ideal for tasks such as facial recognition or medical image analysis, where spatial patterns are critical. RNNs, on the other hand, are useful for tasks that involve temporal sequences, such as video frame prediction or action recognition. Supermicro systems with NVIDIA H200 and H100 Tensor Core GPUs, available in HGX or PCIe configurations, provide the computational power needed to accelerate these deep learning workloads. With high throughput for tensor operations, these GPUs enable faster training and inference for CNNs and RNNs. Advanced features like FP8 precision and NVLink interconnects enhance memory bandwidth and GPU-to-GPU communication, ensuring efficient processing of large-scale image and video datasets. These capabilities are particularly valuable in healthcare, where CNNs and RNNs are deployed for applications such as diagnostics and medical image analysis, delivering high accuracy and speed to advance patient outcomes.
7. **Gen AI model for images:** Generative AI models are revolutionizing how images are created and manipulated, enabling rapid production of high-quality visuals from text prompts. These models, such as Stable Diffusion, use advanced neural networks to generate realistic images by interpreting detailed input and iteratively refining noisy data into coherent visuals. Supermicro systems powered by NVIDIA Tensor Core GPUs, featuring the latest H200 and H100 models in HGX or PCIe configurations, provide the computational power and efficiency needed for fine-tuning and inference of these models. With features such as FP8 precision and enhanced GPU memory bandwidth, these GPUs accelerate image generation tasks while efficiently managing the large datasets required for creating diverse and highly detailed visuals. Generative AI models powered by these systems are transforming industries like retail, enabling personalized advertising and virtual try-ons, and healthcare, supporting the creation of detailed medical imagery and educational materials to enhance patient care.

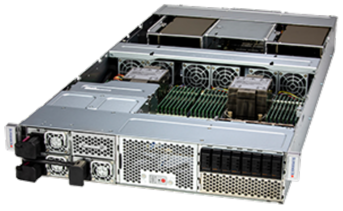
### e) Supermicro systems SuperServer ARS-221GL-NR, and SuperServer SYS-221GE-NR enabling these workloads

The 2U NVIDIA MGX System from Supermicro provides an adaptable platform with two configuration options, supporting both the NVIDIA Grace CPU Superchip and x86 processors.

**Key features:**

- The systems integrate a 144-core NVIDIA Grace CPU within a single chip or dual 4th and 5th Generation Intel® Xeon® Scalable processors, offering up to 60 cores per socket for robust computational power.
- ARS-221GL-NR features up to 960GB of onboard ECC LPDDR5X memory, ensuring minimal latency and maximum power efficiency. Whereas, SYS-221GE-NR, offers 32 DIMM slots supporting up to 8TB of 5600MT/s ECC DDR5 memory, providing substantial capacity and high-speed performance.

**ARS-221GL-NR**



**SYS-221GE-NR**

- The system also supports up to four double-width GPUs, including the NVIDIA H100 NVL (PCIe GPU), and offers up to 8x hot-swappable E1.S drives, 2x M.2 NVMe drives, and three PCIe 5.0 x16 slots compatible with NVIDIA BlueField-3 or ConnectX-7.
- Enhanced with up to three 2000W Titanium Level redundant power supplies, this system ensures high reliability for demanding workloads across AI, data analytics, and HPC applications.
- For detailed technical specification, click on the respective system names, ARS-221-GL-NR, SYS-221GE-NR to access the product pages.

### Summary of Supermicro systems with NVIDIA H200 and H100 Tensor Core GPUs

| System Name | Form Factor | Processor | GPU | System Cooling |
|---|---|---|---|---|
| SYS-821GE-TNHR (Intel CPU) | 8U Rackmount | CPU: Dual Socket E (LGA-4677)<br>Core Count: Up to 64C/128T; Up to 320MB Cache per CPU | Max GPU Count: Up to 8 onboard GPUs<br>Supported GPU: NVIDIA HGX H100 8-GPU (80GB), HGX H200 8-GPU (141GB)<br>CPU-GPU Interconnect: PCIe 5.0 x16 CPU-to-GPU Interconnect<br>GPU-GPU Interconnect: NVIDIA® NVLink® with NVSwitch™ | Air |
| AS -8125GS-TNHR (AMD CPU) | | Dual processor(s)<br>AMD EPYC™ 9004/9005 Series Processors<br>(* AMD EPYC™ 9005 Series drop-in support requires board revision 2.x) | Max GPU Count: Up to 8 onboard GPUs<br>Supported GPU: NVIDIA SXM: HGX H100 8-GPU (80GB), HGX H200 8-GPU (141GB)<br>CPU-GPU Interconnect: PCIe 5.0 x16 CPU-to-GPU Interconnect<br>GPU-GPU Interconnect: NVIDIA® NVLink® with NVSwitch™ | |
| SYS-421GE-TNHR2-LCC (Intel CPU) | 4U Rackmount | CPU: Dual Socket E (LGA-4677)<br>Core Count: 64C/128T; 320MB Cache per CPU | Max GPU Count: Up to 8 onboard GPUs<br>Supported GPU: NVIDIA SXM: HGX H100 8-GPU (80GB), HGX H200 8-GPU (141GB)<br>CPU-GPU Interconnect: PCIe 5.0 x16 CPU-to-GPU Interconnect<br>GPU-GPU Interconnect: NVIDIA® NVLink® with NVSwitch™ | Liquid |
| AS -4125GS-TNHR2-LCC (AMD CPU) | | CPU: Dual processor(s)<br>AMD EPYC™ 9004/9005 Series Processors<br>Core Count: Up to 128C/256T | | |
| SYS-521GE-TNRT (Intel CPU) | 5U Rackmount | CPU: Dual Socket E (LGA-4677)<br>5th Gen Intel® Xeon®/4th Gen Intel® Xeon® Scalable processors | Max GPU Count: Up to 10 double-width or 10 single-width GPUs<br>Supported GPU: NVIDIA PCIe: H100 NVL, H100, L40S, A100 | Air and liquid |

| System Name | Form Factor | Processor | GPU | System Cooling |
|---|---|---|---|---|
| | | Core Count: 64C/128T; 320MB Cache per CPU | CPU-GPU Interconnect: PCIe 5.0 x16 Switch Dual-Root<br>GPU-GPU Interconnect: NVIDIA® NVLink® Bridge (optional) | |
| SYS-522GA-NRT (Intel CPU) | 5U Rackmount | CPU: Dual Socket BR (LGA-7529)<br>Intel® Xeon® 6900 series processors with P-cores | Max GPU Count: Up to 10 double-width or 10 single-width GPUs<br>Supported GPU: NVIDIA PCIe: H100 NVL, L40S, L4<br>CPU-GPU Interconnect: PCIe 5.0 x16 Switch Dual-Root<br>GPU-GPU Interconnect: NVIDIA® NVLink® Bridge (optional) | Air and liquid |
| AS -5126GS-TNRT (AMD CPU) | 5U Rackmount | CPU: Dual processor(s)<br>AMD EPYC™ 9005/9004 Series Processors | Max GPU Count: Up to 8 double-width GPUs<br>Supported GPU: NVIDIA PCIe: H100 NVL, H200 NVL (141GB), L40S<br>CPU-GPU Interconnect: Direct Attached<br>GPU-GPU Interconnect: NVIDIA® NVLink® Bridge (optional) | Air |
| ARS-221GL-NR | 2U Rackmount | CPU: Dual processor(s)<br>Dual 72-core CPUs on a NVIDIA Grace CPU Superchip | Max GPU Count: Up to 2 double width GPUs<br>Supported GPU: NVIDIA PCIe: H100 NVL, L40S<br>CPU-GPU Interconnect: PCIe 5.0 x16 CPU-to-GPU Interconnect<br>GPU-GPU Interconnect: PCIe | Air |
| SYS-221GE-NR | 2U Rackmount | CPU: Dual Socket E (LGA-4677)<br>5th Gen Intel® Xeon®/4th Gen Intel® Xeon® Scalable processors<br>Core Count: Up to 64C/128T; Up to 320MB Cache per CPU | Max GPU Count: Up to 4 double-width GPUs<br>Supported GPU: NVIDIA PCIe: H100 NVL, H100, L40S<br>CPU-GPU Interconnect: PCIe 5.0 x16 CPU-to-GPU Interconnect<br>GPU-GPU Interconnect: NVIDIA® NVLink® Bridge (optional) | Air |

### NVIDIA AI Enterprise: Maximizing the Power of NVIDIA AI with Supermicro Systems

Organizations can unlock even greater value from their NVIDIA-powered infrastructure by integrating NVIDIA AI Enterprise, a full-stack software solution that optimizes AI development and deployment. Paired with NVIDIA Certified Systems, such as Supermicro's HGX H100 systems, this platform ensures seamless scalability and efficiency for enterprise-grade AI applications. Additionally, NVIDIA GH200 supports all NVIDIA software platforms, including NVIDIA AI Enterprise, and the HPC SDK. NVIDIA H200 NVL and H100 NVL come with a five-year NVIDIA AI Enterprise subscription, simplifying the path to building an enterprise AI-ready platform.

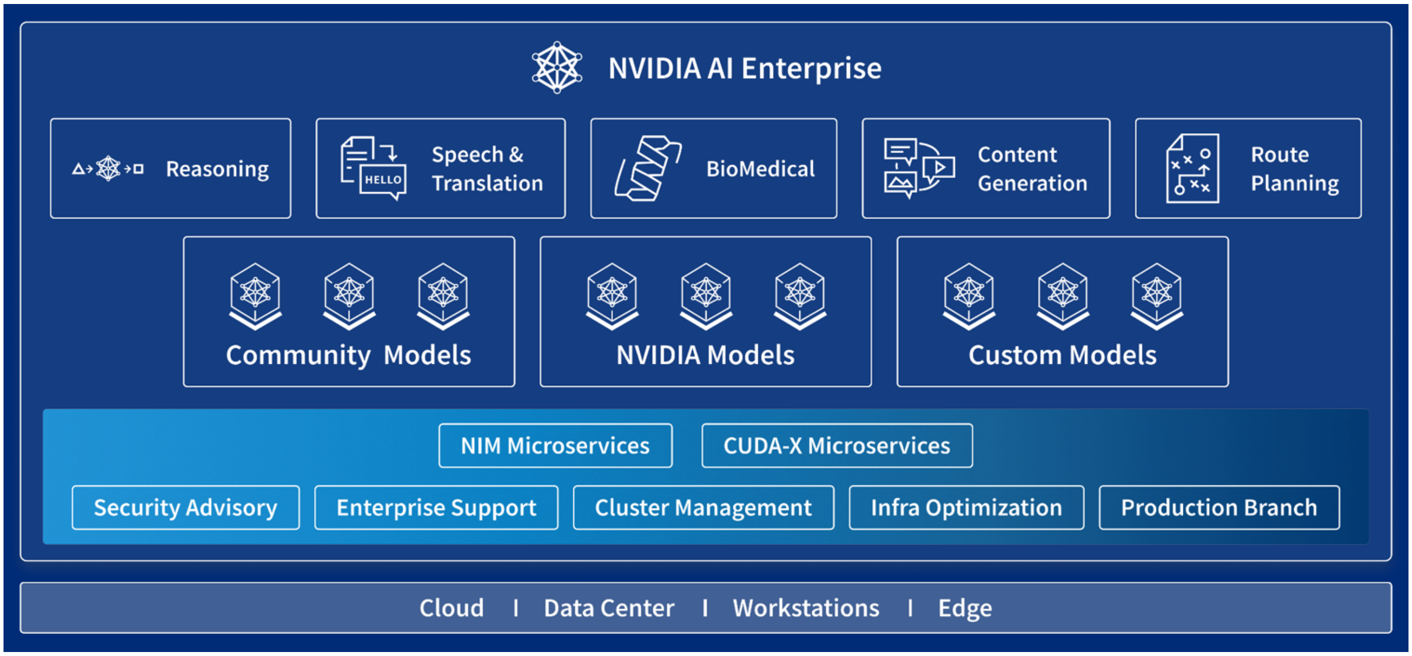
Many of Supermicro systems are NVIDIA-Certified or being certified to be fully optimized for NVIDIA AI Enterprise software out of the box. For the complete list of NVIDIA Certified systems, click here.

**Features:**

- Optimized NIM inference microservices, which enhance model performance and speed time to deployment.
- NVIDIA® Riva, NeMo, RAPIDS™, and other frameworks and libraries across many domains.
- Tools and libraries that accelerate data analytics, AI model training and customization, and AI model optimization and deployment.
- Infrastructure software to help manage AI clusters at scale, across the edge and data center, both bare-metal and virtualized.

**Benefits**

- **Optimize Performance:** NVIDIA NIM microservices designed for secure and reliable AI deployment that speed LLM throughput by up to 5X and improve retrieval throughput by 2X and accuracy by 30%.
- **Accelerate Time to Deployment:** Production-ready AI software containers accessible via industry-standard APIs and reference architecture workflows for a broad array of end-to-end AI solutions.
- **Run Anywhere:** Standards-based and containerized microservices are certified to run in the cloud, in the data center, and on workstations.
- **Enterprise-Grade:** Predictable production software branches for API stability, proactive security remediation, and NVIDIA Enterprise Support.



## Supermicro systems accelerating your AI workloads

Supermicro's cutting-edge AI-ready infrastructure solutions help with **large-scale training to enterprise level inferencing** enabling organizations to streamline and accelerate AI deployment. Their AI-infrastructure empowers workloads with **optimal performance and scalability** while **optimizing costs and minimizing environmental impact.** Supermicro's flexible range of solutions ensures that organizations implementing AI solutions can scale up their implementation as much as needed. Whatever the requirements, solutions are available to expand memory, processing power, and storage to meet any situation.

With the addition of NVIDIA AI Enterprise, Supermicro can offer organizations even greater value by providing a full-stack software solution that optimizes AI development and deployment, integrating seamlessly with NVIDIA GPUs for enterprise-grade applications.

Supermicro's AI infrastructure is purpose-built for exceptional performance, scalability, and efficiency, providing organizations with robust, adaptable solutions to meet the demands of modern AI workloads and accelerate innovation.

For more information

To learn more about our AI solutions, visit https://www.supermicro.com/en/solutions/ai-deep-learning.

### 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

Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com.