SOLUTION BRIEF

SUPERMICRO RACK SCALE SOLUTIONS: LARGE SCALE AI TRAINING WITH LIQUID COOLING

Embrace an Order-of-Magnitude Leap In Performance With Supermicro Rack Scale AI Solutions

TABLE OF CONTENTS

Executive Summary

Turbocharge AI infrastructures with Supermicro's rack-scale complete plug-and-play AI solutions powered by Supermicro SYS-821GE-TNHR or the AS -8125GS-TNHR GPU Servers. These application-optimized servers are ideal for medium to very large AI training scenarios and contain dual AMD or Intel CPUs and eight high-performance NVIDIA GPUs. This purpose-built highest density turnkey rack scale solution is extremely scalable and customizable to meet any scale of Deep Learning workload demands. Leveraging NVIDIA's cutting-edge NVIDIA H100 SXM GPUs and harnessing the power of Supermicro's supreme building blocks, these rack-scale AI solutions deliver unprecedented Deep Learning performance.

Solution Highlights

Supreme AI Cluster for Exascale Computing

The Supermicro Rack Scale AI Solution is powered by Supermicro GPU servers - the highest density and compact computation powerhouse. The cluster utilizes the latest NVIDIA HGX™ H100 GPUs to deliver incomparable performance. The design features 32 GPUs in the Base Package (Scalable Unit-SU), scaling up to 128 GPUs per POD (4 racks of servers) and 256 GPUs per SuperPOD (8 racks of servers).

Scalable Design achieving unprecedented peak performance

Supermicro Rack Scale AI Solution is designed to provide outstanding scalability for faster and easier future deployments. Starting with the Base Package (Scalable Unit-SU) delivering 1.1 PetaFlops, the cluster is seamlessly scalable to reach close to 4.5 Peta Flops (FP64) per POD and up to 8.7 Peta Flops (FP64) per SuperPOD. In addition, Supermicro offers to deploy Rack Scale AI Solutions with 1 to 4 nodes as a proof of concept (POC). It provides flexibility to quickly scale to hundreds of servers via SuperPODs to meet workload demands.

Most Advanced Processors & Networking

The clusters feature the latest and state-of-the-art CPUs (both 4th Gen Intel® Xeon® Platinum Processors or AMD EPYC™ 9004 Processors), achieving an unprecedented number of cores. Furthermore, each cluster is powered by NDR InfiniBand, allowing virtually unlimited scalability for large data aggregation through the network. Along with three terabytes per second (TB/s) of memory bandwidth per GPU and scalability with NVLink (900 GB/sec) and NVSwitch (up to 256 GPU connections)—further enhancing system performance under heavy Deep Learning workloads.

Customizable Highest Quality Storage Options

Depending on the scale and end application requirements, Supermicro's AI Solution offers additional Storage Solutions that are seamlessly integrated and completely tested with the compute cluster. Utilizing Supermicro's application-optimized, high performance storage blocks, along with storage software integration (parallel filing systems like WekaIO, BeeGFS, etc.), Supermicros AI solutions are complete offerings – capable of meeting any scale of DL workloads.

Flexible and Superior Cooling Options

With the rising number of TDPs for both CPUs and GPUs, large-scale AI clusters will soon demand superior cooling technologies compared to air cooling. Supermicro Rack Scale AI Solution offers air and liquid cooling options, which include Direct To Chip, Rear Door Heat Exchangers, and Immersion Cooling. Powered by Supermicro's high-quality liquid cooling components, Supermicro's AI solution provides dramatic savings in PUE and OPEX. In addition, the building blocks for this solution can be either air cooled or liquid cooled.

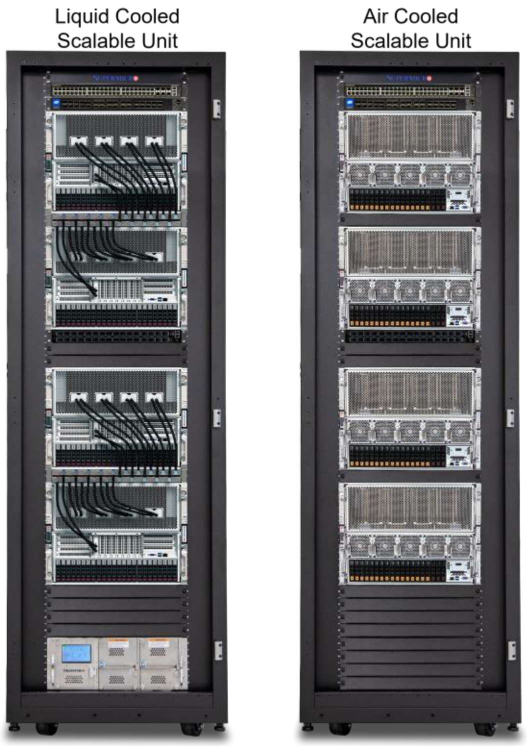


Figure 1- Liquid Cooled or Air Cooled Supermicro 8U Servers with NVIDIA H100 GPUs

Use Cases

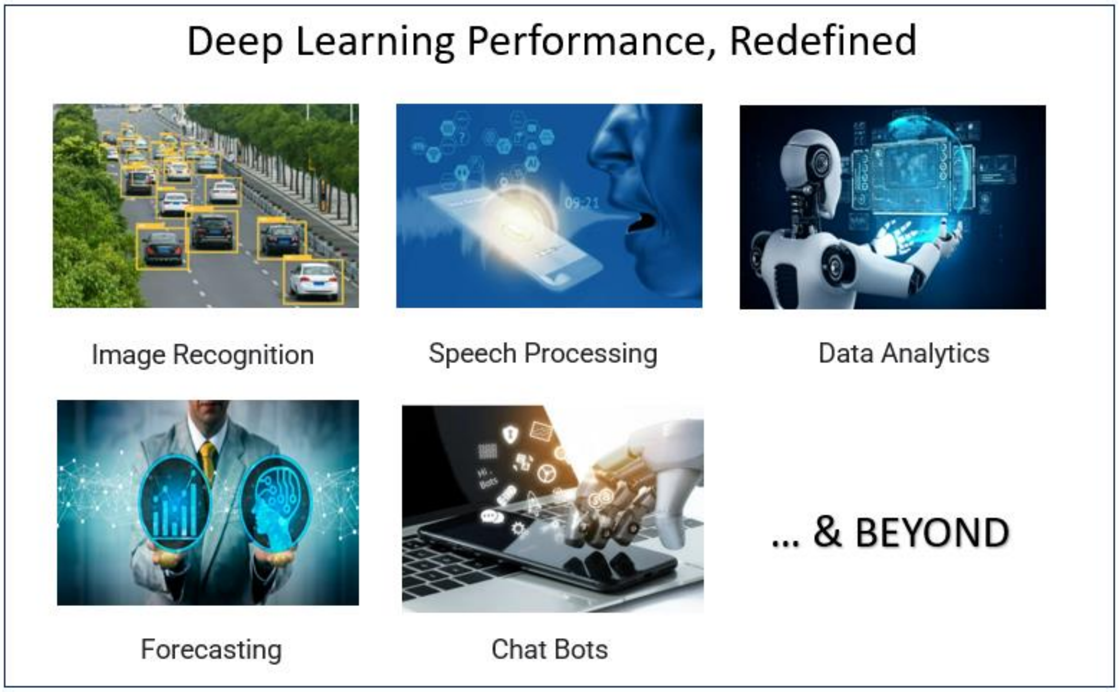
Supermicro’s rack-scale AI solutions are designed to remove AI infrastructure obstacles and bottlenecks, accelerating Deep Learning (DL) performance to the max.

Primary Use Case – Large Scale Distributed DL Training

Deep Learning Training requires high-efficiency parallelism and extreme node-to-node bandwidth to deliver faster training times. Supermicro’s rack-scale design facilitates training massive neural network models with millions of training instances and billions of parameters - in the most optimized and cost-efficient manner. While the Base Package (SU) is a great starting point for training DL models, the PODs and SuperPODs (and beyond) ensure to shorten enterprise level DL training times to a minimum.

Secondary Use Cases – DL Inference & Hyperparameter Search

In addition to parallel computing, production Inferencing requires deploying models at scale and high availability. Supermicro’s GPU servers include (N+N) power redundancy, capable of delivering seamless inferencing performance. Regarding Hyperparameter search, where parallel computation matters but node-to-node bandwidth is not critical, Supermicro’s customizable networking options make these solutions an excellent fit for this application.



Deep Learning Performance, Redefined

Image Recognition

Speech Processing

Data Analytics

Forecasting

Chat Bots

… & BEYOND

Supermicro’s versatile array of end-application and delivery focused total AI solutions offer great flexibility to choose from the latest and greatest compute platforms. Both ML training and inferencing serving applications can scale from a single rack to a SuperPOD. Generative AI (such as GPT models) and Large Language Models are examples of applications that can take advantage of the scalability of this solution.

The Supermicro Scalable Rack Scale AI Solutions are based on a single rack as a building block, referred to as a Scalable Unit, which can then be expanded to four or eight racks.

	SU/Base Package SRS-42UGPU-AI-SU1	POD SRS-42UGPU-AI-SU2	SuperPOD SRS-42UGPU-AI-SU3
GPU Server (8U 8GPU)	4x SYS-821GE-TNHR / 4x AS -8125GS-TNHR	16x SYS-821GE-TNHR / 16x AS -8125GS-TNHR	32x SYS-821GE-TNHR / 32x AS -8125GS-TNHR
Total CPUs	8x Intel® Xeon® Platinum 8480+ Processors or 8x AMD EPYC™ 9004 Processors	32x Intel® Xeon® Platinum 8480+ Processors or 16x AMD EPYC™ 9004 Processors	64x Intel® Xeon® Platinum 8480+ Processors or 64x AMD EPYC™ 9004 Processors
Total GPUs	32x NVIDIA HGX H100 SXM5	128x NVIDIA HGX H100 SXM5	256x NVIDIA HGX H100 SXM5
Rack	1x 42U (Optional 48U)	4x 42U (Optional 48U)	8x 42U (Optional 48U)
Memory	32TB DDR5 (X13) 24TB DDR5 (H13)	128TB DDR5 (X13) 96TB DDR5 (H13)	256TB DDR5 (X13) 192TB DDR5 (X13)
Estimated Total Power Per Rack	Max 45 kW	Max 180 kW	Max 360 kW
Networking	1x 400G 64-port NDR IB: SSE-MQM9700-NS2F	1x 400G 64-port NDR IB: SSE-MQM9700-NS2F	3x 400G 64-port NDR IB: SSE-MQM9700-NS2F
	1x SMC 100G Eth Switch (Storage)	1x SMC 100G Eth Switch (Storage)	1x SMC 100G Eth Switch (Storage)
	1x SMC 1G/25G MGT Switch	1x SMC 1G/25G MGT Switch	2x SMC 1G/25G MGT Switch

Customer Benefits

Rapid Deployment

Rack scale solutions can lower the deployment time of an average IT system from 3 months down to 2 weeks. We attribute this to the fact that our rack assembly team stocks an inventory of parts and can start assembly of a rack cabinet of systems immediately upon receiving an order.

Rather than customers waiting and stock-piling many components until everything arrives, Supermicro can begin assembly quickly and ship the final product to customers in a single box. Supermicro Plug and Play (PnP) solution can save money and accelerate the time-to-market approach for timely deployment.

Industry Standard Components

Supermicro builds the entire Rack Scale solution. That means we audit all hardware components, including 3rd party components, to give confidence that a completely tested and standardized total solution. Evaluating factors such as processing power, memory capacity, network connectivity options, storage capacity, and reliability ensures that the selected hardware meets the specific needs of the rack integration project. It's all guaranteed to work right after delivery.

Energy Efficiency

Supermicro can adapt our rack solutions to whatever power configuration our customers have.   The manufacturing facility supports 208, 230, 415, or 480VAC. Single or Three Phase, and the facility is 48VDC-ready.   Most importantly, though, the rack scale solutions are energy efficient. Supermicro significantly reduces energy consumption through more efficient power supplies in our products, liquid cooling capabilities, and even immersion cooling.

Supermicro considers factors such as power consumption, heat dissipation, cooling capacity, and available space for IT solutions. For example, determining proper ventilation requires assessing components' heat output and airflow patterns and implementing appropriate cooling mechanisms to prevent overheating and ensure efficient cooling throughout the rack.

Liquid and immersion cooling significantly reduce the energy required to cool IT equipment.   Because liquid creates much better thermal transfer than air, the cost to our customers to cool a rack cabinet of IT equipment can be a tenth of what an air-cooled system might require.

Supermicro's liquid cooled racks are optimized for high coolant temperatures offering unmatched efficiency. The solution can sustain a 100% server uptime with the new Supermicro Coolant Distribution Unit, which integrates redundant and hot swappable pump modules and Power supplies. The integrated software suite lets customers control the entire system from a single interface. This solution also comes with best-in-class after-sales services dispensed by our local experts.

Lower Carbon Footprint

Along with lower energy consumption comes lower carbon emissions. A hallmark of Supermicro is the desire to deliver green products to our customers, and Supermicro has done that repeatedly over the past several years. By reducing power consumption with Supermicro liquid cooling and immersion cooling capabilities, customers can better meet their reduced carbon footprint goals.

Scalable

A large value proposition of our rack scale systems is our ability to scale-up and scale-out with our customers' expanding IT needs. Because Supermicro uses a building-block approach to our rack solutions, the manufacturing facility can easily add new systems as our customers' requirements grow. The components work together seamlessly and can be added cumulatively.

Testing/Validation Expertise

Supermicro has developed its own unique quality assurance testing processes that will thoroughly validate the operational effectiveness of the entire integrated rack solution (pre-shipment). As a result, customers are delivered an entire solution, thoroughly tested as a working unit at the rack or multi-rack level.

Representative Performance Benchmarks

X13-H100 GPU Super Server: Training Performance - ResNet-50 v1.5 for MXNet (Config 1)

Performance Unit: Images/ sec (Higher is better) –

# of GPUs	SYS-821GE-TNHR	Ref NVIDIA DGX A100
1	6385	3411
4	25022	13443
8	49433	26674

H13-H100 GPU Super Server: Training Performance - ResNet-50 v1.5 for MXNet (Config 2)

Performance Unit: Images/ sec (Higher is better)

# of GPUs	AS-8125GS-TNHR	Ref NVIDIA DGX A100
1	6393	3411
4	24705	13443
8	49057	26674

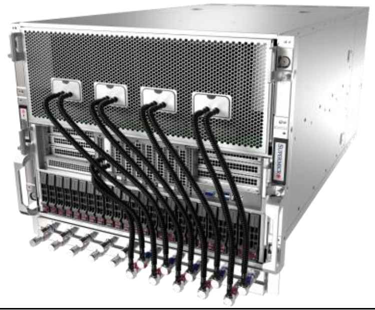
H13-H100 GPU Super Server: TensorRT BERT Large Inference Performance (Higher is Better)

Data Type	Batch Size	Sequence Length	AS-8125GS-TNHR	Ref A100 SXM4 80GB	Ref H100 SXM5 80GB
INT8	128	128	9743	4887	9622
INT8	8	128	5056	2679	5024
INT8	128	384	2826	1412	2819
INT8	8	384	2047	1071	2016

Building Block Server: GPU Super Server SYS-821GE-TNHR

Overview	8U Dual Socket (4th Gen Intel® Xeon® Scalable Processors), up to 8 SXM5 GPUs
CPU	2x 4th Gen Intel Xeon Scalable Processors
Memory (additional memory available)	32 DIMM slots Up to 8TB: 32x 256 GB DRAM
Graphics	8x HGX H100 SXM5 GPUs (80GB, 700W TDP)
Storage (additional storage available)	8x 2.5" SATA 8x 2.5" NVMe U.2 Via PCIe Switches Additional 8x 2.5" NVMe U.2 Via PCIe Switches (option) 2x NVMe M.2
Power	3+3 Redundant 6x 3000W Titanium Level Efficiency Power Supplies *These are max Turbo frequencies

Building Block Server: GPU Super Server AS -8125GS-TNHR



Overview	8U Dual Socket (4th Gen AMD EPYC™), up to 8 SXM5 GPUs
CPU	2x 4th Gen AMD EPYC™ Processors
Memory (additional memory available)	24 DIMM slots Up to 6TB ECC DDR5-4800 RDIMM
Graphics	8x HGX H100 SXM5 GPUs (80GB, 700W TDP)
Storage (additional storage available)	8x 2.5" SATA 8x 2.5" NVMe U.2 Via PCIe Switches Additional 8x 2.5" NVMe U.2 Via PCIe Switches (option) 2x NVMe M.2
Power	3+3 Redundant 6x 3000W Titanium Level Efficiency Power Supplies

Building Block Switch: NDR 400G IB SSE-MQM9700-NS2F

Overview	64-ports NDR, 32 x NDR 400Gb/s OSFP ports, managed, power-to-connector (P2C) airflow (forward)
Max Throughput	51.2 Tb/s
Power	Typical power with passive cables (ATIS): 747W Max power with active cables: 1,703W

Supermicro Liquid Cooling Solution

A critical component of an advanced AI solution is a liquid cooling system that works at the rack level, which can decrease the PUE of a data center. The Supermicro Liquid Cooled Solution consists of several components, all available from Supermicro. (www.supermicro.com/liquid-cooling) The main components of an effective liquid cooling solution include:

- **CDU – Cooling Distribution Unit:**

The Supermicro CDU is at the heart of the system. The Supermicro CDU is designed to ensure that high-performing servers' entire racks perform as expected. This CDU is designed with dual redundant and hot-swappable coolant pumps and power supplies.

Figure 1 - Supermicro CDU

- **Cold Plates:**

To cool the hot CPU and GPU chips, the heat must be removed from the packages. This is accomplished by applying a device to the top of the chip, through which the liquid will flow and carry away the heat. The cool liquid is passed over the CPU or GPU, warming up as the liquid traverses the chip, removing the heat.

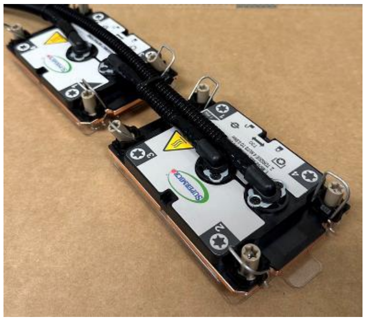


Figure 2 - Supermicro Cold Plates in Parallel

- **CDM – Cooling Distribution Manifold:**

  The purpose of the CDM is to deliver cold liquid from the CDU to each server and send the hot liquid back to the CDU. CDMs can take a number of different forms. Below is a horizontal CDM that serves two servers, one above the CDM and one below.

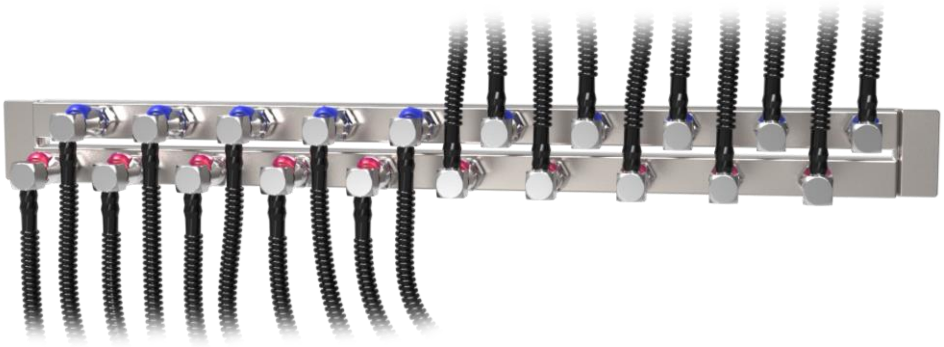


Figure 3 - Horizontal CDM

- **Leak Proof Connectors**

  When working with a liquid cooled system, it is essential to ensure that liquid does not come in contact with electronic surfaces that it is not supposed to. Leak-proof connectors allow for maintenance, with the possibility of leakage.

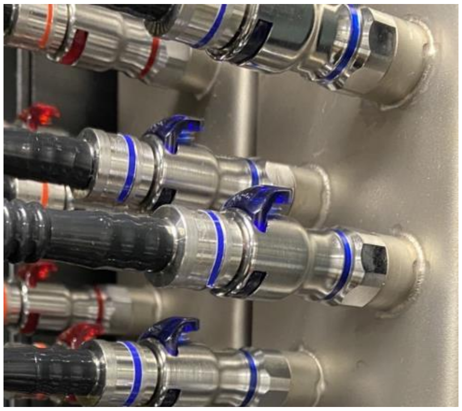


Figure 4 - Leak Proof Connectors

- **Optimal Hoses**

  With the Supermicro Rack Scale Liquid Cooling solution, hose lengths are optimized for the server that is being liquid cooled and the position of the server within the rack.

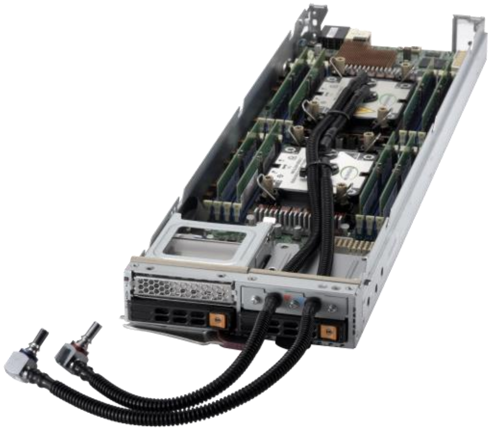


Figure 5 - Supermicro BigTwin(R) with Liquid Cooling Cold Plates and Hoses

Supermicro Advantages with Scale AI Solutions Plug and Play

One-Stop-Shop: From initial cluster design (extreme optimization for end user DL applications), assembly and configuration, testing and validation, delivery and deployment, all the way up to support and service – Supermicro is the ultimate one-stop-shop for AI infrastructure building.

Supermicro's comprehensive AI packages are entirely tested and validated at rack scale. Extensive testing includes L10 (system level tests), L11 (cluster level tests), and L12 (application level optimization and benchmarking).

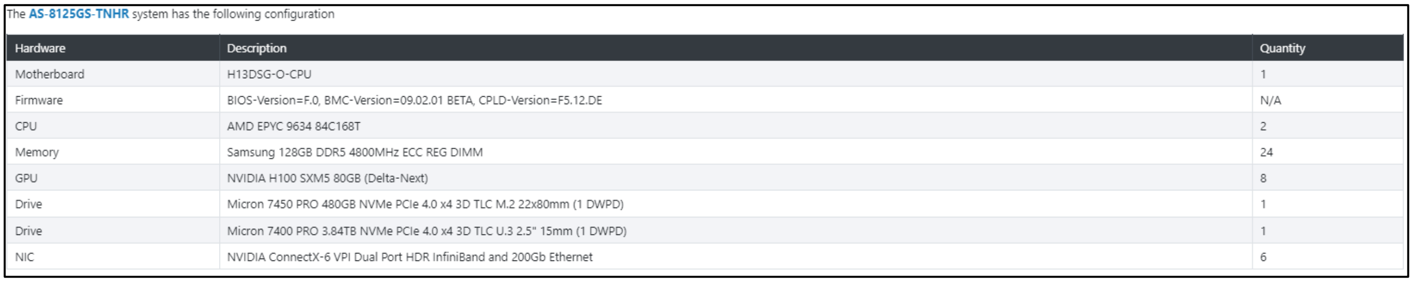
Further Information

To learn more about Supermicro's Rack Scale AI Solutions, please visit:

- Supermicro Rack Integration Services: https://www.supermicro.com/en/solutions/rack-integration
- Supermicro GPU Server Product Lines: https://www.supermicro.com/en/products/gpu
- Intel 4th Gen Xeon Scalable Processor Lineup: https://www.supermicro.com/en/support/resources/cpu-4th-gen-intel-xeon-scalable

Appendix

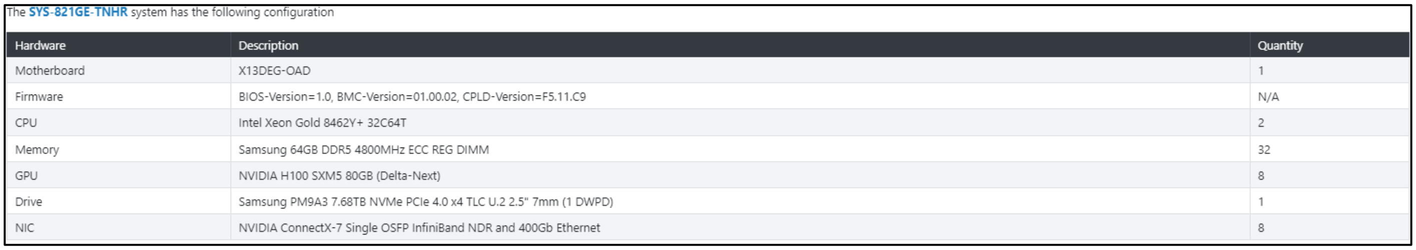
Config 1 -



The AS-8125GS-TNHR system has the following configuration

Hardware	Description	Quantity
Motherboard	H13DSG-O-CPU	1
Firmware	BIOS-Version=F.0, BMC-Version=09.02.01 BETA, CPLD-Version=F5.12.DE	N/A
CPU	AMD EPYC 9634 84C168T	2
Memory	Samsung 128GB DDR5 4800MHz ECC REG DIMM	24
GPU	NVIDIA H100 SXM5 80GB (Delta-Next)	8
Drive	Micron 7450 PRO 480GB NVMe PCIe 4.0 x4 3D TLC M.2 22x80mm (1 DWPD)	1
Drive	Micron 7400 PRO 3.84TB NVMe PCIe 4.0 x4 3D TLC U.3 2.5" 15mm (1 DWPD)	1
NIC	NVIDIA ConnectX-6 VPI Dual Port HDR InfiniBand and 200Gb Ethernet	6

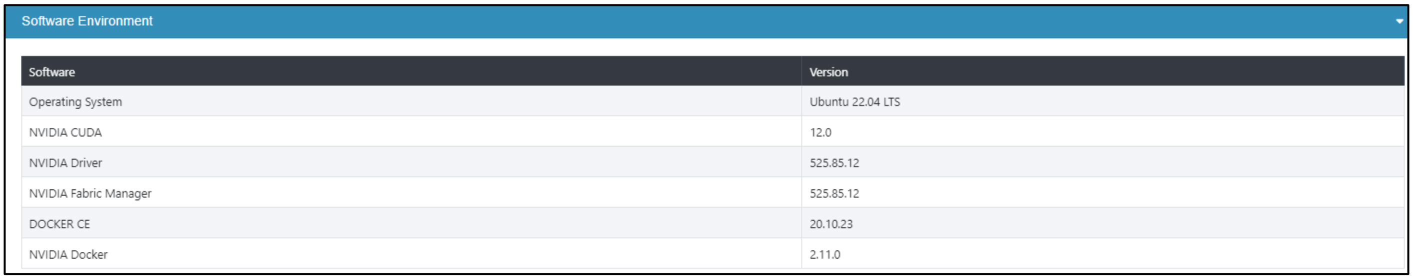
Config 2 -



The SYS-821GE-TNHR system has the following configuration

Hardware	Description	Quantity
Motherboard	X13DEG-OAD	1
Firmware	BIOS-Version=1.0, BMC-Version=01.00.02, CPLD-Version=F5.11.C9	N/A
CPU	Intel Xeon Gold 8462Y+ 32C64T	2
Memory	Samsung 64GB DDR5 4800MHz ECC REG DIMM	32
GPU	NVIDIA H100 SXM5 80GB (Delta-Next)	8
Drive	Samsung PM9A3 7.68TB NVMe PCIe 4.0 x4 TLC U.2 2.5" 7mm (1 DWPD)	1
NIC	NVIDIA ConnectX-7 Single OSFP InfiniBand NDR and 400Gb Ethernet	8

Software specifications for benchmarks



Software Environment

Software	Version
Operating System	Ubuntu 22.04 LTS
NVIDIA CUDA	12.0
NVIDIA Driver	525.85.12
NVIDIA Fabric Manager	525.85.12
DOCKER CE	20.10.23
NVIDIA Docker	2.11.0