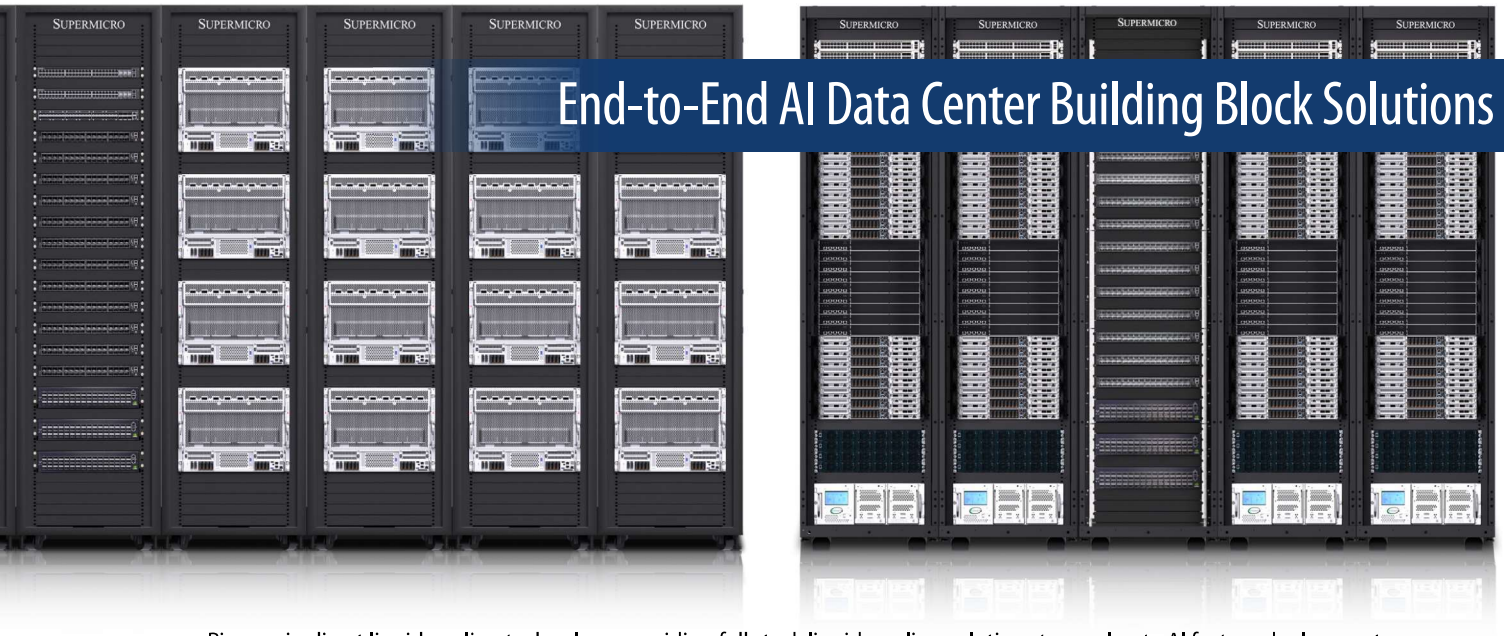
NVIDIA Blackwell Journey Starts Here

The complete multi-GPU scalable compute units built for trillion parameter AI models, directly available from Supermicro.

Complete NVIDIA Blackwell System Portfolio, Now with Blackwell Ultra

AI's transformative moment is here. Evolving scaling laws and the rise of AI reasoning continue pushing data center capabilities to new limits. Supermicro's latest NVIDIA Blackwell-powered solutions, developed through close collaboration with NVIDIA, deliver unprecedented computational performance, density, and efficiency with next-generation air-cooled and liquid-cooled architecture. With our readily deployable Data Center Building Block Solutions® (DCBBS), Supermicro is your premier partner for your NVIDIA Blackwell journey, providing sustainable, cutting-edge solutions that accelerate AI innovation.

End-to-End AI Data Center Building Block Solutions



Pioneer in direct liquid cooling technology providing full stack liquid-cooling solutions to accelerate AI factory deployment

End-to-End AI Data Center Building Block Solutions Advantage

Choose from a broad range of air-cooled and liquid-cooled systems with multiple CPU, memory, storage, networking, I/O configuration options. The comprehensive solutions include a complete data center management software suite, turn-key rack and cluster level integration with network topology design and cabling, L11/L12 validation, global delivery, on-site services and support.

Vast Experience

Supermicro Data Center Building Block Solutions power some of the largest liquid-cooled AI data center deployment in the world.

Flexible Offerings

Air or liquid-cooled, GPU-optimized, multiple system and rack form factors, CPUs, storage, and networking options, optimized for your needs.

Liquid-Cooling Pioneer

Proven, scalable, and plug-and-play liquid-cooling solutions to sustain the AI revolution. Designed specifically for NVIDIA Blackwell.

Fast Time-to-Online

Accelerated delivery with global capacity, world-class deployment expertise, and on-site services, to bring your AI to production, fast.

Front I/O 4U Liquid-cooled HGX B300 System

Networking

- NVIDIA Quantum-X800 InfiniBand or NVIDIA Spectrum-X™ Ethernet for up to 800Gb/s compute fabric
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network

Compute

- 8x SYS-422GS-NB3RT-ALC or SYS-422GS-NB3RT-LCC per rack
- 64x NVIDIA Blackwell Ultra B300 GPUs per rack
- 18.4TB of HBM3e per rack
- Dedicated storage fabric options with full NVIDIA GPUDirect RDMA and Storage or RoCE support

Liquid-Cooling

- Supermicro 250kW capacity Coolant Distribution Unit (CDU) with redundant PSU and dual hot-swap pumps
- Vertical Coolant Distribution Manifolds (CDM)

The Most Advanced DLC Technology for NVIDIA Blackwell Ultra

Pre-validated at system, rack, and cluster scale before shipping, Supermicro's NVIDIA Blackwell Ultra solutions enable turn-key day-one operation of the industry's highest performance, compute-dense AI infrastructure through Supermicro Data Center Building Block Solutions® (DCBBS). The new 4U liquid-cooled system features Supermicro's DLC-2 technology with up to 98% heat capture to achieve up to 40% data center power savings. DCBBS, combined with Supermicro's expertise in on-site deployments, provide a total solution encompassing liquid-cooling technology, network topology and cabling, power delivery, and thermal management to accelerate AI factory time-to-online.

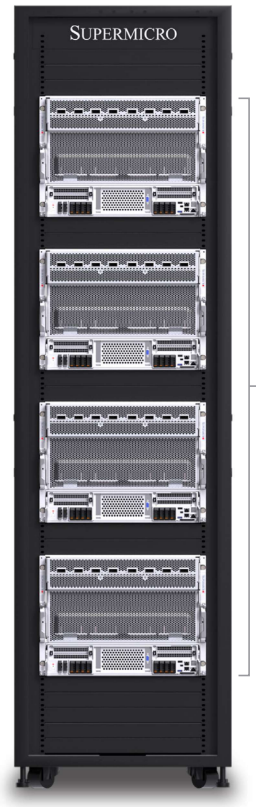
64-GPU Scalable Unit	SRS-48UDLC2-4U8N-R0
GPUs	8x NVIDIA HGX B300 8-GPU (64 GPUs)
CPUs	16x Intel® Xeon® or AMD EPYC™ processors
GPU Systems	8x SYS-422GS-NB3RT-ALC / SYS-422GS-NB3RT-LCC / AS -4126GS-NB3RT-LCC
Networking*	NVIDIA Quantum-X800 InfiniBand 800G XDR or NVIDIA Spectrum-X Ethernet 800Gb/s Ethernet ToR management switches
Rack Dimension*	48U x 800mm x 1470mm
Liquid Cooling Options	1 in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and dual hot-swap pumps Optional: 1.8MW capacity in-row CDU

*Recommended configuration. Other network switch options and rack dimensions and layouts are available.
Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

4U 8-GPU System	SYS-422GS-NB3RT-ALC / SYS-422GS-NB3RT-LCC / AS-4126GS-NB3RT-LCC
Overview	4U liquid-cooled system with front I/O NICs, DPUs, storage, and management
CPU	Dual Intel® Xeon® 6700 series processors with P-cores (SYS-422GS-NB3RT) Dual AMD EPYC™ 9005/9004 Series Processors (AS -4126GS-NB3RT)
Memory	32 DIMMs, up to 8TB DDR5-5200 or up to 4TB DDR5-6400 (SYS-422GS-NB3RT) 24 DIMMs, up to 6TB DDR5-6400 (AS -4126GS-NB3RT)
GPU	NVIDIA HGX B300 8-GPU (288GB HBM3e per GPU*) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking	8 integrated NVIDIA ConnectX®-8 SuperNICs, up to 800Gb/s 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 hot-swap E1.S NVMe drive bays 2 M.2 NVMe slots
Power Supply	4 redundant (2+2) 6600W Titanium Level power supplies

*Physical GPU memory

Front I/O 8U Air-Cooled HGX B300 System



Networking

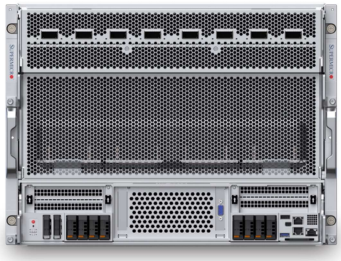
- NVIDIA Quantum-X800 InfiniBand or NVIDIA Spectrum-X Ethernet for up to 800Gb/s compute fabric
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network

Compute

- 4x SYS-822GS-NB3RT or AS -8126GS-NB3RT
- 32x NVIDIA Blackwell Ultra B300 GPUs per rack
- 9.2TB HBM3e per rack
- Dedicated storage fabric options with full NVIDIA GPUDirect RDMA and Storage or RoCE support

New Air-Cooled System Design with Ultra Performance

The Supermicro NVIDIA HGX platform is the building block of many of the world's largest AI clusters, delivering the immense computational output required for powering today's transformative AI applications. Now featuring NVIDIA Blackwell Ultra, the 8U air-cooled system is designed to maximize performance of eight 1100W TDP NVIDIA HGX B300 GPUs with up to 2.3TB of total HBM3e memory*. The front eight OSFP ports supporting integrated NVIDIA ConnectX®-8 SuperNICs at 800 Gb/s enable plug-and-play deployment with NVIDIA Quantum-X800 InfiniBand or NVIDIA Spectrum-X Ethernet compute fabric.



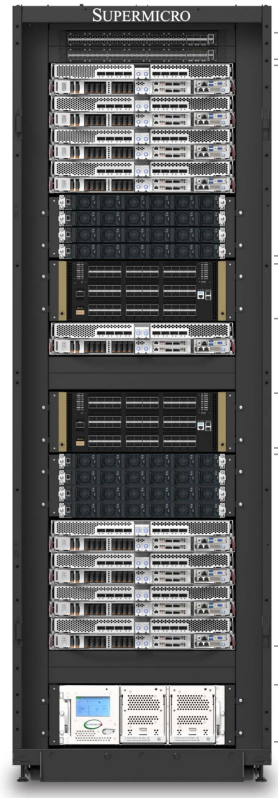
32-GPU Scalable Unit	SRS-48UAC-8U4N-R0
GPUs	4x NVIDIA HGX B300 8-GPU (32 GPUs)
CPUs	8x Intel® Xeon® or AMD EPYC™ processors
GPU Systems	4x SYS-822GS-NB3RT / AS -8126GS-NB3RT
Networking*	NVIDIA Quantum-X800 InfiniBand 800G XDR or NVIDIA Spectrum-X Ethernet 800Gb/s Ethernet ToR management switches
Rack Dimension*	48U x 750mm x 1400mm

*Recommended configuration. Other network switch options and rack dimensions and layouts are available. Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

8U 8-GPU System	SYS-822GS-NB3RT / AS -8126GS-NB3RT
Overview	8U air-cooled system with front I/O NICs, DPUs, storage, and management
CPU	Dual Intel® Xeon® 6700 series processors with P-cores (SYS-822GS-NB3RT) Dual AMD EPYC™ 9005/9004 Series Processors (AS -8126GS-NB3RT)
Memory	32 DIMMs, up to 8TB DDR5-5200 or up to 4TB DDR5-6400 (SYS-822GS-NB3RT) 24 DIMMs, up to 6TB DDR5-6400 (AS -8126GS-NB3RT)
GPU	NVIDIA HGX B300 8-GPU (288GB HBM3e per GPU*) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking	8 integrated NVIDIA ConnectX®-8 SuperNICs, up to 800Gb/s 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 front hot-swap E1.S NVMe drive bays 2 M.2 NVMe slots
Power Supply	6 redundant (3+3) 6600W Titanium Level power supplies

*Physical GPU memory

2-OU (OCP) Liquid-Cooled HGX B300 System



Management Networking
- In-band and out-of-band management switches

Compute Nodes and Power Shelves
- 5x SYS-222GS-NB3OT-ALC
- 4x power shelves with OCP ORV3 44-OU rack busbar
- 72x NVIDIA Blackwell Ultra B300 GPUs per rack (total of 9 compute nodes)

Compute Fabric Networking
- 2x NVIDIA Quantum-X800 InfiniBand switches for up to 800Gb/s compute fabric (leaf and spine switch)
- Dedicated storage fabric options with full NVIDIA GPUDirect RDMA and Storage or RoCE support

Compute Nodes and Power Shelves
- 4x SYS-222GS-NB3OT-ALC
- 4x power shelves with OCP ORV3 44-OU rack busbar

Liquid-Cooling
- Supermicro 250kW capacity Coolant Distribution Unit (CDU) with redundant PSU and dual hot-swap pumps

Hyperscale Performance in Ultra-Compact Design

Built to the 21-inch OCP Open Rack V3 (ORV3) specification, Supermicro's 2-OU liquid-cooled NVIDIA HGX B300 system sets a new standard for GPU density and efficiency. Each compact node features eight NVIDIA Blackwell B300 GPUs operating at up to 1,100W TDP, cooled by state-of-the-art liquid cooling with blind-mate manifold connections and modular GPU/CPU tray architecture. This innovative design delivers exceptional serviceability while dramatically reducing power consumption and rack footprint—enabling hyperscale and cloud providers to maximize compute performance in space-constrained data centers.

New

72-GPU Scalable Unit	SRS-440UDLC-20U9N-L1
GPUs	9x NVIDIA HGX B300 8-GPU (72 GPUs)
CPUs	18x Intel® Xeon® processors
GPU Systems	9x SYS-222GS-NB3OT-ALC
Networking*	NVIDIA Quantum-X800 InfiniBand 800G XDR NVIDIA Spectrum Ethernet management switches
Power Shelves	8x 1U 33kW (6x 5.5kW PSUs) power shelves with built-in capacitor, total 132kW (n+n redundancy)
Rack Dimension*	44-OU (OCP) x 750mm x 1200mm
Liquid Cooling Options	1x in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and n+1 hot-swap pumps Optional: 1.8MW capacity in-row CDU

*Recommended configuration. Other network switch options and rack dimensions and layouts are available. Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

2-OU 8-GPU System	SYS-222GS-NB3OT-ALC
Overview	2-OU liquid-cooled system with front I/O NICs, DPUs, storage, and management
CPU	Dual Intel® Xeon® 6700 series processors with P-cores
Memory	32 DIMMs, up to 8TB DDR5-5200 or up to 4TB DDR5-6400
GPU	NVIDIA HGX B300 8-GPU (288GB HBM3e per GPU*) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking	8 integrated NVIDIA ConnectX®-8 SuperNICs, up to 800Gb/s 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 front hot-swap E1.S NVMe drive bays 2 M.2 NVMe slots
Power Supply	Shared power through 4+4 rack power shelves with ORV3 busbar design

*Physical GPU memory

2-OU (OCP) Liquid-Cooled HGX B300 System

Management Networking

- In-band and out-of-band management switches

Compute Nodes and Power Shelves

- 18x SYS-222GS-NB3OT-ALC
- 10x power shelves with OCP ORV3 48-OU rack busbar
- 144x NVIDIA Blackwell Ultra B300 GPUs per rack

Compute Fabric Networking

- Centralized networkng racks with NVIDIA Quantum-X800 InfiniBand switches for up to 800Gb/s compute fabric (leaf and spine switches)
- Dedicated storage fabric options with full NVIDIA GPUDirect RDMA and Storage or RoCE support

Liquid-Cooling

- Supermicro 1.8MW capacity Coolant Distribution Unit (CDU) with redundant PSU and n+1 hot-swap pumps

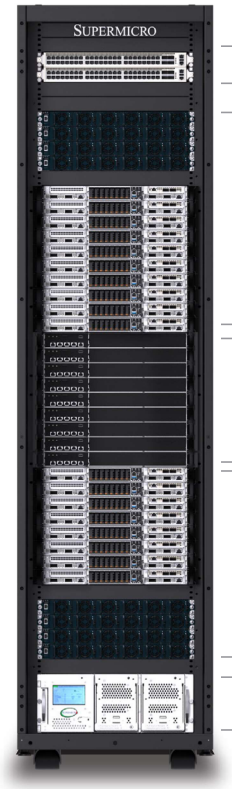
144 GPUs per Rack: Industry-Leading Density at Scale

Scale seamlessly from node to rack to cluster with Supermicro's complete NVIDIA HGX B300 architecture. A single ORV3 rack supports up to 18 nodes with 144 GPUs total, integrated with NVIDIA Quantum-X800 InfiniBand switches and Supermicro's 1.8MW in-row coolant distribution units (CDUs). At full scale, eight HGX B300 compute racks, three NVIDIA Quantum-X800 networking racks, and two Supermicro CDUs form a SuperCluster with 1,152 GPUs—delivering unmatched performance density for the most demanding AI workloads in hyperscale data centers and AI factories.

144-GPU Scalable Unit	SRS-480UDLC-20U18N-L1
GPUs	18x NVIDIA HGX B300 8-GPU (144 GPUs)
CPUs	36x Intel® Xeon® processors
GPU Systems	18x SYS-222GS-NB3OT-ALC
Networking*	NVIDIA Quantum-X800 InfiniBand 800G XDR NVIDIA Spectrum Ehternet management switches
Power Shelves	10x 1U 33kW (6x 5.5kW PSUs) power shelves with built-in capacitor, total 165kW (n+2 redundancy)
Rack Dimension*	48-OU (OCP) x 750mm x 1200mm
Liquid Cooling	Supermicro 1.8MW capacity in-row CDU with redundant PSU and n+1 hot-swap pumps (2x in-row CDUs required for 1152-GPU SuperCluster with 8x compute racks and 3x networking racks)

*Recommended configuration. Other network switch options and rack dimensions and layouts are available.
Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

NVIDIA GB300 NVL72



Management Networking
- In-band and out-of-band management switches

Compute Trays and Power Shelves
- 10x compute tray with 4x NVIDIA Blackwell Ultra B300 GPUs and 2x NVIDIA Grace CPUs per tray
- 4x power shelves

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

Compute Trays and Power Shelves
- 8x compute tray with 4x NVIDIA Blackwell Ultra B300 GPUs and 2x NVIDIA Grace CPUs per tray
- 4x power shelves

Liquid-Cooling Options
- Supermicro 250kW capacity coolant distribution unit (CDU) with redundant PSU and hot-swap pumps
- Optional 200kW capacity liquid-to-air sidecar CDU (no facility water required)

An Exascale Compute in a Rack

The Supermicro NVIDIA GB300 NVL72 tackles AI computational demands from training foundational models to large-scale reasoning model inference. It combines high AI performance with Supermicro's direct liquid cooling technology, enabling maximum computing density and efficiency. Based on NVIDIA Blackwell Ultra, a single rack integrates 72 NVIDIA B300 GPUs with 288GB HBM3e memory each. With 1.8TB/s NVLink interconnects, the GB300 NVL72 operates as an exascale supercomputer in a single node. Upgraded networking doubles performance across compute fabric, supporting 800 Gb/s speeds. Supermicro's manufacturing capacity and end-to-end services accelerate liquid-cooled AI factory deployment and speed time-to-market for GB300 NVL72 clusters.

NVIDIA GB300 NVL72	SRS-GB300-NVL72-M1 / SRS-GB300-NVL72-M0
GPUs	72x NVIDIA Blackwell Ultra B300 GPUs
CPUs	36x 72-core NVIDIA Grace Arm Neoverse V2 CPUs
Total GPU Memory	18x 1.15TB HBM3e
Total System Memory	18x 960GB LPDDR5X
NVLink Switch Trays	9x NVLink Switches, 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) power shelves with built-in capacitor, total 132kW (n+n redundancy)
Operating Power	132kW to 140kW
Rack Dimension	2236mm x 600 mm x 1068mm
Liquid Cooling Options	1x in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and dual hot-swap pumps (SRS-GB300-NVL72-M1) Optional:1.8MW capacity in-row CDU with redundant pumps Optional: 200kW capacity liquid-to-air sidecar CDU for facilities without cooling tower and water supply

Compute Tray	
Overview	1U Liquid-cooled System with 2x NVIDIA GB300 Grace Blackwell Superchips
CPU and GPU	2 72-core NVIDIA Grace Arm Neoverse V2 CPUs 4 NVIDIA Blackwell Tensor Core GPUs
GPU Memory	1.15 HBM3e per Compute Tray
CPU Memory	960GB LPDDR5X per Compute Tray
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking	4 NVIDIA NVLink Switch ports (rear) 4 single-port NVIDIA ConnectX®-8 SuperNICs (front), up to 800Gb/s 1 dual-port NVIDIA BlueField®-3 DPUs (front)
Storage	Up to 8 E1.S PCIe 5.0 drives
Power Supply	Shared power through 4+4 rack power shelves

Total Liquid-Cooling Offerings for a Wide Range of AI Data Center Environments

The NVIDIA GB300 NVL72 and GB200 NVL72 deliver exascale computing capabilities in a single rack with fully integrated liquid cooling. Supermicro's Data Center Building Block Solutions® (DCBBS) accelerate time-to-online by offering a total solution, along with on-site deployment services and support. From individual GPUs to full racks and facility-side infrastructure, Supermicro enables end-to-end deployment with ultimate flexibility. DCBBS provides a comprehensive data center-level solution stack for liquid cooling with multiple options suitable for a wide range of data center environments.

In-Rack CDU
250 kW in-rack CDU with easy controls through touchscreen and web interface. Ensures uptime with n+1 redundant pumps.

In-Row CDU
1.8 MW in-row CDU enables a single CDU to cool multiple racks of systems. Ensures uptime with dual redundant pumps.

L2A Sidecar CDU
200 kW Liquid-to-air CDU dissipates heat from liquid to air, allowing for simple deployment when full liquid cooling is not possible.

Cooling Tower
Efficiently dissipates heat to the outside environment with a highly efficient design that can adapt to any deployment size

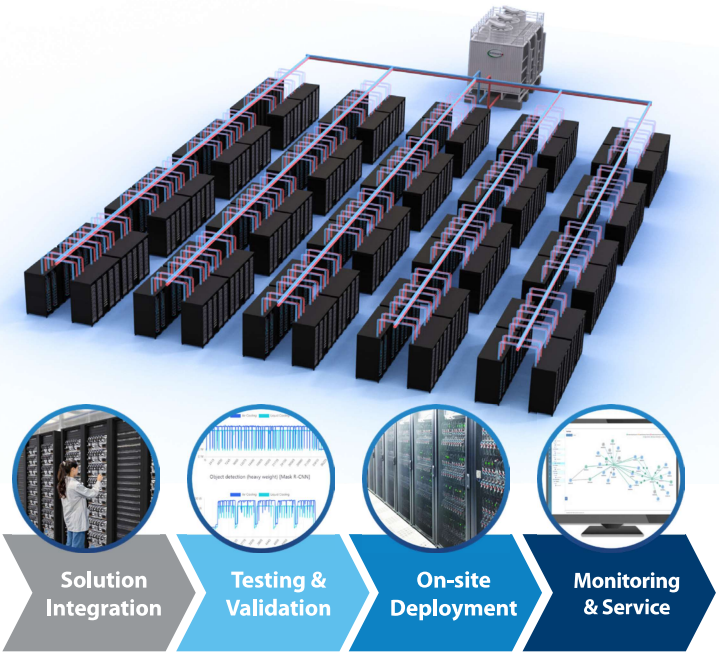
Dry Cooler
Delivers air-based heat rejection to adapt to any deployment size with a flexible, modular design

Rear Door Heat Exchanger
Mounts to rack, providing passive and active heat rejection to ensure full thermal coverage

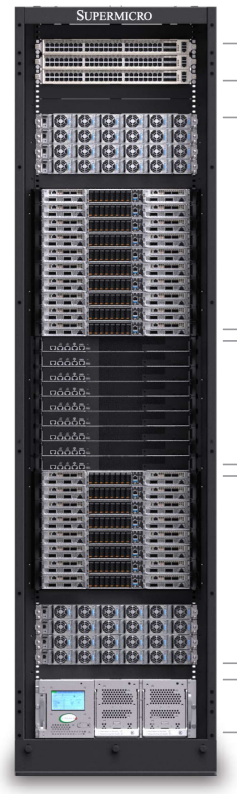
End-to-end Data Center Solution and Deployment Services for NVIDIA Blackwell

Supermicro serves as a comprehensive one-stop solution provider with global manufacturing scale, delivering data center-level solution design, liquid-cooling technologies, switching, cabling, a full data center management software suite, L11 and L12 solution validation, on-site installation, and professional support and service. With production facilities across San Jose, Europe, and Asia, Supermicro offers unmatched manufacturing capacity for liquid-cooled or air-cooled rack systems, ensuring timely delivery, reduced total cost of ownership (TCO), and consistent quality.

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.



NVIDIA GB200 NVL72



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

Compute Trays and Power Shelves
- 10x compute tray with 4x NVIDIA Blackwell B200 GPUs and 2x NVIDIA Grace CPUs per tray
- 4x power shelves

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

Compute Trays and Power Shelves
- 8x compute tray with 4x NVIDIA Blackwell B200 GPUs and 2x NVIDIA Grace CPUs per tray
- 4x power shelves

Liquid-Cooling Options
- Supermicro 250kW capacity coolant distribution unit (CDU) with redundant PSU and dual hot-swap pumps
- Optional 200kW capacity liquid-to-air sidecar CDU (no facility water required)

72-GPU Unified Architecture in a Rack

The Supermicro NVIDIA GB200 NVL72 delivers unprecedented performance for training and deploying large language models and generative AI applications. Built on NVIDIA Blackwell architecture, a single rack integrates 72 NVIDIA B200 GPUs with 192GB HBM3e memory each, connected via 1.8TB/s NVLink to function as a unified exascale AI supercomputer. Supermicro's advanced direct liquid cooling technology maintains optimal thermal performance while maximizing computing density. With complete rack-level integration and global manufacturing capacity, Supermicro ensures rapid deployment of production-ready GB200 NVL72 infrastructure for AI factories.

NVIDIA GB200 NVL72	SRS-GB200-NVL72-M1 / SRS-GB200-NVL72-M0
GPUs	72x NVIDIA Blackwell B200 GPUs
CPUs	36x 72-core NVIDIA Grace Arm Neoverse V2 CPUs
Total GPU Memory	18x 744GB HBM3e
Total System Memory	18x 960GB LPDDR5X
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) power shelves, total 132kW (n+n redundancy)
Operating Power	125kW to 135kW
Rack Dimension	2236mm x 600 mm x 1068mm
Liquid Cooling Options	1 in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and dual hot-swap pumps (SRS-GB200-NVL72-M1) Optional:1.8MW capacity in-row CDU with redundant pumps Optional: 200kW capacity liquid-to-air sidecar CDU for facilities without cooling tower and water supply

Compute Tray	
Overview	1U Liquid-cooled System with 2x NVIDIA GB200 Grace Blackwell Superchips
CPU and GPU	2 72-core NVIDIA Grace Arm Neoverse V2 CPUs 4 NVIDIA Blackwell Tensor Core GPUs
GPU Memory	744GB HBM3e per Compute Tray
CPU Memory	960GB LPDDR5X per Compute Tray
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking	4 NVIDIA NVLink Switch ports (rear) 4 single-port NVIDIA ConnectX®-7 SuperNICs or 2 dual-port ConnectX®-8 (front) 1 dual-port NVIDIA BlueField®-3 DPUs (front)
Storage	Up to 8 E1.S PCIe 5.0 drives
Power Supply	Shared power through 4+4 rack power shelves

NVIDIA GB200 NVL4

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

Compute Networking
- NVIDIA Quantum-X800 InfiniBand for up to 800Gb/s compute fabric
- Centralized networking switch rack

Compute Nodes and Power Shelves
- Up to 26x nodes per rack with in-rack coolant distribution unit (CDU) and 8x power shelves
- Up to 32x nodes per rack with in-row CDU
- 4x NVIDIA Blackwell B200 GPUs and 2x NVIDIA Grace CPUs per node
- 8x power shelves with busbar (10x power shelves required for 32x nodes)

Liquid-Cooling Options
- Supermicro 250kW capacity in-rack CDU for up to 104 GPUs per rack
- 1.8MW capacity in-row CDU for up to 128 GPUs per rack

Designed for Accelerated HPC

Supermicro's NVIDIA GB200 NVL4 rack-scale platform delivers exceptional performance for GPU-accelerated HPC and AI science workloads including molecular simulation, weather modeling, fluid dynamics, and genomics. Each node unifies four NVLink-connected NVIDIA Blackwell B200 GPUs with two NVIDIA Grace CPUs over NVLink-C2C, featuring direct-to-chip liquid cooling and four ports of 800G NVIDIA Quantum InfiniBand networking with 800G dedicated to each GPU. Scale up to 32 nodes per rack for 128 GPUs in a 48U NVIDIA MGX rack with busbar power distribution and flexible in-rack or in-row CDU options—delivering unmatched rack density for scientific computing and advanced AI research applications.

104-GPU Scalable Unit	SRS-48LB26-NVL4-M1
GPUs	104x NVIDIA Blackwell B200 GPUs
CPUs	52x 72-core NVIDIA Grace Arm Neoverse V2 CPUs
GPU Systems	26x SYS-121GL-NB2B-LCC
Networking*	NVIDIA Quantum-X800 InfiniBand 800G XDR or NVIDIA Spectrum-X Ethernet 800Gb/s Ethernet ToR management switches
Power Shelves	8x 1U 33kW (6x 5.5kW PSUs) power shelves, total 159kW (6+2 redundancy)
Rack Dimension*	2236mm x 600mm x 1068mm
Liquid Cooling Options	1 in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and dual hot-swap pumps Optional: 1.8MW capacity in-row CDU

*Recommended configuration. Other network switch options and rack dimensions and layouts are available. Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

NVIDIA GB200 NVL4	ARS-121GL-NB2B-LCC
Overview	1U Liquid-cooled System with 2x NVIDIA GB200 Grace Blackwell Superchips
CPU and GPU	2 72-core NVIDIA Grace Arm Neoverse V2 CPUs 4 NVIDIA Blackwell Tensor Core GPUs
GPU Memory	744GB HBM3e
CPU Memory	960GB LPDDR5X
NVLink	5th Generation NVIDIA NVLink at 600GB/s
Networking*	2 dual-port NVIDIA ConnectX®-8 SuperNICs (front) 1 dual-port NVIDIA BlueField®-3 DPUs (front)
Storage	Up to 8 E1.S PCIe 5.0 drives
Power Supply	Shared power through 6+2 rack power shelves

*Optional NIC customization available.

Front I/O 4U Liquid-Cooled HGX B200 System

Networking

- NVIDIA Quantum-2 InfiniBand or NVIDIA Spectrum-X Ethernet for up to 400Gb/s compute fabric
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network

Compute

- 8x SYS-422GS-NBRT-LCC per rack
- 64x NVIDIA Blackwell B200 GPUs per rack
- 11.5TB of HBM3e per rack
- Flexible storage options with dedicated storage fabric supporting full NVIDIA GPUDirect RDMA and Storage or RoCE

Liquid-Cooling with DLC-2

- Up to 98% system heat capture by liquid-cooling CPUs, GPUs, DIMMs, PCIe switches, VRMs, power supplies, and more
- Supermicro 250kW capacity Coolant Distribution Unit (CDU) with redundant PSU and dual hot-swap pumps
- Supermicro vertical Coolant Distribution Manifolds (CDM)

Front I/O Liquid-Cooled System with Supermicro DLC-2 Technology

The new front I/O liquid-cooled 4U System designed for NVIDIA HGX B200 8-GPU system features Supermicro's DLC-2 technology. The direct liquid cooling now captures up to 98% of the heat generated by server components, such as CPU, GPU, PCIe switch, DIMM, VRM, and PSU, allowing up to 40% data center power savings and as low as a 50dB noise level. Supermicro's Front I/O NVIDIA HGX B200 systems simplify the deployment, management, and maintenance of liquid-cooled AI infrastructure, allowing easy front I/O access, simplifying cabling, improving thermal efficiency and compute density, and reducing operational expenses (OPEX).

64-GPU Scalable Unit	SRS-48UDLC-4U8N-R0
GPUs	8x NVIDIA HGX B200 8-GPU (64 GPUs)
CPUs	16x Intel® Xeon® processors
GPU Systems	8x SYS-422GS-NBRT-LCC
Networking*	NVIDIA Quantum-2 InfiniBand 400G NDR or NVIDIA Spectrum-X Ethernet 400Gb/s Ethernet ToR management switches
Rack Dimension*	48U x 800mm x 1470mm
Liquid Cooling Options	1 in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and dual hot-swap pumps Optional: 1.8MW capacity in-row CDU

*Recommended configuration. Other network switch options and rack dimensions and layouts are available. Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

Front I/O 4U 8-GPU System	SYS-422GS-NBRT-LCC
Overview	4U DLC-2 Liquid-cooled System with front I/O NICs, DPUs, Storage, and Management
CPU	Dual Intel® Xeon® 6700 series processors with P-cores
Memory	32 DIMMs, up to 8TB DDR5-5200 or up to 4TB DDR5-6400 RDIMM
GPU	NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking*	8 single-port NVIDIA ConnectX®-7 NICs or NVIDIA BlueField®-3 SuperNICs Up to 400Gb/s 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 Hot-swap E1.S NVMe storage drive bays and 2 M.2 NVMe slots
Power Supply	4 redundant (2+2) 6600W Titanium Level power supplies

*Recommended configuration.

Front I/O 8U Air-Cooled HGX B200 System

Networking

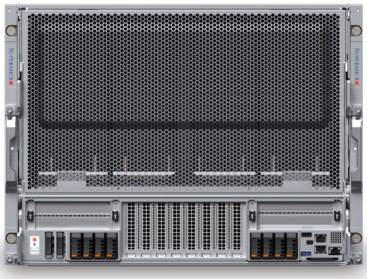
- NVIDIA Quantum-2 InfiniBand or NVIDIA Spectrum-X Ethernet for up to 400Gb/s compute fabric
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network

Compute

- 4x SYS-822GS-NBRT per rack
- 32x NVIDIA Blackwell B200 GPUs
- 5.76TB HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage or RoCE support

Front I/O, Enhanced Air-Cooled System for AI Factories

Similar to the front I/O liquid-cooled system, the newly introduced front I/O air-cooled system features front-accessible (or cold aisle accessible) NICs, DPUs, storage, and management components to streamline AI factory deployment without liquid-cooling infrastructure. It features a compact 8U form factor while providing system memory expansion with 32 DIMM slots to deliver greater flexibility. The larger system memory complements the NVIDIA HGX B200's HBM3e GPU memory by reducing CPU-GPU bottlenecks, and enhancing multi-job efficiency in virtualized environments, and accelerating data processing.



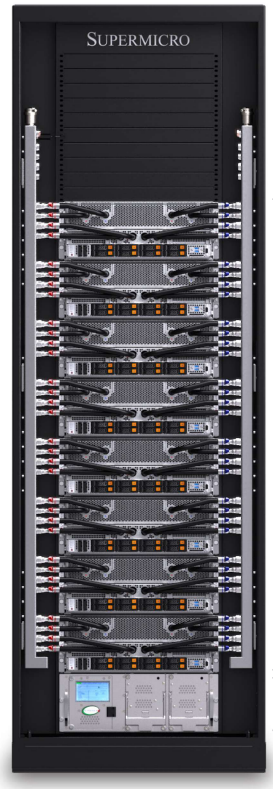
32-GPU Scalable Unit	SRS-48UAC-8U4N-R0
GPUs	4x NVIDIA HGX B200 8-GPU (32 GPUs)
CPUs	8x Intel® Xeon® processors
GPU Systems	4x SYS-822GS-NBRT
Networking*	NVIDIA Quantum-2 InfiniBand 400G NDR NVIDIA Spectrum-X Ethernet 400Gb/s Ethernet ToR management switch
Rack Dimension*	48U x 750mm x 1400mm

*Recommended configuration. Other network switch options and rack dimensions and layouts are available. Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

Front I/O 8U 8-GPU System	SYS-822GS-NBRT
Overview	8U Air-cooled System with Front I/O NICs, DPUs, Storage, and Management
CPU	Dual Intel® Xeon® 6700 series processors with P-cores
Memory	32 DIMMs, up to 8TB DDR5-5200 or up to 4TB DDR5-6400 RDIMM
GPU	NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking*	8 single-port NVIDIA ConnectX®-7 NICs or NVIDIA BlueField®-3 SuperNICs Up to 400Gb/s 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 Hot-swap E1.S NVMe storage drive bays and 2 M.2 NVMe slots
Power Supply	6 redundant (3+3) 6600W Titanium Level power supplies

*Recommended configuration.

Rear I/O 4U Liquid-Cooled HGX B200 System



Networking

- NVIDIA Quantum-2 InfiniBand or NVIDIA Spectrum-X Ethernet for up to 400Gb/s compute fabric
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network

Compute

- 8x SYS-422GA-NBRT-LCC, AS -4126GS-NBR-LCC, or SYS-421GE-NBRT-LCC per rack
- 64x NVIDIA Blackwell B200 GPUs
- 11.5TB of HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage or RoCE support

Liquid-Cooling

- Supermicro 250kW capacity Coolant Distribution Unit (CDU) with redundant PSU and dual hot-swap pumps
- Supermicro vertical Coolant Distribution Manifolds (CDM)

Proven Density and Efficiency

The liquid-cooled 4U NVIDIA HGX B200 8-GPU system features cold plates for CPUs, GPUs, and DIMMs (optional), and advanced tubing design paired with the new 250kW coolant distribution unit (CDU) in the 4U form factor. The liquid-cooling architecture optimized specifically for NVIDIA HGX B200 8-GPU enhances efficiency and serviceability of the predecessor that are designed for NVIDIA HGX H100/ H200 8-GPU. Available in 42U, 48U or 52U configurations, the rack scale design with the vertical coolant distribution manifolds (CDM) means that horizontal manifolds no longer occupy valuable rack units. This enables 8 systems, 64 NVIDIA Blackwell GPUs in a 42U rack and all the way up to 12 systems with 96 NVIDIA GPUs in a 52U rack.

64-GPU Scalable Unit	SRS-48UDLC-4U8N-L1
GPUs	8x NVIDIA HGX B200 8-GPU (64 GPUs)
CPUs	16x Intel® Xeon® or AMD EPYC™ processors
GPU Systems	8x SYS-422GA-NBRT-LCC / AS -4126GS-NBR-LCC / SYS-421GE-NBRT-LCC
Networking*	NVIDIA Quantum-2 InfiniBand 400G NDR or NVIDIA Spectrum-X Ethernet 400Gb/s Ethernet ToR management switches
Rack Dimension*	48U x 800mm x 1470mm
Liquid Cooling Options	1 in-rack Supermicro 4U 250kW capacity CDU with redundant PSU and dual hot-swap pumps Optional: 1.8MW capacity in-row CDU

*Recommended configuration. Other network switch options and rack dimensions and layouts are available. Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

4U 8-GPU System	SYS-422GA-NBRT-LCC / AS-4126GS-NBR-LCC / SYS-421GE-NBRT-LCC
Overview	4U Liquid-cooled System with NVIDIA HGX B200 8-GPU
CPU	Dual Intel® Xeon® 6900 series processors with P-cores (SYS-422GA-NBRT-LCC) Dual AMD EPYC™ 9005/9004 Series Processors (AS -4126GS-NBR-LCC) Dual 5th/4th Gen Intel® Xeon® Scalable processors (SYS-421GE-NBRT-LCC)
Memory	24 DIMMs, up to DDR5-6400 RDIMM (SYS-422GA-NBRT-LCC) 24 DIMMs, up to DDR5-6000 RDIMM (AS -4126GS-NBR-LCC) 32 DIMMs, up to DDR5-5600 RDIMM (SYS-421GE-NBRT-LCC)
GPU	NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking*	8 single-port NVIDIA ConnectX®-7 NICs or NVIDIA BlueField®-3 SuperNICs Up to 400Gb/s 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 front hot-swap 2.5" NVMe drive bays 2 M.2 NVMe slots
Power Supply	4 redundant (2+2) 6600W Titanium Level power supplies

*Recommended configuration.

Rear I/O 10U Air-Cooled HGX B200 System

Networking

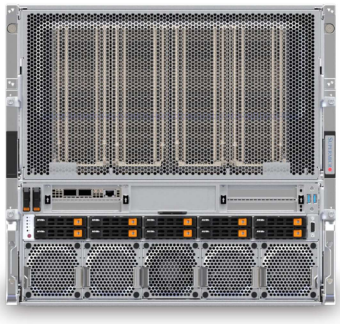
- NVIDIA Quantum-2 InfiniBand or NVIDIA Spectrum-X Ethernet for up to 400Gb/s compute fabric
- Ethernet leaf switches for in-band management
- Out-of-band 1G/10G IPMI switch
- Non-blocking network

Compute

- 4x SYS-A22GA-NBRT, AS -A126GS-TNBR, or SYS-A21GE-NBRT per rack
- 32x NVIDIA Blackwell B200 GPUs
- 5.76TB HBM3e per rack
- Flexible storage options with local or dedicated storage fabric with full NVIDIA GPUDirect RDMA and Storage or RoCE support

Advanced Air-Cooling Technology

The air-cooled 10U NVIDIA HGX B200 system features a redesigned chassis with expanded thermal headroom to accommodate eight 1000W TDP Blackwell GPUs. Up to 4 of the new 10U air-cooled systems can be installed and fully integrated in a rack, the same density as the previous generation, while providing up to 15x inference and 3x training performance. All Supermicro NVIDIA HGX B200 systems are equipped with a 1:1 GPU-to-NIC ratio supporting NVIDIA BlueField®-3 or NVIDIA ConnectX®-7 for scaling across a high-performance compute fabric.



32-GPU Scalable Unit	SRS-48UAC-10U4N-A1
GPUs	4x NVIDIA HGX B200 8-GPU (32 GPUs)
CPUs	8x Intel® Xeon® or AMD EPYC™ processors
GPU Systems	4x SYS-A22GA-NBRT / AS -A126GS-TNBR / SYS-A21GE-NBRT
Networking*	NVIDIA Quantum-2 InfiniBand 400G NDR NVIDIA Spectrum-X Ethernet 400Gb/s Ethernet ToR management switch
Rack Dimension*	48U x 750mm x 1400mm

*Recommended configuration. Other network switch options and rack dimensions and layouts are available. Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional.

10U 8-GPU System	SYS-A22GA-NBRT / AS-A126GS-TNBR / SYS-A21GE-NBRT
Overview	10U Air-cooled System with NVIDIA HGX B200 8-GPU
CPU	Dual Intel® Xeon® 6900 series processors with P-cores (SYS-A22GA-NBRT) Dual AMD EPYC™ 9005/9004 Series Processors (AS -A126GS-TNBR) Dual 5th/4th Gen Intel® Xeon® Scalable processors (SYS-A21GE-NBRT)
Memory	24 DIMMs, up to DDR5-6400 RDIMM (SYS-A22GA-NBRT) 24 DIMMs, up to DDR5-6000 RDIMM (AS -A126GS-TNBR) 32 DIMMs, up to DDR5-5600 RDIMM (SYS-A21GE-NBRT)
GPU	NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) 1.8TB/s NVLink GPU-GPU interconnect with NVSwitch
NVLink	5th Generation NVIDIA NVLink at 1.8TB/s
Networking*	8 single-port NVIDIA ConnectX®-7 NICs or NVIDIA BlueField®-3 SuperNICs Up to 400Gb/s 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	10 front hot-swap 2.5" NVMe drive bays 2 M.2 NVMe slots
Power Supply	6 redundant (3+3) 5250W redundant Titanium Level power supplies

*Recommended configuration.