# Supermicro NVIDIA

## Ultra-Performance for AI with 72 Liquid-Cooled NVIDIA B300 GPUs in a Rack

### The Highest Level of AI Performance and Scalability

- **Rack-scale solution** with NVIDIA GB300 Grace™ Blackwell Superchip providing 72 NVIDIA B300 GPUs and 36 Grace CPUs per rack
- **NVIDIA Blackwell Ultra Platform** with 288GB HBM3e per GPU*
- **Supermicro Direct Liquid-Cooling** total solution stack for fast deployment of liquid cooled data centers
- **Comprehensive Service** from consultation to full-scale deployment, providing all necessary parts, networking solutions, and onsite installation services
- **Up to 800Gb/s** NVIDIA Quantum-X800 InfiniBand or Spectrum-X Ethernet with integrated NVIDIA ConnectX®-8 SuperNICs

## An Exascale AI Supercomputer in a Rack, Doubled in Performance

The Supermicro NVIDIA GB300 NVL72 is built to tackle the rapidly-scaling computational demands of AI, ranging from training massive foundational models to employing large scale reasoning model inference that require greater test-time computing resources. It combines the highest level of AI performance and scalability together with Supermicro's leading direct liquid cooling technology, allowing enterprises to push the limits of computing density and efficiency. Based on the new NVIDIA Blackwell Ultra platform, a single rack integrates 72 NVIDIA B300 GPUs, each with the highest-available 288GB HBM3e memory capacity*. With each GPU interconnected via 1.8TB/s NVLink, the GB300 NVL72 effectively serves as an exascale supercomputer, operating as a single node. The upgraded networking doubles performance when scaling across the compute fabric, supporting speeds up to 800 Gb/s. Supermicro's leading manufacturing capacity and end-to-end services accelerate liquid-cooled AI factory buildouts and speeds up time-to-market in deploying GB300 NVL72 clusters of virtually any size.

## Max Speeds, from the Chip to Cluster-Level

The GB300 NVL72 maximizes overall AI factory performance by addressing key bottlenecks in AI computing. At the chip-level, the densely co-packaged HBM3e increases memory capacity to 288GB per GPU. Each of the 72 GPUs per rack is connected with 1.8TB/s NVLink to form a massive pool of 21TB HBM3e, an unprecedented capacity to store massive AI models within the fastest regions of the memory hierarchy. Cluster-level speeds are doubled via the GB300 NVL72's integrated NVIDIA ConnectX®-8 NICs, with up to 800Gb/s throughput across the compute fabric. At the data center-level, Supermicro's leading direct liquid cooling ensures thermal stability for the most demanding AI workloads. In particular, the GB300 NVL72 offers dramatic speedups for AI training applications with elevated memory requirements.

## Ready for Plug-and-Play AI Factories

Supermicro's Data Center Building Block Solutions® accelerate time-to-market and time-to-online by offering a total solution with all critical computing and cooling infrastructure, along with on-site services and support. From individual GPUs to full racks and facility-side infrastructure, Supermicro enables end-to-end deployment with ultimate flexibility.

### Supermicro Data Center Building Block Solutions®:

- **Speed up Time-to-Market and Time-to-Online**
- **One-Stop Shop with Optimized Quality and Performance**
- **Save Power and Cost with Reduced OPEX and CAPEX**

*Physical GPU memory

## Rack Scale Design Close-up


### 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

### 10 Compute Trays

- 4x NVIDIA B300 GPUs per tray
- 2x NVIDIA Grace CPUs per tray

### Compute Interconnect

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

### 8 Compute Trays

- 4x NVIDIA B300 GPUs per tray
- 2x NVIDIA Grace CPUs per tray

### Liquid-Cooling Options

- In-Rack CDU: Up to 250kW capacity CDU with N+1 redundant pumps
- In-Row CDU: Up to 1.8MW capacity CDU with N+1 redundant pumps (supports up to 8 racks)
- L2A Sidecar CDU: Up to 200kW capacity CDU with N+1 redundant pumps (no facility water required)

## Direct Liquid Cooling and Supermicro Data Center Building Block Solutions® (DCBBS)

Supermicro's DCBBS covers all critical data center components, including air and liquid cooling 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**
200kW sidecar 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

| NVIDIA GB300 NVL72 | SRS-GB300-NVL72-M1 / SRS-GB300-NVL72-M0 |
|---|---|
| GPUs | 72x NVIDIA B300 Tensor Core GPUs |
| CPUs | 36x NVIDIA 72-core Grace ARM Neoverse V2 CPUs |
| Compute Trays | 18x 1U ARS-121GL-NB3-LC |
| 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) with built-in capacitor, total 132kW |
| Operating Power | 132kW to 140kW |
| Rack Dimensions (mm) | 2236mm x 600 mm x 1068mm |
| Liquid Cooling Options | In-Rack CDU: Up to 250kW capacity with N+1 redundant pumps<br>In-Row CDU: Up to 1.8MW capacity with N+1 redundant pumps (supports up to 8 racks)<br>L2A Sidecar CDU: Up to 200kW capacity CDU with N+1 redundant pumps (no facility water required) |

| Compute Tray | |
|---|---|
| Overview | 1U liquid-cooled with 2 NVIDIA GB300 Grace Blackwell Superchips |
| CPU and GPU | 2 72-core NVIDIA Grace ARM Neoverse V2 CPUs<br>4 NVIDIA B300 Tensor Core GPUs |
| GPU Memory | 1.15TB HBM3e per Compute Tray |
| CPU Memory | 960GB LPDDR5X per Compute Tray |
| Networking | 4 NVIDIA NVLink Switch ports<br>4 integrated NVIDIA ConnectX®-8 SuperNICs, up to 800Gb/s<br>Up to 2 NVIDIA BlueField®-3 DPUs |
| Storage | Up to 8 E1.S PCIe 5.0 drives |
| Power Supply | Shared power through 4+4 rack power shelves |