Enterprise 3D + AI SuperCluster

For NVIDIA Omniverse™ with 256 NVIDIA L40S GPUs

Scalable Compute Unit Built For 3D & AI Workflows

- Maximize multi-workload performance for enterprise AI-enabled workflows. Optimized for NVIDIA Omniverse™ with OpenUSD.
- 256 NVIDIA L40S GPUs in one scalable unit
- 12TB of GPU memory and 32TB of system memory in one scalable unit
- Scale-out with 400Gb/s NVIDIA Spectrum™-X Ethernet
- Customizable data storage fabric with industry leading parallel file system options
- Certified for NVIDIA Omniverse™ Enterprise with included Enterprise Support Services

Enterprise-Grade 3D + AI Performance for NVIDIA Omniverse™

A wide range of professionals depend on compute-intensive 3D workflows, with use-cases ranging from animated film production to industrial digital twins. AI has augmented existing 3D workflows with a full Generative AI toolkit. Supermicro's SuperCluster for NVIDIA Omniverse™ is a rack solution designed to simplify the deployment of scale-out infrastructure for the multi-workload needs of 3D and AI. Supermicro's SuperCluster for NVIDIA Omniverse™ delivers enterprise-grade performance for demanding 3D workloads in addition to providing the TFLOPs required for integrating cutting-edge AI inference models for real-time Generative AI. With up to 256 NVIDIA L40S GPUs per scalable unit, Supermicro's SuperCluster can serve the GPU-computing needs for the entire organization.

4U PCIe GPU System Nodes

Supermicro 4U PCIe GPU Systems compose the compute nodes of SuperCluster for NVIDIA Omniverse™. Each Supermicro 4U PCIe GPU system is equipped with up to 8 NVIDIA L40S GPUs powered by 4x 2700W Titanium Level PSUs, all within a high-airflow chassis.

Supermicro 4U PCIe GPU System nodes maximize the performance of NVIDIA L40S GPUs across the diverse range of tools and applications within the NVIDIA Omniverse development platform, including the world-building OpenUSD ecosystem, and generative AI technologies through Omniverse Cloud APIs.

System nodes scale across a high-performance network fabric via up to 4x 400Gb/s BlueField®-3 SuperNICs or NVIDIA ConnectX®-7 NICs.

Plug-and-Play, Reduced Lead-time

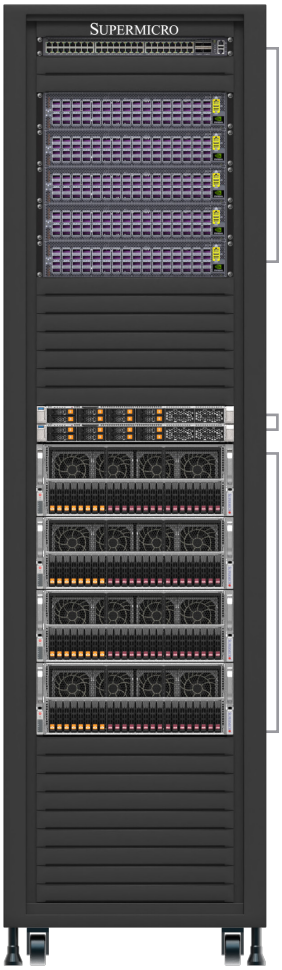
SuperCluster for NVIDIA Omniverse™ is an integrated software and hardware solution that simplifies complex scale-out 3D and AI infrastructure projects, ensuring interoperability and performance of all of its components, including systems, PDUs, networking, and more.

Supermicro's team of engineers design, build, and deploy the solution from end-to-end. SuperCluster for NVIDIA Omniverse™ can be deployed from a range of available sizes and options, tailored to the customer's requirements.

The result is a plug-and-play data center deployment experience, with Supermicro overseeing the delivery, cabling, configuration, testing, and support with a team of on-site engineers.

Rack Scale Design Close-up

(Layout and component quantities vary based on deployment size and configuration).



Networking

- Compute Fabric: 3x 400G NVIDIA SN5600 Switches (1 switch required for single-rack configuration)
- In-Band Management: 2x 400G NVIDIA SN5600 Switches (1 switch required for single-rack configuration)
- 1x NVIDIA SN2201 Switch (Out-of-band management)

Control

- 1x SYS-121H-TNR

Compute

- 4x or 8x 4U PCIe GPU Systems per rack
- Up to 64x NVIDIA L40S GPUs per rack (8 GPUs per system)
- Up to 45.6TB storage per rack
- Flexible storage options with local or dedicated storage fabric

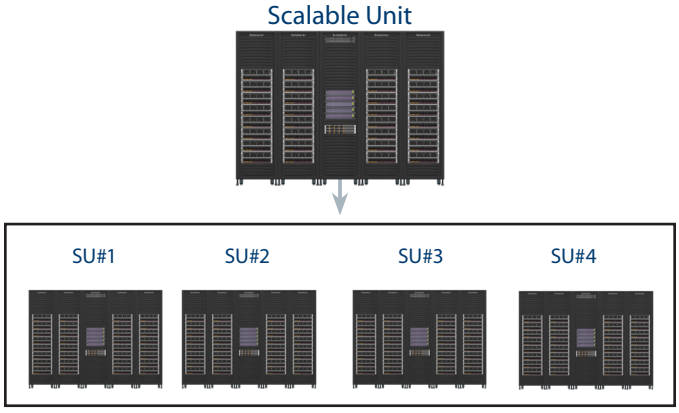
32-Node SuperCluster Scalable Unit

SuperCluster for NVIDIA Omniverse™ is a fully interconnected infrastructure solution that ensures designers, artists, and engineers, and more can access the highest level of GPU computing at the time-of-need by accessing virtual GPUs or full bare-metal system nodes.

The 400Gb/s high-performance network fabric, supporting NVIDIA Spectrum™-X Ethernet and Quantum-2 InfiniBand, unifies the cluster's high-performance compute fabric and supports the capability of leveraging GPU resources across nodes for a combined pool of GPU memory, essential for AI training applications.

Supermicro's validated rack solutions range in size from four GPU system nodes to a 256-GPU Scalable Unit, which can be further multiplied to fit enterprises of any size.

Scalable Unit



Node Configuration — SYS-421GE-TNRT/SYS-421GE-TNRT3

Overview	4U PCIe GPU System with NVIDIA L40S GPUs
CPU	Dual 5th/4th Gen Intel® Xeon® 8480+ or Intel® Xeon® 8462Y+ processors
Memory	1TB DDR5-4800 via 16 DIMMs (32 DIMM slots total)
GPU	Up to 8x NVIDIA L40S 48GB GDDR6 GPUs
Networking	Up to 4x NVIDIA BlueField®-3 B3140H SuperNIC or NVIDIA ConnectX®-7 400G NICs 1x NVIDIA BlueField®-3 DPU Dual Port 200G
Storage	3.8TB NVMe Storage drives (2x 3.8TB U.3) 1.9TB NVMe M.2, Boot drive
Power Supply	4x 2700W Redundant Titanium Level power supplies

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

32-Node Scalable Unit — SRS-48U0VX-8GXL-03 / SRS-48U0VX-4GXL-03

Overview	32-node cluster with up to 256 NVIDIA L40S GPUs
Compute System Nodes	32x SYS-421GE-TNRT (8 GPUs per node) or SYS-421GE-TNRT3 (4 GPUs per node)
Control System Nodes	3x SYS-121H-TNR
Compute Fabric Switches	3x 400G 64-port NVIDIA SN5600 Ethernet Switches
In-Band Management Switches	2x 400G 64-port NVIDIA SN5600 Ethernet Switches
Out-of-Band Management Switches	2x 1Gbps 48-port NVIDIA SN2201 Ethernet Switches
Rack / PDU	Rack: 5x 48U 750mmx1200mm. PDU: 18x 415V 60A 3Ph

*Recommended configuration, other network switch options and rack layouts are available, including configuration supporting NVIDIA Quantum-2 InfiniBand
*Login node may be required. NVIDIA Unified Fabric Manager (UFM) node optional