Lingvo 24

Business Assistance Centre

presents:

Your Trusted Partner for Supermicro AI Solutions

We are representing

Taiwan-based company specialising in high-performance computing solutions, with a primary focus on NVIDIA GPUs as the core component of our offerings. As authorised Supermicro resellers, we deliver the best server solutions, known for their superior quality, lowest defective rates, and exceptional stability. We bring cutting-edge AI infrastructure to the British and European market, ensuring faster delivery times than many competitors while providing powerful, scalable, and reliable solutions for businesses, research institutions, and innovators.

Based in Taiwan — a global hub for advanced technology and manufacturing—we combine expertise, quality, and speed to efficiently serve our European customers.

What We Offer

We specialise in delivering NVIDIA-powered solutions designed to accelerate AI innovation

- NVIDIA-Powered Solutions
- Supermicro Servers
- Streamlined Hardware Procurement and Delivery
- Authorized Supermicro Reseller
- Faster Delivery Times
- Taiwanese Expertise
- British Sales Support
- Future-Readyабзаца

Generative AI SuperCluster

With 256 NVIDIA HGX™ H100/H200 GPUs, 32 8U Air-cooled Systems

Industry leading Scalable Compute Unit Built For Large Language Models

- Proven industry leading architecture for large scale AI infrastructure deployments
- 256 NVIDIA H100/H200 GPUs in one scalable unit
- 20TB of HBM3 with H100 or 36TB of HBM3e with H200 in one scalable unit
- 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage for training large language model with up to trillions of parameters
- Customizable AI data pipeline storage fabric with industry leading parallel file system options
- Supports NVIDIA Quantum-2 InfiniBand and Spectrum™-X Ethernet platform
- Certified for NVIDIA AI Enterprise Platform including NVIDIA NIM microservices

Building Blocks for Highest Density Generative Al Infrastructure Deployment

In the era of Al, a unit of compute is no longer measured by just the number of servers. Interconnected GPUs, CPUs, memory, storage, and these resources across multiple nodes in racks construct today's artificial Intelligence. The infrastructure requires high-speed and low-latency network fabrics, and carefully designed cooling technologies and power delivery to sustain optimal performance and efficiency for each data center environment. Supermicro's SuperCluster solution provides foundational building blocks for rapidly evolving Generative Al and Large Language Models (LLMs). The full turn-key data center solution accelerates time-to-delivery for mission-critical enterprise use cases, and eliminates the complexity of building a large cluster, that used to be only achievable through intensive design tuning and time-consuming optimization of supercomputing.

8U 8-GPU System

Supermicro's proven industry-leading 8U system is powering NVIDIA HGX H100/H200 8-GPU at its full potential. 8 of PCIe 5.0 slots are dedicated to 1:1 400Gb/s networking for GPUs. Each GPU is paired with 400Gb/s networking such as NVIDIA ConnectX-7 to enable NVIDIA GPUDirect RDMA and Storage so that the data flows directly to the GPU memory with the lowest latency possible.

The NVIDIA HGX H100/H200 8-GPU equipped system is ideal for training Generative Al. The high-speed interconnected GPUs through NVIDIA® NVLink®, high GPU memory bandwidth and capacity are the keys to running large language (LLM) models cost-effectively. The SuperCluster creates a massive pool of GPU resources acting as one AI supercomputer.

Plug-and-Play, Reduced Lead-time

The SuperCluster design with the 8U air-cooled systems comes with 400Gb/s networking fabrics and non-blocking architecture. The 4 nodes per rack and 32-node cluster operate as a scalable unit of compute providing a foundational building block for Generative AI Infrastructure.

Whether fitting an enormous foundation model trained on a dataset with trillions of tokens from scratch, or building a cloud- scale LLM inference infrastructure, the spine and leaf network topology allows it to scale from 32 nodes to thousands of nodes seamlessly. Supermicro's proven testing processes thoroughly validate the operational effectiveness and efficiency before shipping. Customers receive plug-and-play scalable units for rapid deployment.

Rack Scale Design Close-up

Net working

- 400G InfiniBand NDR leaf switches dedicated for compute and storage
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network
- Leaf switches in the dedicated networking rack or in the individual compute racks

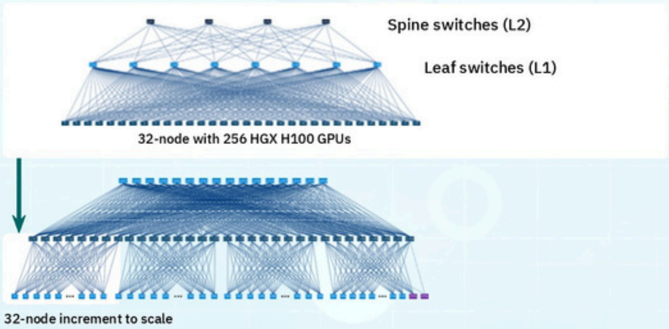
Compute and Storage

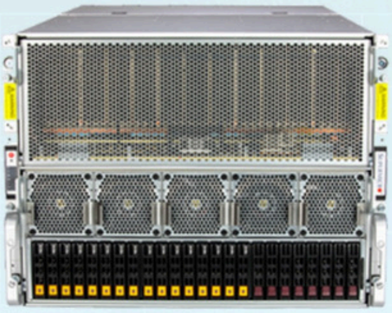
- 4x SYS-821GE-TNHR or AS -8125GS- TNHR per rack
- 4x NVIDIA HGX H100/H200 8-GPU per rack
- 32x NVIDIA H100/H200 Tensor Core GPUs
- 5TB of HBM3 or 9TB of HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage support

32-Node LLM Scalable Unit

The spine-leaf network fabric allows 32-node compute unit as a increment to scale to thousands of nodes. With highest network performance achievable for GPU-GPU connectivity, the SuperCluster is optimized for LLM training and high volume, high batch size inference. Plus, our L11 and L12 validation testing, and on-site deployment service provides seamless experience.

Network Fabrics





Node Configuration

SYS-821GE-TNHR / AS -8125GS-TNHR

Overview	8U Air-cooled System with NVIDIA HGX H100/H200 8-GPU
CPU	Dual 5th/4th Gen Intel® Xeon® or AMD EPYC 9004 Series Processors
Memory	2TB DDR5 (recommended)
GPU	NVIDIA HGX H100/H200 8-GPU (80GB HBM3 or 141GB HBM3e per GPU) 900GB/s NVLink GPU-GPU interconnect with NVSwitch
Networking	8x NVIDIA ConnectX®-7 Single-port 400Gbps/NDR OSFP NICs 2x NVIDIA ConnectX-7 Dual-port 200Gbps/NDR200 QSFP112 NICs 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage
Storage	30.4TB NVMe (4x 7.6TB U.3) 3.8TB NVMe (2x 1.9TB U.3, Boot) [Optional M.2 available]
Power Supply	6x 3000W Redundant Titanium Level power supplies

*Recommended configuration, other system memory, networking, storage options are available.

32-Node Scalable Unit

SRS-48UGPU-AI-ACSU

Overview	Fully integrated air-cooled 32-node cluster with 256 H100/H200 GPUs
Compute Fabric Leaf	8x SSE-MQM9700-NS2F, 64-port NVIDIA Quantum-2 InfiniBand 400G NDR, 32 OSFP ports switch
Compute Fabric Spine	4x SSE-MQM9700-NS2F, 64-port NVIDIA Quantum-2 InfiniBand 400G NDR, 32 OSFP ports switch
In-band Management Switch	2x SSE-MSN4600-CS2FC 64-port 100GbE QSFP28, 2U switch
Out-of-band Management Switch	2x SSE-G3748R-SMIS, 48-port 1Gbps Ethernet ToR management switch 1x SSE-F3548SR, 48-port 10Gbps Ethernet ToR management switch
Rack	9x 48U 750mm x 1200mm
PDU	34x 208V 60A 3Ph

*Recommended configuration, other network switch options and rack layouts are available, including configuration supporting NVIDIA Spectrum-X Ethernet.
*Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional

GPU A+ Server AS -8125GS-TNHR

DP AMD 8U System with NVIDIA HGX H100/H200 8-GPU

Key Applications

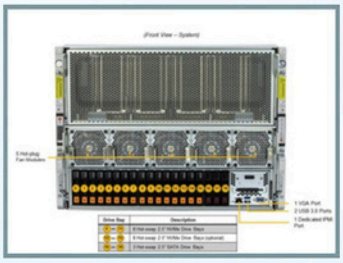
High Performance Computing, AI/Deep Learning Training, Industrial Automation, Retail, Climate and Weather Modeling,

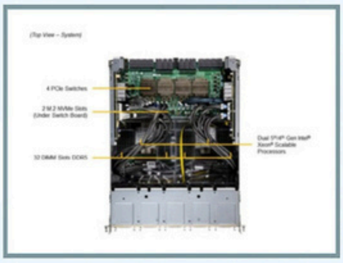
Key Features

- High density 8U system for NVIDIA® HGX™ H100/H200 8-GPU Highest GPU communication using NVIDIA® NVLINK™ + NVIDIA® NVSwitch™ 8 NIC for GPU direct RDMA (1:1 GPU Ratio);
- 24 DIMM slots DDR5; up to 6TB 4800MT/s ECC LRDIMM/RDIMM;
- Up to 8 PCIe 5.0 x16 LP + 4 PCIe 5.0 x16 FHFL slots;
- Flexible networking options;
- 12 Hot-swap 2.5" NVMe drive bays + 2 hot-swap 2.5" SATA drive bays
+ 4 hot-swap 2.5" NVMe drive bays (optional)
1 M.2 NVMe for boot drive only;
- 10 heavy duty fans with optimal fan speed control;
- 6x 3000W redundant Titanium level power supplies;

Form Factor	8U Rackmount Enclosure: 437 x 355.6 x 843.28mm (17.2" x 14" x 33.2") Package: 698 x 750 x 1300mm (27.5" x 29.5" x 51.2")
Processor	Dual processor(s) AMD EPYC™ 9004/9005 Series Processors (* AMD EPYC™ 9005 Series drop-in support requires board revision 2.x) Up to 128C/256T
GPU	Max GPU Count: Up to 8 onboard GPUs Supported GPU: NVIDIA SXM: HGX H100 8-GPU (80GB), HGX H200 8-GPU (141GB) CPU-GPU Interconnect: PCIe 5.0 x16 CPU-to-GPU Interconnect GPU-GPU Interconnect: NVIDIA® NVLink® with NVSwitch™
System Memory	Slot Count: 24 DIMM slots Max Memory (1DPC): Up to 6TB 4800MT/s ECC DDR5 RDIMM
Drive Bays Configuration	Default: Total 18 bays • 2 front hot-swap 2.5" SATA drive bays • 4 front hot-swap 2.5" NVMe* drive bays • 12 front hot-swap 2.5" NVMe drive bays (*NVMe support may require additional storage controller and/or cables) M.2: 1 M.2 NVMe slot (M-key)
Expansion Slots	Default • 8 PCIe 5.0 x16 LP slots • 2 PCIe 5.0 x16 FHFL slots Option A • 8 PCIe 5.0 x16 LP slots • 4 PCIe 5.0 x16 FHFL slots
On-Board Devices	AMD SP5
Input / Output	1 VGA port
System Cooling	Fans: 10 heavy duty fans with optimal fan speed control
Power Supply	6x 3000W Redundant Titanium Level (96%) power supplies

System BIOS	6x 3000W Redundant Titanium Level (96%) power supplies
Management	SuperCloud Composer; Supermicro Server Manager (SSM); Supermicro Update Manager (SUM); Supermicro SuperDoctor® 5 (SD5); Super Diagnostics Offline (SDO); Supermicro Thin-Agent Service (TAS); SuperServer Automation Assistant (SAA) New!
PC Health Monitoring	CPU: Monitors for CPU Cores, Chipset Voltages, Memory 7 +1 Phase-switching voltage regulator FAN: Fans with tachometer monitoring Status monitor for speed control Temperature: Monitoring for CPU and chassis environment Thermal Control for fan connectors
Dimensions and Weight	Weight: Gross Weight: 225 lbs (102.1 kg) Net Weight: 166 lbs (75.3 kg) Available Color: Black front & silver body
Operating Environment	Operating Temperature: 10°C ~ 35°C (50°F ~ 95°F) Non-operating Temperature: -40°C to 60°C (-40°F to 140°F) Operating Relative Humidity: 8% to 90% (non-condensing) Non-operating Relative Humidity: 5% to 95% (non-condensing)
Motherboard	**Super H13DSG-O-CPU-D**
Chassis	CSE-GP801TS




(Top View – System)
4 PCIe Switches
2 M.2 NVMe Slots (Under Switch Board)
32 DIMM Slots DDR5
Dual 5th/4th Gen Intel® Xeon® Scalable Processors

Generative AI SuperCluster

With 256 NVIDIA HGX™ H100/H200 GPUs, 32 4U Liquid-cooled Systems

Scalable Compute Unit Built For Large Language Models - Available from Favortron

- Doubling compute density through Supermicro's custom liquid-cooling solution with up to 40% reduction in electricity cost for data center
- 256 NVIDIA H100/H200 GPUs in one scalable unit
- 20TB of HBM3 with H100 or 36TB of HBM3e with H200 in one scalable unit
- 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage for training large language model with up to trillions of parameters
- Customizable AI data pipeline storage fabric with industry leading parallel file system options
- Supports NVIDIA Quantum-2 InfiniBand and Spectrum™-X Ethernet platform
- Certified for NVIDIA AI Enterprise Platform including NVIDIA NIM microservices

Building Blocks for Highest Density Generative AI Infrastructure Deployment

In the era of AI, a unit of compute is no longer measured by just the number of servers. Interconnected GPUs, CPUs, memory, storage, and these resources across multiple nodes in racks construct today's artificial Intelligence. The infrastructure requires high-speed and low-latency network fabrics, and carefully designed cooling technologies and power delivery to sustain optimal performance and efficiency for each data center environment. Supermicro's SuperCluster solution provides foundational building blocks for rapidly evolving Generative AI and Large Language Models (LLMs). The full turn-key data center solution accelerates time-to-delivery for mission-critical enterprise use cases, and eliminates the complexity of building a large cluster, that used to be only achievable through intensive design tuning and time-consuming optimization of supercomputing.

4U 8-GPU System, Liquid-cooled

Supermicro 4U liquid-cooled system with NVIDIA HGX H100/ H200 8-GPU doubles the density of the 8U air-cooled system. Our custom direct-to-chip (D2C) cold plates keep both GPUs and CPUs at optimal temperature for sustained maximum performance. Supermicro cooling distribution unit (CDU) and manifold (CDM) are the main arteries for distributing cooled liquid to the cold plates, enabling up to 40% reduction in electricity costs for the entire data center, reducing server noise, and saving data center space.

The NVIDIA HGX H100/H200 8-GPU equipped system is ideal for training Generative AI. The high-speed interconnected GPUs through NVIDIA® NVLink®, high GPU memory bandwidth and capacity are the key for running large language (LLM) models cost effectively. The SuperCluster creates a massive pool of GPU resources acting as one AI supercomputer.

Plug-and-Play, Reduce Lead-time

The SuperCluster design with the 4U liquid-cooled systems comes with 400Gb/s networking fabrics and non-blocking architecture. The 8 nodes per rack and 32-node cluster operate as a scalable unit of compute providing a foundational building block for Generative AI Infrastructure.

Whether fitting an enormous foundation model trained on a dataset with trillions of tokens from scratch, or building a cloud-scale LLM inference infrastructure, the spine and leaf network topology allows it to scale from 32 nodes to thousands of nodes seamlessly. With fully integrated liquid-cooling out of the box, Supermicro's proven testing processes thoroughly validate the operational effectiveness and efficiency before shipping. Customers receive plug-and-play scalable units for rapid deployment.

Rack Scale Design Close-up

Net working

- 400G InfiniBand NDR leaf switches dedicated for compute and storage
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch Non-blocking network

Compute and Storage

- 8x SYS-421GE-TNHR2-LCC or AS -4125GS-TNHR2-LCC per rack
- 8x NVIDIA HGX H100/H200 8-GPU per rack
- 64x NVIDIA H100/H200 Tensor Core GPUs
- 5TB of HBM3 or 9TB of HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and storage support

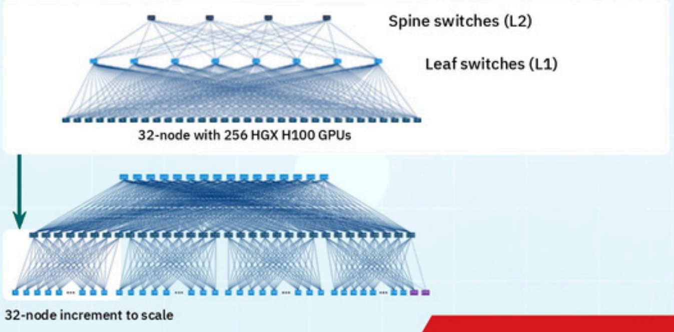
CDU and CDM

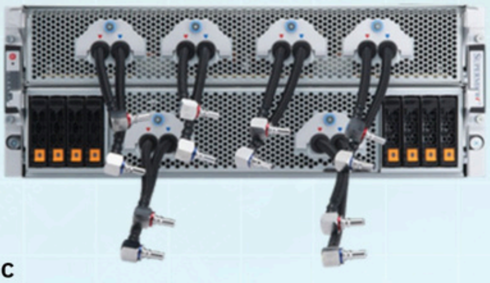
- Supermicro 100kW capacity Cooling Distribution Unit with redundant PSU and dual hot-swap pumps
- 8x 1U Supermicro Cooling Distribution Manifold

32-Node LLM Scalable Unit

The spine-leaf network fabric allows 32-node compute unit as a increment to scale to thousands of nodes. With highest network performance achievable for GPU-GPU connectivity, the SuperCluster is optimized for LLM training and high volume, high batch size inference. Plus, our L11 and L12 validation testing, and on-site deployment service provides seamless experience.

Network Fabrics





Node Configuration

SYS-421GE-TNHR2-LCC / AS-4125GS-TNHR2-LCC

Overview	4U Liquid-cooled System with NVIDIA HGX H100/H200 8-GPU
CPU	Dual 5th/4th Gen Intel® Xeon® or AMD EPYC™ 9004 Series Processors
Memory	2TB DDR5 (recommended)
GPU	NVIDIA HGX H100/H200 8-GPU (80GB HBM3 or 141GB HBM3e per GPU) 900GB/s NVLink GPU-GPU interconnect with NVSwitch
Networking	8x NVIDIA ConnectX®-7 Single-port 400Gbps/NDR OSFP NICs 2x NVIDIA ConnectX®-7 Dual-port 200Gbps/NDR200 QSFP112 NICs 1:1 networking to each GPU to enable NVIDIA GPUDirect RDMA and Storage
Storage	30.4TB NVMe (4x 7.6TB U.3) 3.8TB NVMe (2x 1.9TB U.3, Boot) [Optional M.2 available]
Power Supply	4x 5250W Redundant Titanium Level power supplies

*Recommended configuration, other system memory, networking, storage options are available.

32-Node Scalable Unit

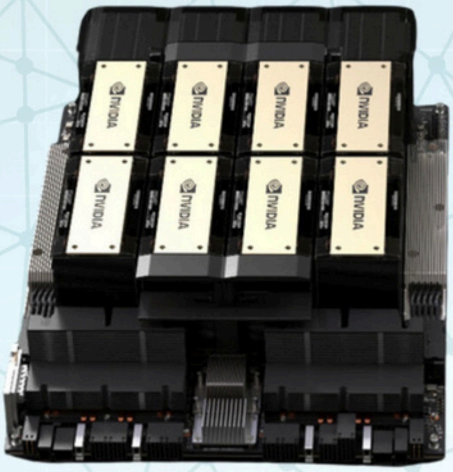
SRS-48UGPU-AI-LCSU

Overview	Fully integrated liquid-cooled 32-node cluster with 256 NVIDIA H100/H200 GPUs
Compute Fabric Leaf	8x SSE-MQM9700-NS2F, 64-port NVIDIA Quantum-2 InfiniBand 400G NDR, 32 OSFP ports switch
Compute Fabric Spine	4x SSE-MQM9700-NS2F, 64-port NVIDIA Quantum-2 InfiniBand 400G NDR, 32 OSFP ports switch
In-band Management Switch	3x SSE-MSN4600-CS2FC 64-port 100GbE QSFP28, 2U switch
Out-of-band Management Switch	8x SSE-MQM9700-NS2F, 64-port NVIDIA Quantum-2 InfiniBand 400G NDR, 32 OSFP ports switch
Rack and PDU	5x 48U 750mm x 1200mg PDU: 18x 415V 60A 3Ph
Liquid Cooing	4x Supermicro 80kW capacity CDU with redundant PSU and dual hot-swap pumps

*Recommended configuration, other network switch options and rack layouts are available, including configuration supporting NVIDIA Spectrum-X Ethernet.
*Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional

NVIDIA H200 Tensor Core GPU

Supercharging AI and HPC workloads.



Higher Performance With Larger, Faster Memory

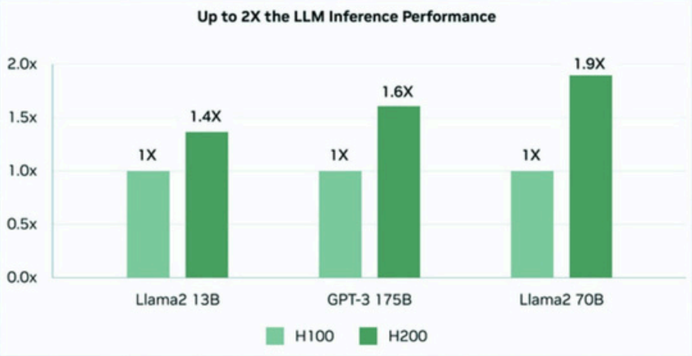
The NVIDIA H200 Tensor Core GPU supercharges generative AI and high- performance computing (HPC) workloads with game-changing performance and memory capabilities. Based on the **NVIDIA Hopper™ architecture**, the NVIDIA H200 is the first GPU to offer 141 gigabytes (GB) of HBM3e memory at 4.8 terabytes per second (TB/s)—that's nearly double the capacity of the **NVIDIA H100 Tensor Core GPU** with 1.4X more memory bandwidth. The H200's larger and faster memory accelerates generative AI and large language models, while advancing scientific computing for HPC workloads with better energy efficiency and lower total cost of ownership.

Key Features

- > 141GB of HBM3e GPU memory
- > 4.8TB/s of memory bandwidth
- > 4 petaFLOPS of FP8 performance
- > 2X LLM inference performance
- > 110X HPC performance

Unlock Insights With High-Performance LLM Inference

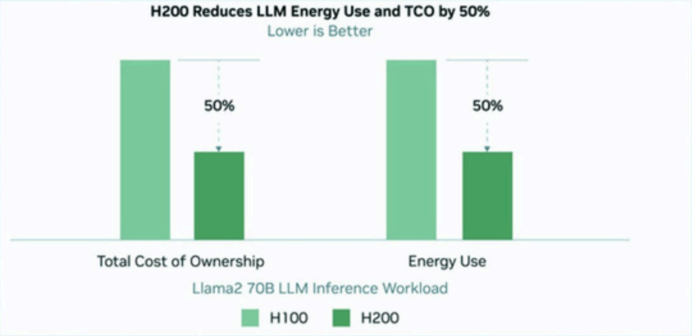
In the ever-evolving landscape of AI, businesses rely on large language models to address a diverse range of inference needs. An **AI inference** accelerator must deliver the highest throughput at the lowest TCO when deployed at scale for a massive user base.
The H200 doubles inference performance compared to H100 GPUs when handling large language models such as Llama2 70B.



Preliminary specifications. May be subject to change. Llama2 13B: ISL 128, OSL 2K | Throughput | H100 SXM 1x GPU BS 64 | H200 SXM 1x GPU BS 128 GPT-3 175B: ISL 80, OSL 200 | x8 H100 SXM GPUs BS 64 | x8 H200 SXM GPUs BS 128 Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1x GPU BS 8 | H200 SXM 1x GPU BS 32.

Reduce Energy and TCO

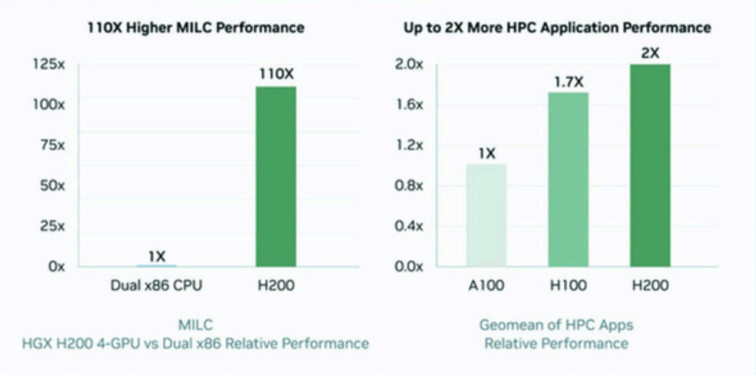
With the introduction of H200, energy efficiency and TCO reach new levels. This cutting-edge technology offers unparalleled performance, all within the same power profile as the **H100 Tensor Core GPU.** AI factories and supercomputing systems that are not only faster but also more eco-friendly deliver an economic edge that propels the AI and scientific communities forward.



Preliminary specifications. May be subject to change. Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 SXM 1x GPU BS 8 | H200 SXM 1x GPU BS 32

Supercharge High-Performance Computing

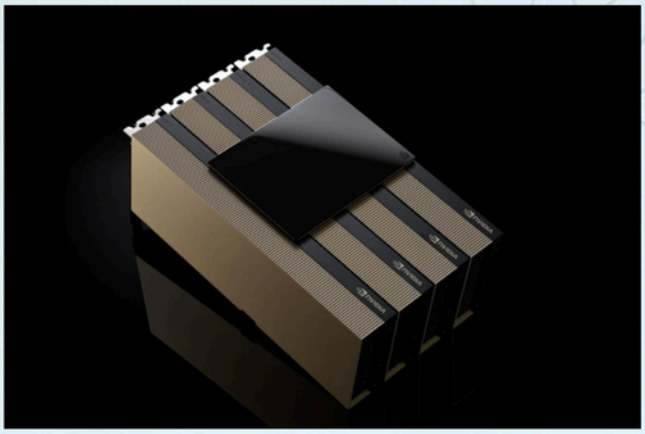
Memory bandwidth is crucial for HPC applications, as it enables faster data transfer and reduces complex processing bottlenecks. For memory-intensive HPC applications like simulations, scientific research, and artificial intelligence, the H200's higher memory bandwidth ensures that data can be accessed and manipulated efficiently, leading to 110X faster time to results.



Preliminary specifications. May be subject to change. HPC MILC- dataset NERSC Apex Medium | HGX H200 4-GPU | dual Sapphire Rapids 8480 HPC Apps- CP2K: dataset H2O-32-RI-dRPA-96points | GROMACS: dataset STMV | ICON: dataset r2b5 | MILC: dataset NERSC Apex Medium | Chroma: dataset HMC Medium | Quantum Espresso: dataset AUSURF112 | 1x H100 SXM | 1x H200 SXM.

AI Acceleration for Mainstream Enterprise Servers With H200 NVL

NVIDIA H200 NVL is ideal for lower-power, air-cooled enterprise rack designs that require flexible configurations, delivering acceleration for every AI and HPC workload regardless of size. With up to four GPUs connected by **NVIDIA NVLink™** and a 1.5X memory increase, large language model (LLM) inference can be accelerated up to 1.7X and HPC applications achieve up to 1.3X more performance over the H100 NVL.

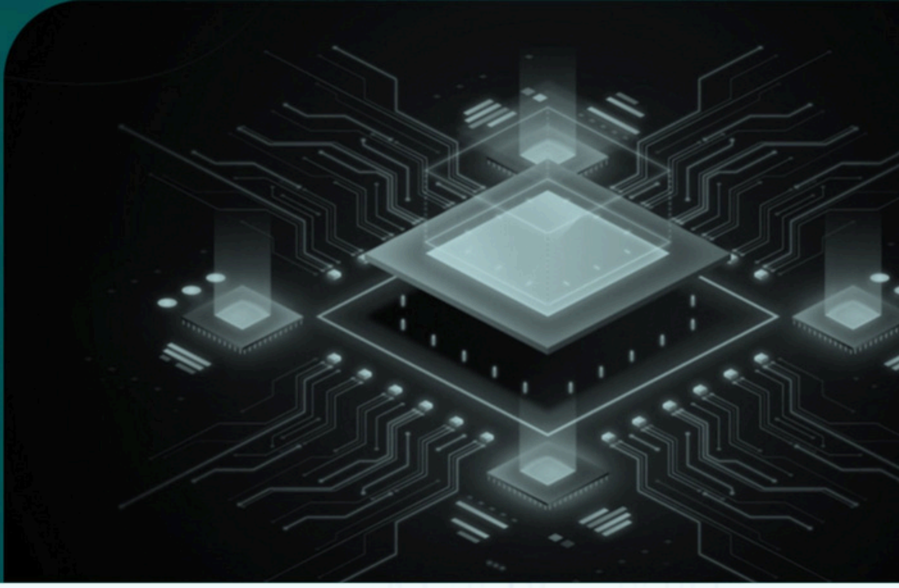


Enterprise-Ready: AI Software Streamlines Development and Deployment

NVIDIA H200 NVL comes with a five-year **NVIDIA AI Enterprise** subscription and simplifies the way you build an enterprise AI-ready platform. H200 accelerates AI development and deployment for production-ready generative AI solutions, including computer vision, speech AI, retrieval augmented generation (RAG), and more. NVIDIA AI Enterprise includes **NVIDIA NIM™**, a set of easy-to-use microservices designed to speed up enterprise generative AI deployment. Together, deployments have enterprise-grade security, manageability, stability, and support. This results in performance-optimized AI solutions that deliver faster business value and actionable insights.

Technical Specifications

	H200 SXM1	H200 NVL1
FP64	34 TFLOPS	30 TFLOPS
FP64 Tensor Core	67 TFLOPS	60 TFLOPS
FP32	67 TFLOPS	60 TFLOPS
TF32 Tensor Core2	989 TFLOPS	835 TFLOPS
BFLOAT16 Tensor Core2	1,979 TFLOPS	1,671 TFLOPS
FP16 Tensor Core2	1,979 TFLOPS	1,671 TFLOPS
FP8 Tensor Core2	3,958 TFLOPS	3,341 TFLOPS
INT8 Tensor Core2	3,958 TFLOPS	3,341 TFLOPS
GPU Memory	141GB	141GB
GPU Memory Bandwidth	4.8TB/s	4.8TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Confidential Computing	Supported	Supported
Max Thermal Design Power (TDP)	Up to 700W (configurable)	Up to 600W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @18GB each	Up to 7 MIGs @16.5GB each
Form Factor	SXM	PCIe Dual-slot air-cooled
Interconnect	NVIDIA NVLink: 900GB/s PCIe Gen5: 128GB/s	2- or 4-way NVIDIA NVLink bridge: 900GB/s per GPU PCIe Gen5: 128GB/s
Server Options	NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs	NVIDIA MGX™ H200 NVL partner and NVIDIA-Certified Systems with up to 8 GPUs
NVIDIA AI Enterprise	Add - on	Included



Supermicro NVIDIA GB200 NVL72

Liquid-cooled Exascale Compute in a Rack with 72 NVIDIA Blackwell GPUs

Scalable Compute Unit Built For Trillion Parameter AI Models

- **72 NVIDIA Blackwell GPUs**: acting as one GPU with a massive pool of HBM3e memory to deliver the most efficient exascale computing in a rack
- **Pioneers in Liquid Cooling**: total liquid-cooling solution with up to 40% reduction in electricity cost for data center
- **Unmatched Manufacturing Scale**: with the largest liquid cooling rack-level manufacturing capacity, Supermicro ensures timely and high-quality deployment of the GB200 NVL72, supported by production facilities in San Jose, CA, Europe, and Asia
- **Comprehensive Service Offering**: from proof of concept to full-scale deployment, Supermicro is one-stop shop, providing all necessary parts, networking solutions, and on-site installation ser vices
- **Advanced Networking Ready**: Supermicro is at the forefront of adopting NVIDIA BlueField®-3 SuperNIC, Spectrum™-X, Quantum-2, and next generation 800 Gb/s networking platforms

An Exascale Supercomputer in a Rack

Supermicro accelerates the industry's transition to liquid-cooled data centers with NVIDIA Blackwell to deliver a new paradigm of energy-efficiency for the rapidly heightened energy demand of AI infrastructure. With extensive experience deploying large scale direct-to-chip (DLC) liquid-cooled AI systems, Supermicro's leading liquid-cooling technology advancement powers NVIDIA GB200 NVL72, an exascale computing in a single rack, providing up to 25 times more energy efficiency than the previous generation.

Powered by Supermicro End-to-End Liquid- cooling Solution

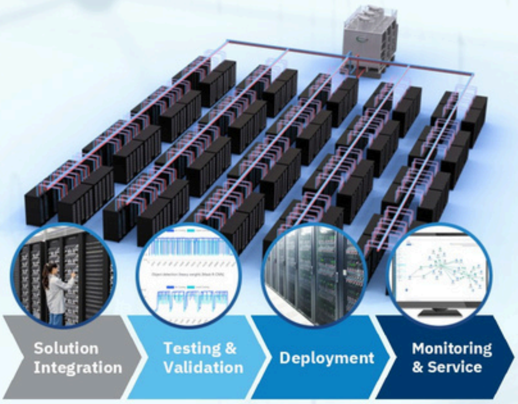
Supermicro NVIDIA GB200 NVL72 SuperCluster features the new advanced in-rack or in-row cooling distribution unit (CDU) and custom coldplates designed for the compute trays housing the NVIDIA GB200 Grace™ Blackwell Superchips. The NVIDIA GB200 NVL72 delivers exascale computing capabilities in a single rack with fully integrated liquid-cooling. It incorporates 72 NVIDIA Blackwell GPUs and 36 Grace CPUs interconnected by NVIDIA's largest NVLink™ network to date. The NVLink Switch System facilitates 130 terabytes per second (TB/s) of total GPU communications with low latency, enhancing performance for AI and high-performance computing (HPC) workloads.

End-to-End Onsite Deployment Services

From proof-of-concept (PoC) to full-scale deployment, Supermicro is a one-stop shop, providing all necessary parts, Liquid-Cooling, networking solutions, management software, and onsite installation services. As a one-stop shop, Supermicro delivers a comprehensive, in-house Liquid-Cooling ecosystem, encompassing custom-designed cold plates optimized for various GPUs, CPUs and memory modules, along with multiple coolant distiribution unit form factors and capacity, manifolds, hoses, connectors, cooling towers, and monitoring and management software. This end-to-end solution seamlessly integrates into rack-level configurations, significantly boosting system efficiency, mitigating thermal throttling, and simultaneously reducing both the Total Cost of Ownership (TCO) and environmental impact of data center operations for the era of AI.

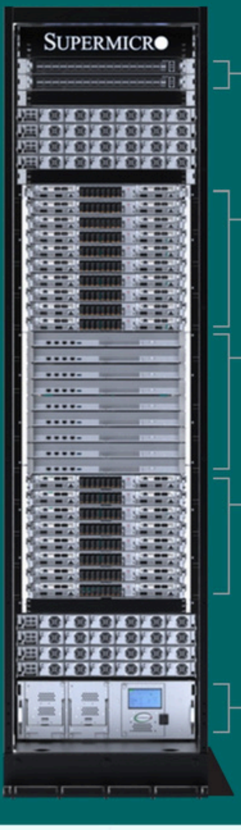
SuperCloud Composer (SCC) for Liquid- Cooled Data Center Management

Supermicro's comprehensive datacenter management platform, SuperCloud Composer software, provides powerful tools to monitor vital information on liquid- cooled systems and racks, coolant distribution units, and cooling towers, including pressure, humidity, pump and valve conditions and more. SuperCloud Composer's Liquid- Cooling Consult Module (LCCM) optimizes the operational cost and manages the integrity of liquid-cooled data centers.



Supermicro NVIDIA GB200 NVL72 SuperCluster

Rack Scale Design Close-up



Management Networking
- In-band management switch
- Out-of-band management switch

10 Compute Trays
- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Compute Interconnect
- 9x NVLink Switches
- 72 GPUs and 36 CPUs inteconnected at 1.8TB/s

Текст абзаца

8 Compute Trays
- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Liquid-Cooling Options
- Supermicro 250kW capacity coolant distribution unit (CDU) with redundant PSU and dual hot-swap pumps
- 240kW or 180kW capacity Liquid-to-air solution (no facility water rquired)

72-GPU Scalable Unit

SRS-GB200-NVL72-M1

GPUs	72x NVIDIA Blackwell B200 GPUs
CPUs	36x NVIDIA 72-core Grace Arm Neoverse V2
Compute Trays	18x 1U ARS-121GL-NBO
NVLink Switch Trays	9x NVLink Switch, 4-ports per compute tray connecting 72 GPUs to provide 1.8TB/s GPU-to-GPU interconnect
Power Shelves	8x 1U 33kW (6x 5.5kW PSUs), total power 132kW
Rack Dimensions (mm)	W 600 x D 1068 x H 2236
Liquid Cooing Options	• 1x in-rack Supermicro 250kW capacity CDU with redundant PSU and dual hot-swap pumps • 1.3MW capacity in-row CDU • 180kW/240kW capacity liquid-to-air solutions for facilities without cooling tower and water supply

Subject to change

Compute Tray

ARS-121GL-NBO

Overview	1U Liquid-cooled System with 2x NVIDIA GB200 Grace Blackwell Superchips
CPU and GPU	• 2x 72-core NVIDIA Grace Arm Neoverse V2 CPU CPU and GPU • 4x NVIDIA Blackwell B200 per Superchip
GPU Memory	Up to 384GB HBM3e per Superchip (768GB per tray)
CPU Memory	Up to 480GB LPDDR5X per Superchip (960GB per tray)
Networking	4x NVIDIA NVLink Switch ports
Storage	8x E1.S PCIe 5.0 drives
Power Supply	Shared power through 4+4 rack power shelves

Subject to change

GPU SuperServer SYS-821GE-TNHR

DP Intel 8U System with NVIDIA HGX H100/H200 8-GPU and Rear I/O

Key Applications

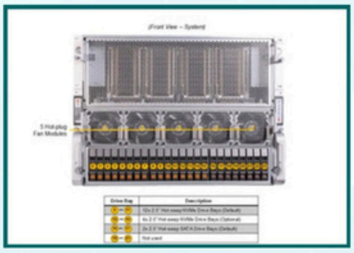
High Performance Computing, AI/Deep Learning Training, Industrial Automation, Retail, Healthcare, Conversational AI, Business Intelligence & Analytics, Drug Discovery, Climate and Weather Modeling, Finance & Economics,

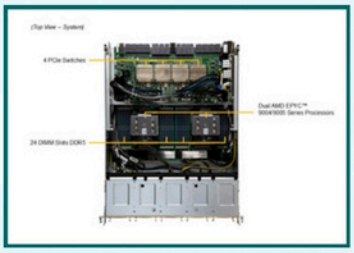
Key Features

- 5th/4th Gen Intel® Xeon® Scalable processor support;
- 32 DIMM slots Up to 8TB: 32x 256 GB DRAM Memory Type: 5600MTs ECC DDR5;
- 8 PCIe Gen 5.0 X16 LP
- 2 PCIe Gen 5.0 X16 FHHL Slots, 2 PCIe Gen 5.0 X16 FHHL Slots (optional);
- Flexible networking options;
- 2 M.2 NVMe for boot drive only 16x 2.5" Hot-swap NVMe drive bays (12x by default, 4x optional) 3x 2.5" Hot-swap SATA drive bays Optional: 8x 2.5" Hot-swap SATA drive bays;
- 10 heavy duty fans with optimal fan speed control;
- Optional: 8x 3000W (4+4) Redundant Power Supplies, Titanium Level 6x 3000W (4+2) Redundant Power Supplies, Titanium Level;

Form Factor	8U Rackmount Enclosure: 437 x 355.6 x 843.28mm (17.2" x 14" x 33.2") Package: 698 x 750 x 1300mm (27.5" x 29.5" x 51.2")
Processor	Dual Socket E (LGA-4677) 5th Gen Intel® Xeon® / 4th Gen Intel® Xeon® Scalable processors Up to 64C/128T; Up to 320MB Cache per CPU
GPU	Max GPU Count: Up to 8 onboard GPUs Supported GPU: NVIDIA SXM: HGX H100 8-GPU (80GB), HGX H200 8-GPU (141GB) CPU-GPU Interconnect: PCIe 5.0 x16 CPU-to-GPU Interconnect GPU-GPU Interconnect: NVIDIA® NVLink® with NVSwitch™
System Memory	Slot Count: 32 DIMM slots Max Memory (1DPC): Up to 4TB 5600MT/s ECC DDR5 RDIMM Max Memory (2DPC): Up to 8TB 4400MT/s ECC DDR5 RDIMM
Drive Bays Configuration	Default: Total 15 bays • 12 front hot-swap 2.5" NVMe drive bays • 3 front hot-swap 2.5" SATA drive bays Option A: Total 19 bays • 12 front hot-swap 2.5" NVMe drive bays • 4 front hot-swap 2.5" NVMe* drive bays • 3 front hot-swap 2.5" SATA drive bays (*NVMe support may require additional storage controller and/or cables, please see the optional parts list for details) M.2: 2 M.2 NVMe slots (M-key)
Expansion Slots	Default • 8 PCIe 5.0 x16 LP slots • 2 PCIe 5.0 x16 FHHL slots
On-Board Devices	Chipset: Intel® C741 Network Connectivity: • 2 RJ45 10GbE with Intel® X550-AT2 (optional) • 2 SFP28 25GbE with Broadcom® BCM57414 (optional) • 2 RJ45 10GbE with Intel® X710-AT2 (optional)
Input / Output	1 VGA port
System Cooling	Fans: 10 heavy duty fans with optimal fan speed control

Power Supply	6x 3000W Redundant (3 + 3) Titanium Level (96%) power supplies
System BIOS	BIOS Type: AMI 32MB SPI Flash EEPROM
Management	SuperCloud Composer; Supermicro Server Manager (SSM); Supermicro Update Manager (SUM); Supermicro SuperDoctor® 5 (SD5); Super Diagnostics Offline (SDO); Supermicro Thin-Agent Service (TAS); SuperServer Automation Assistant (SAA) New!
PC Health Monitoring	CPU: Monitors for CPU Cores, Chipset Voltages, Memory 8+4 Phase-switching voltage regulator FAN: Fans with tachometer monitoring Status monitor for speed control Pulse Width Modulated (PWM) fan connectors Temperature: Monitoring for CPU and chassis environment Thermal Control for fan connectors
Dimensions and Weight	Weight: Gross Weight: 225 lbs (102.1 kg) Net Weight: 166 lbs (75.3 kg) Available Color: Black front & silver body
Operating Environment	Operating Temperature: 10°C ~ 35°C (50°F ~ 95°F) Non-operating Temperature: -40°C to 60°C (-40°F to 140°F) Operating Relative Humidity: 8% to 90% (non-condensing) Non-operating Relative Humidity: 5% to 95% (non-condensing)
Motherboard	**Super X13DEG-OAD**
Chassis	CSE-GP801TS





About Lingvo 24 from the Founder.

LINGVO 24
Multilingual Investment and Business Assistance Centre

Dear Partners,

My name is Denys Yuzhakov. Nice to meet you. I'm based in the UK.
Our team has experience in online business assistance for our clients since 2010.
And we will be happy to arrange your problems in the corporate field.
We manage projects and provide many corporate services, including investment and fundraising.

We are happy to help you in the following areas:

- Investment and fundraising (we have a good investors database to work with)
- Project management
- Phone services - Call Answering
- Customer care (different languages support)
- Marketing Services (video production, social media, Youtube channels)
- Design Services
- Website Services
- Copywriting Services/Content creation (different languages)
- AI services (different languages)
- Translation and Interpreting services

PLEASE CLAIM OUR BROCHURE FOR MORE INFORMATION.

Please contact us:
marketing@lingvo24.info
https://lingvo24.net

Estonia +372 8804141 (call-back)
UK +441615540150 (call-back)