

Product Portfolio

The Supermicro Advantage



Agenda

About Supermicro

The Supermicro
Advantage

Page 6

Product Portfolio

- Rackmount
- Multi-Node
- GPU Servers
- Storage
- Edge
- Workstations

Page 22

Liquid Cooling

Page 75

System Management

Page 81

Services

Page 84

Summary

Page 90

Disclaimer

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

Super micro computer, inc. Makes no representations or warranties with respect to the contents hereof and assumes no responsibility for any inaccuracies, errors or omissions that may appear in this information.

Super micro computer, inc. Specifically disclaims any implied warranties of merchantability or fitness for any particular purpose. In no event will super micro computer, inc. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, inc. Is expressly advised of the possibility of such damages.

Attribution

© 2026 Super Micro Computer, Inc. All rights reserved.

Supermicro Overview

Fortune 500 and S&P 500

\$22B Revenue & 134% YoY Growth

Fastest Time-to-Market

Founded and Headquartered in Silicon
Valley for 30 years



Source: IDC



Most Extensive Workload Optimized Portfolio



First-to-Market

We deploy the latest technology faster and recently deployed the world's largest liquid-cooled AI cluster in just 122 days

Broadest IT Product Portfolio

Application-optimized solutions serving all markets. For example, AI, cloud, data center, enterprise, big data, HPC, 5G, IoT, embedded, edge.

Vertically Integrated

Our complete control of our supply chain improves quality, reduces cost, and shortens lead times for you.

Universal GPU, PCIe GPU, SuperBlade, Hyper, Hyper-E, CloudDC, WIO, Petascale All-Flash, FlexTwin, GrandTwin, BigTwin, IoT/5G

18MW

Mass Production Center

High efficiency, reduced greenhouse gas emissions, minimized air pollutants and reduced water use.

5,000+/mo.

Rack Capacity

Production, testing and shipping for integrated full-scale rack solutions.

Up to **800Gb/s**

Network Speed Testing Environment

Complete testing environment from a wide range of network requirements.

Up to **480VAC**

Power at Scale

Including 208V, 415V, 277V/480VAC for single and 3-phase & 48VDC-ready.

Global Production at Scale



In-House Design with Global Production at Scale

4 Production Locations

USA, Europe, Taiwan, Malaysia

Operations in 100+ Countries

2000+ Liquid Cooled Racks per Month

1200+ Manufacturing Capacity for NVIDIA
GB200 NVL72 per Month



One-Stop Shop for Full Data Center Solutions

Supermicro Data Center Building Block Solutions



Data Center End-to-End Liquid-Cooling

Up to 40% More AI per Watt with The Next Generation Direct-to-Chip Liquid Cooling Solutions

Power Savings

up to **40%**

Savings in entire data center
(vs. Air-cooling) by using
Supermicro DLC-2

Water Savings

up to **40%**

Savings with 45 °C warm water
operation and eliminating
chilled water and compressor

System Heat Capture

up to **98%**

Heat capture in DLC-2 Liquid-cooling
with CPU, GPU, PCIe Switch, DIMM,
VRM, PSU, and more

Quiet Data Center

as low as **50dB**

Significantly reduces noise with
less fans and fan speed. As
quiet as a library

Space Savings

up to **60%**

Savings with more than 2.5x
compute density compared to
air-cooled systems

Supermicro DLC-2 reduces power costs by up to 40%, accelerate time-to-deployment and time-to-online, and allow data centers to run more efficiently with lower PUE, while lowering TCO by up to 20%

SUPERMICRO

Reference Customers Installations



- MEMX
- Pollen
- Lamini
- Intel
- Applied Digital
- Atlantic.Net
- InMotion
- VEXXHOST
- HelpMeSee
- Lawrence Livermore National Laboratory

- NetProtect
- COTAS
- Vexata
- Esgyn
- NetFire
- Universidad Nacional de Córdoba
- Petrobras
- APAC
- SAKURA Internet
- Getworks

- Onetwork
- Ahrefs
- NEC
- Osaka University
- Preferred Networks
- Getworks
- Eviden
- Genesis Cloud
- Absolute Hosting
- Seeweb

- ElioVP
- CERN
- GleSYS
- Nitrado
- IDLab
- Mainstreaming
- Queenstown Educational Foundation
- SZTAKI

Largest AI Supercomputer Powered by Supermicro

World's Largest Liquid-Cooled AI Cluster

- 100,000 NVIDIA H100 Tensor Core GPUs
- Multi-billion-dollar cluster, deployed in 122 days
- The basic building block for this Supercomputer is the Supermicro liquid-cooled rack
- 8x 4U servers each with 8 GPUs, for a total of 64 GPUs per rack, plus a CDU
- 1,500+ Racks, 1,500+ CDUs, 300K+ cables
- Spectrum-X (400GB) Ethernet
- Supermicro's design is from the ground up to be liquid-cooled, and all from one vendor



Supermicro NVIDIA
HGX Systems

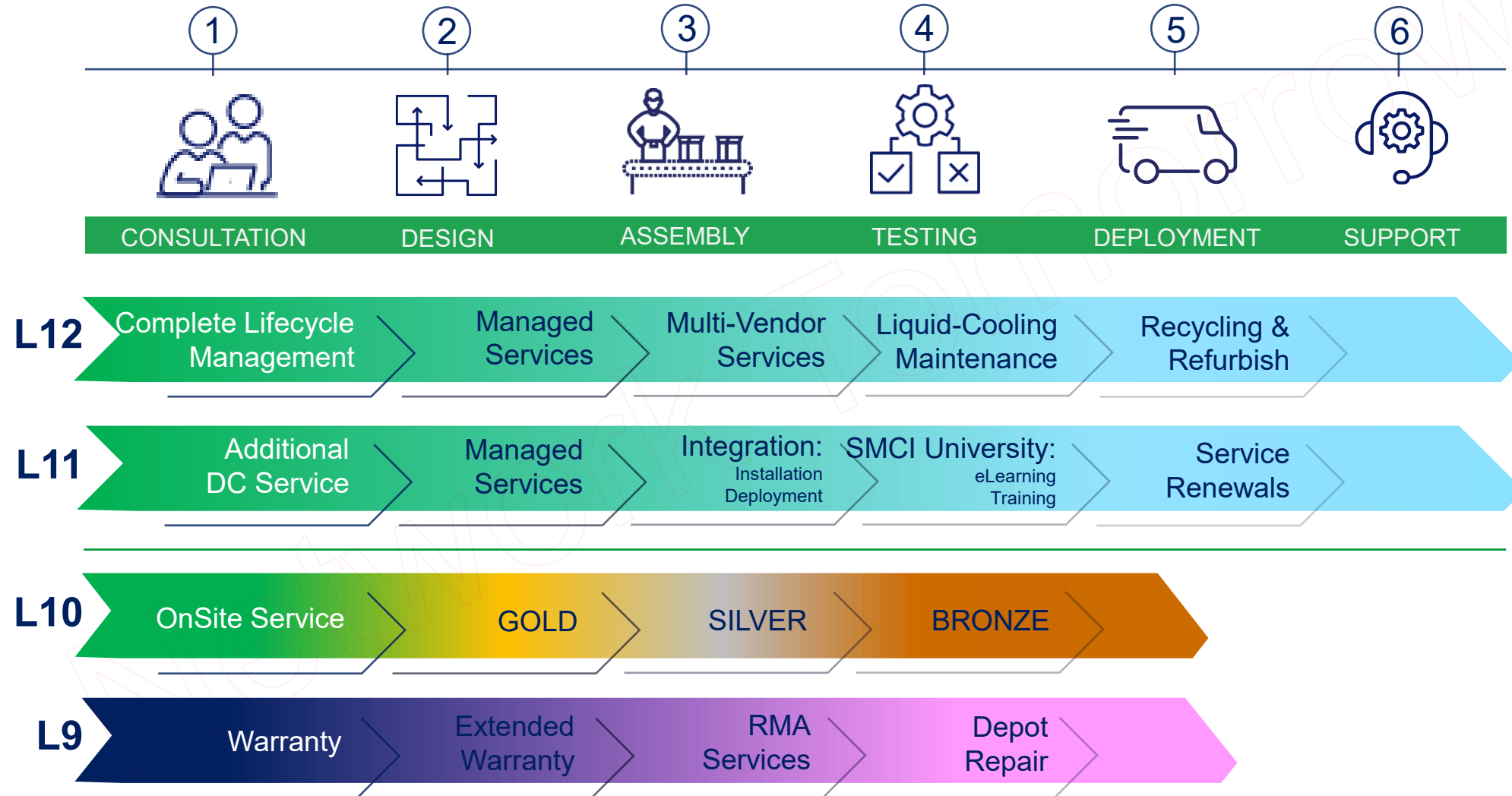


Supermicro Liquid
Cooling Total Solution



NVIDIA Spectrum-X
Ethernet networking

End-to-End Services for your Business Needs



The Supermicro Advantage

- 1 Most Extensive Workload-optimized Portfolio
- 2 In House Design with Global Production at Scale
- 3 One-stop-Shop for Full Data Center Solutions
- 4 First-to-Market Innovation
- 5 End-to-End Services

Engineering Focused... Customer Driven!



Data Center Building Block Solutions®(DCBBS)

End-to-End Solutions from System to Rack to Data Center Scale



Supermicro DCBBS delivers complete, modular AI infrastructure—from GPUs and networking to racks, site infrastructure, management software, and services.



Rapid Deployment
With First-to-Market



Customized
to Your Workloads



Total Solutions
with Onsite Services

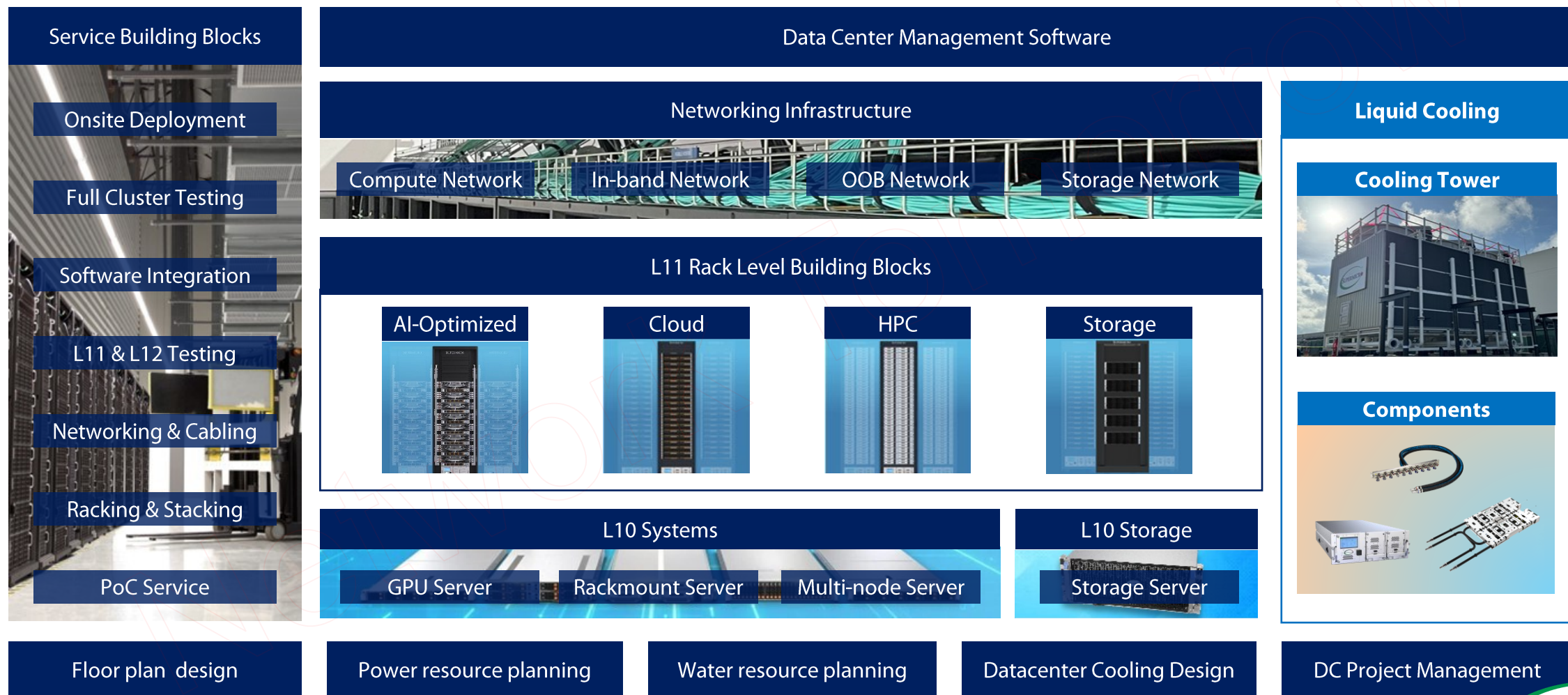


Tested and Validated
Before Shipping

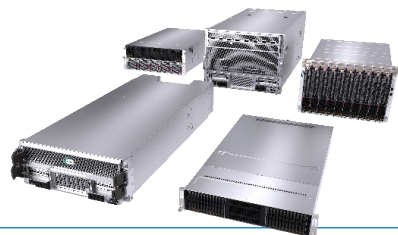
Data Center Building Block Solutions



Data Center Building Block Solutions



DCBBS Components



AI and Compute Systems

Industry's broadest AI and compute system portfolio with unmatched customization—from GPU systems to petascale storage, multi-node systems, and liquid cooling.



In-Rack Solutions

Complete in-rack integration—cooling distribution units & manifolds, hose kits, networking switches, power shelf, and battery backup unit, RDHx for maximum density and efficiency.



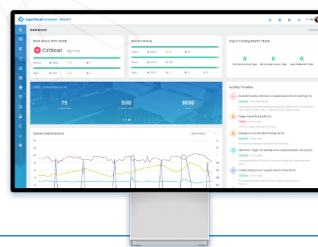
In-Row Solutions

Scalable liquid cooling from 200kW Liquid-to-Air Sidecars to 1.8MW In-Row CDUs, plus pre-validated SuperCluster solutions with integrated networking for turnkey multi-rack deployment.



Site Infrastructure Solutions

Site-level cooling infrastructure with scalable water and dry cooling towers for thermal and power management. Optimized cabling design and implementation for high-performance data pipeline.



Management Software Suite

End-to-end SuperCloud software from bare metal to AI workloads—delivering unified infrastructure control, deployment automation, developer tools, and multi-tenant GPU cloud management.

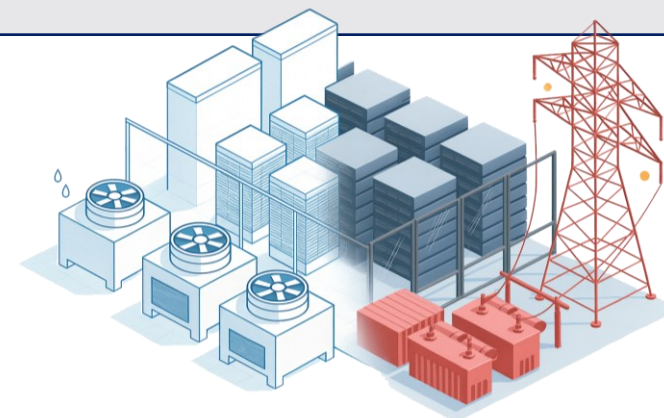


Services and Onsite Deployment

Comprehensive professional services from data center design and solution validation through on-site deployment and ongoing support, with 4-hour response options for mission-critical uptime.

Data Center Solution Offering

Components	Fit-Out	Colo	Build / Retrofit	
Design	- Architecture design - Engineering services - Turn-key solution	Turn-Key Data Hall Ready	- Architecture design - Engineering services - Turn-key solution	- Architecture design - Engineering services - Turn-key solution
Construction	Permitting Services	Turn-Key Direct Liquid Cooling Data Hall	- General contractor - Permitting Services	Permitting Services
Electrical	- Fit out from the data center PDUs to the servers. - Busway, Whips, Rack PDUs	Turn-Key Power to the server ready for operations	- Redundancy scheme - Blocks/systems	Evaluate and update Redundancy Scheme, and existing systems
Mechanical	Install racks, CDU, and DLC piping for complete operational readiness	- Turn-Key CDUs connected and tested. - Water Treatment managed and monitored	- Redundant loop for chilled water - DLC piping	- Direct Liquid Cooling installation from existing chilled loops - Connect CDUs ready for operation
Plumbing	Updates as required	Updates as required	Building infrastructure	Updates as required



Business Protection from Cradle to Grave

Zero Trust: Assume nothing is secured until verified. Every component, process and product is designed and validated for security across the full lifecycle



Trusted and Verified as a U.S. DoD Top-Tier Supplier

Secure Sourcing of Components

Visual inspection of components to insure no rogue components

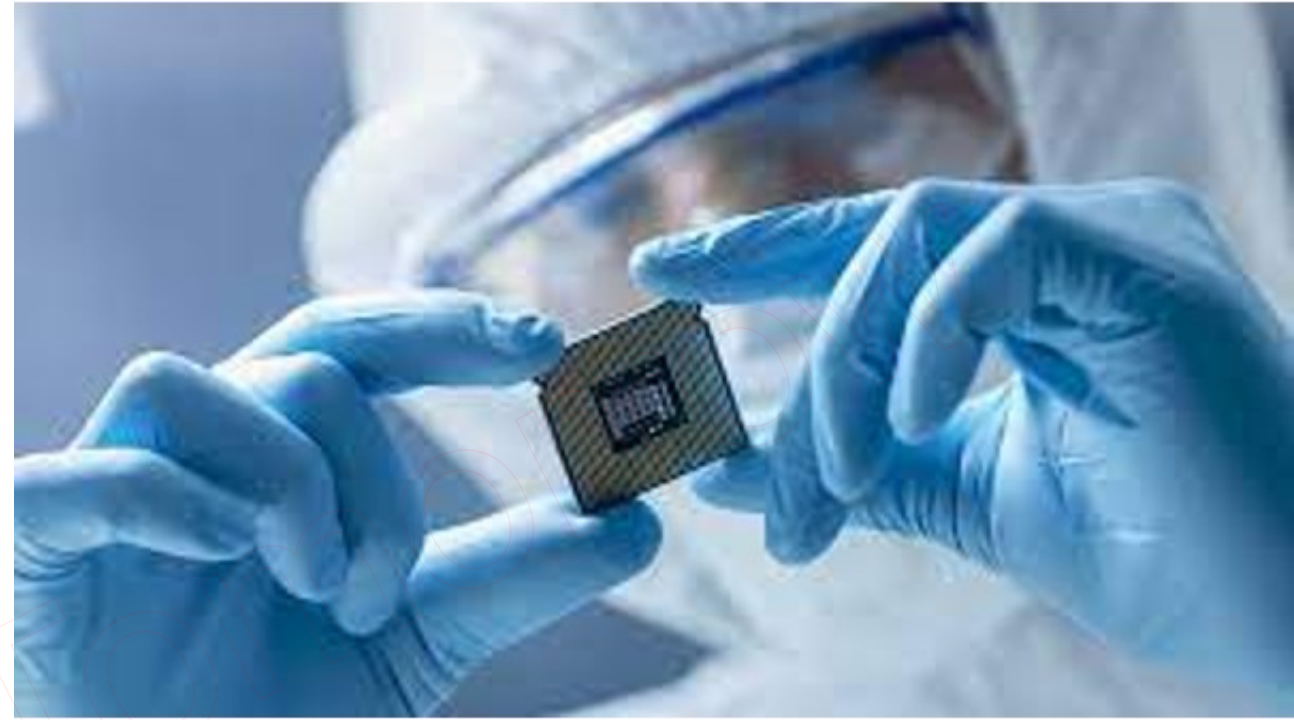
X-ray inspection of selected motherboards for greater scrutiny

Traceability of supplier components directly to our factory

Certification process that requires all suppliers to pass rigorous inspections from Supermicro

Management process that requires suppliers to receive regular training from Supermicro

Suppliers required to pass international Organization for Standards (ISO) Certification



Supply Chain Assurance for X14/H14

Systematic Asset & Configuration Validation Throughout the Server Lifecycle

Supermicro Production Asset Manifest

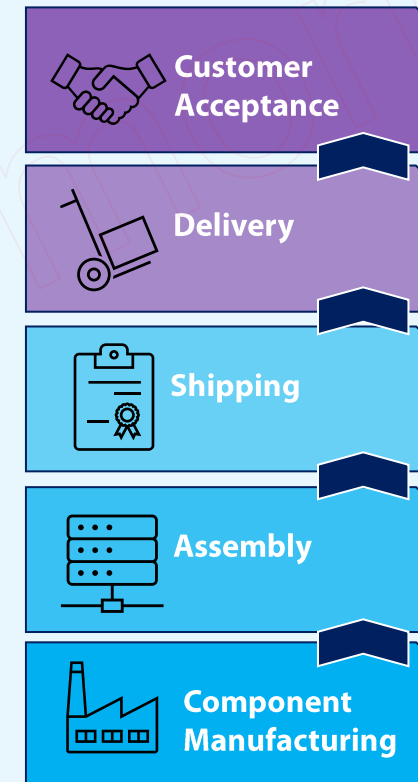
- Motherboard component-level manifest
- Server-level inventory manifest is generated by production & securely stored within Supermicro

Customer Asset Manifest Validation

- Online tools to securely access manifest via APIs
- Cryptographically* verify the integrity of server received matches the production manifest.

Post-Production Configuration Changes

- Customers' post-production configuration changes can be logged and stored in the manifest.



*Requires SDDC License

Secure Cryptographic Inventory/Asset Validation

Hardware Root of Trust (RoT 2.0)

Secure, Silicon-Based Cryptographic Foundation

NIST SP 800-193 Compliance

- Protection
- Detection & Logging
- Recovery

CNSA Approved & Recommended

- NIST Approved SHA 384, 512
- NIST Approved RSA 4096

Immutable Silicon-based Cryptographic Security

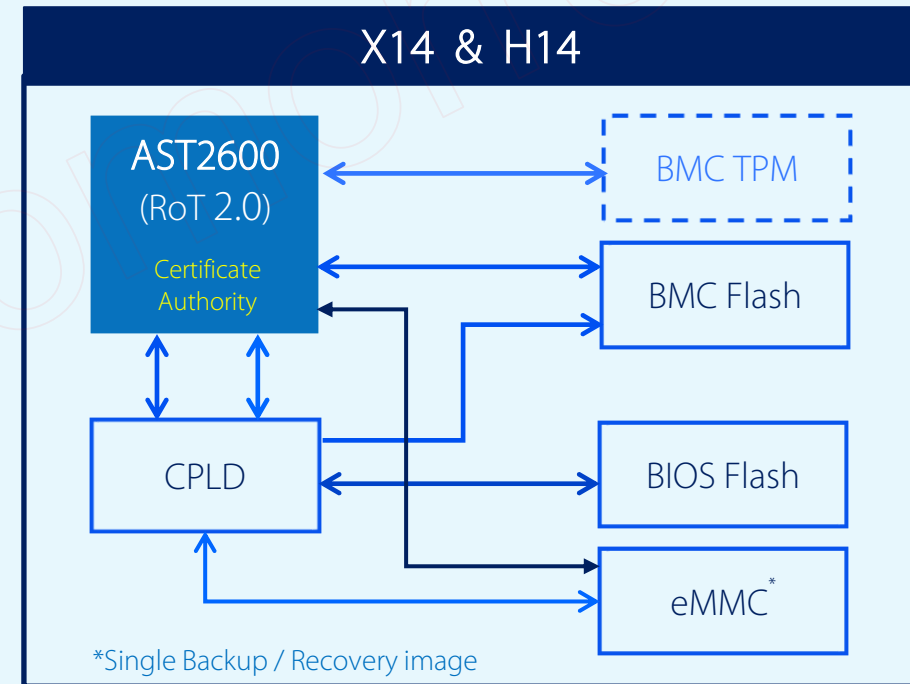
- Secure/Trusted boot process
- Foundation for the Chain of Trust

Secure Firmware Update

- Roll-back Prevention
- Web UI, Redfish, SUM, SMC IPMI tool

Secure Firmware Recovery

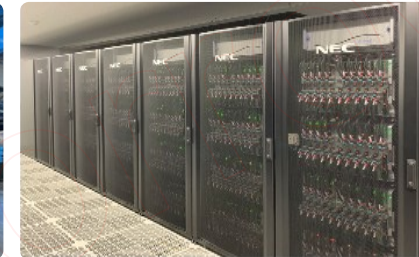
- Automatic BIOS and BMC recovery from backup or golden firmware image



Foundational to Secure HW and SW Chain of Trust

Supermicro Solution Integration Centers

- Top global solution vendor for Cloud, AI, HPC, Enterprise, 5G, IoT, embedded and edge computing
- 6M sq. ft. global manufacturing w/continuous capacity of 5,000+ “plug-n-play” integrated racks per month
- 10MW+, 208v, 230/405v, 277/480v capability. Air cooling, liquid cooling, immersion cooling support
- Quality Management System certified to ISO 9001:2015 and ISO 13485:2016 standards
- Secure integration/validation area; authorized access only
- Secret clearance for on-site services of U.S. DOE facilities



DESIGN

ASSEMBLY

CONFIGURATION

TESTING

DEPLOY & SERV



Green Computing Park,
USA

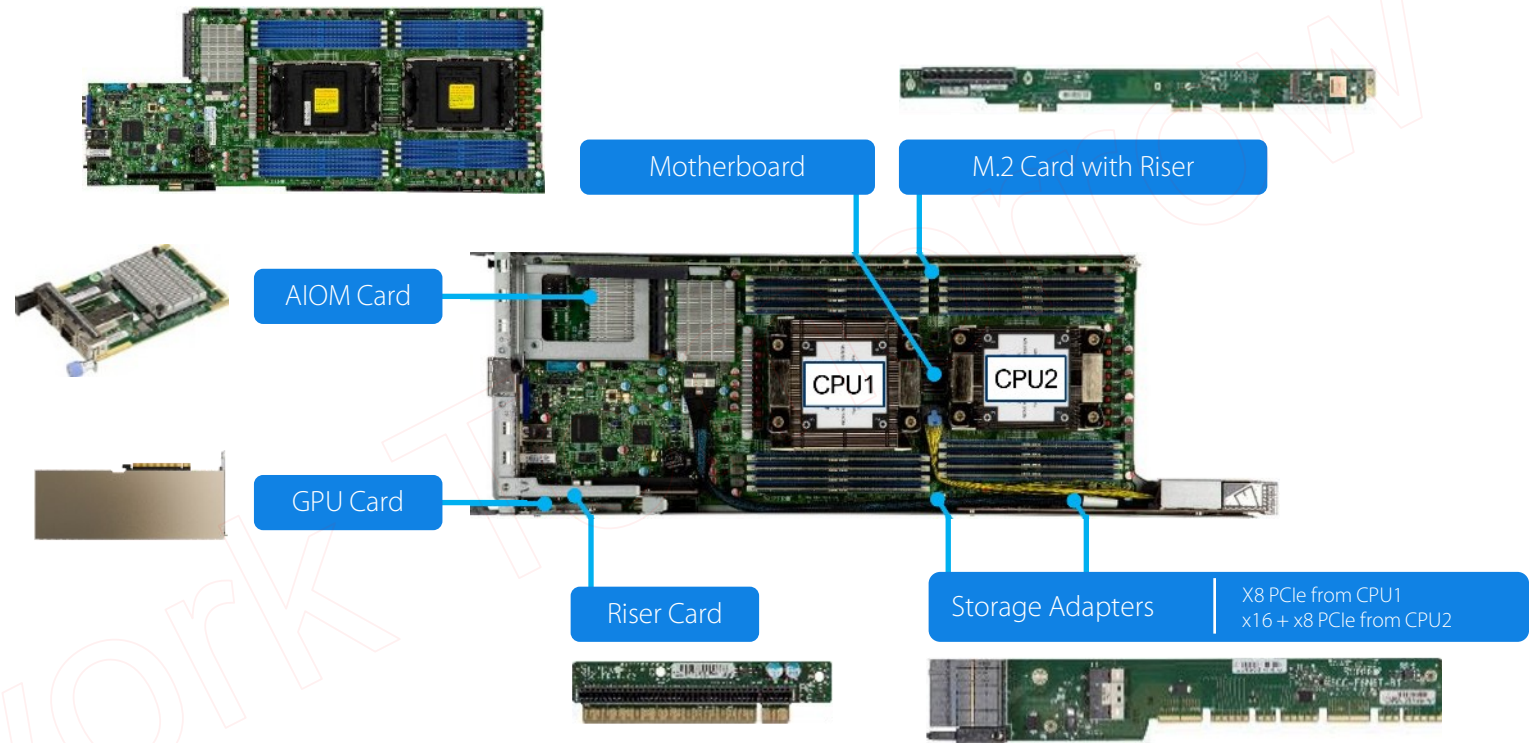


EMEA Operations Park,
Netherlands



Science & Tech Center,
Taiwan

Engineering Focused... Customer Driven!



Exceedingly Fast

- ✓ Time-to-Market
- ✓ Time-to-Solution



Building Block Solutions

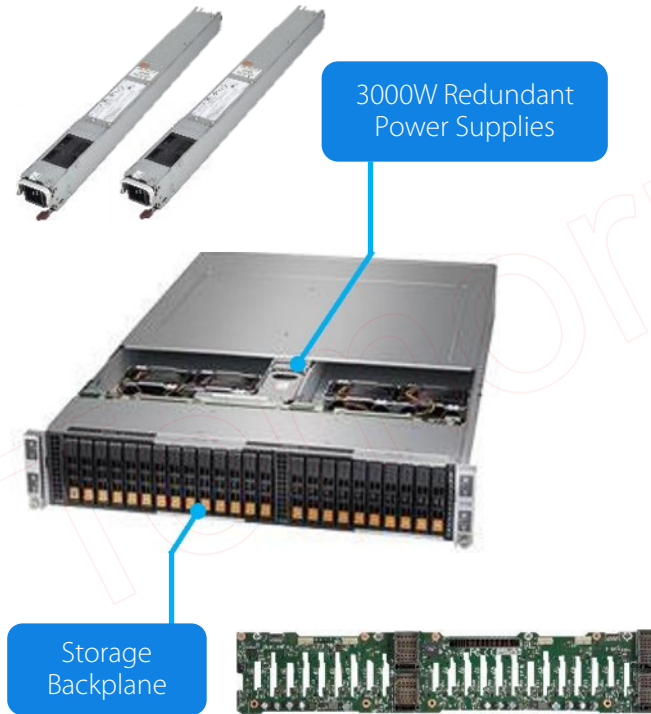
- ✓ Price/ Performance
- ✓ High-Quality



Unmatched...

- ✓ Design Control
- ✓ Flexibility

Engineering Focused... Customer Driven!



Exceedingly Fast

- ✓ Time-to-Market
- ✓ Time-to-Solution



Building Block Solutions

- ✓ Price/ Performance
- ✓ High-Quality



Unmatched...

- ✓ Design Control
- ✓ Flexibility

Supermicro Comprehensive Portfolio

Rackmount

Hyper



Cloud DC



WIO



MegaDC



MP



Multi-node

FlexTwin™



GrandTwin®



BigTwin®



SuperBlade®



MicroCloud



GPU Servers

HGX™ Platforms



PCIe GPU Platform



MGX™ Platforms



Edge/IoT

Rackmount Edge Compact Edge



Workstations



Petascake



Simply Double



Top Loading



Rackmount System Product Families

Hyper



Best-in-class Performance and Flexibility Rackmount Server

- Supermicro's next generation flagship high-performance, high-configurable 1U & 2U rackmount servers
- Single & Dual Intel® Xeon® or AMD EPYC™ processors
- Modular design to optimize customer's requirements: storage, I/O expansion, networking

CloudDC



All-in-one platform for cloud datacenters

- 1U & 2U rackmounts providing balanced performance & efficiency for cloud data center workloads
- Single & Dual Intel® Xeon® or AMD EPYC™ processors
- X14/H14 CloudDC based on OCP's Data Center Modular Hardware Systems (DC-MHS) for standardized server building block designs

WIO



General purpose

- Cost-effective 1U & 2U rackmount servers
- Single Intel® Xeon® or AMD EPYC™ processors
- Minimal memory DIMM slot configurations
- Supports full height full length PCIe expansion cards for I/O flexibility

MegaDC



Commercial Off The Shelf Modular Flexibility

- Best performance per watt with predictable processor performance and more energy efficient
- Optimized component integration for increased cost-effectiveness and reliability
- Modular Design for multiple platforms with rear I/O and front I/O configurations
- Flexibility for better operations efficiency

Multi-Processor (MP)



Mission Critical

- Multi-Processor high density rackmount servers
- Single node 2U, 4U & 6U form factor
- Quad or Octa Intel Xeon® processors (4-way / 8-way)
- Highest Density with largest memory capacity and CPU core count in a single node server.

Hyper: Enterprise Rackmounts

Best-in-Class Performance and Flexibility Rackmount Server

Designed to deliver best-in-class computing performance for the most demanding workloads

Offering both single and dual sockets to meet wider compute workloads with up to 500W TDP support for both Intel and AMD processors

Higher memory bandwidth and capacity

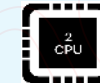
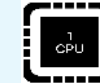
Intel Hyper supports up to 32 DDR5 RDIMM (6400MT/s) as well as DDR5 MRDIMM (8800MT/s/8000MT/s), delivering up to 37% higher memory bandwidth. AMD Hyper supports up to 24 DDR5 RDIMM (6000MT/s)

Large density of hybrid storage

Direct connect NVMe offering improved throughput and lower latency with PCIe Gen 5

Flexible expansion slot and networking configurations

Configurable x16 and x8 slots supporting the latest GPU, FPGA along with our AIOM (OCP3.0) for seamless network integration to provide the ultimate flexibility with tool-less design



Product Specs

- Dual and Single Intel® Xeon® 6900/6700/6500 series processors and AMD EPYC™ 9005/9004 series processors
- 32 DIMM (Dual Xeon® 6700/6500), 24 DIMM (Dual Xeon® 6900, Dual/Single AMD EPYC™ 9005/9004), 16 DIMM (Single Xeon® 6700/6500), 12 DIMM (Single Xeon® 6900, Single AMD EPYC™ 9005/9004)
- NVMe SSD support with up to 24 drives in 2U (Optional SAS and SATA configurations)
- Up to 3 PCIe 5.0 x16 slots in 1U or 8 PCIe 5.0 x8 (or 4 PCIe 5.0 x16) slots in 2U
- Up to 2 AIOM* slots for flexible networking options



SYS-122HA-TN-LCC / SYS-112HA-TN / AS -1126HS-TN
SYS-222HA-TN / SYS-212HA-TN / AS -2126HS-TN
SYS122H-TN / SYS-112H-TN
SYS-222H-TN / SYS-212H-TN



CloudDC: DC MHS Rackmounts

Ideal for Cloud Service Providers (CSPs) and Hyperscalers
Utilizing Multi-vendor Hardware Platforms

OCP Data Center Modular Hardware Systems (DC-MHS)

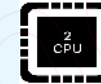
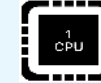
Standards for common server building blocks and interconnects to ensure interoperability of multi-vendor hardware and software within the data center

Data Center Secure Control Module (DC-SCM)

Provides common management and security infrastructure across multiple platforms within the data center and simplifies the adoption of open-source management firmware, such as OpenBMC

Enhance the sustainability of server hardware

Common server building blocks allow for component upgrades in existing infrastructures, enhancing compute density, reducing carbon footprint, and minimizing waste.



Product Specs

- Dual and Single Intel® Xeon® 6700/6500 series processors and AMD EPYC™ 9005/9004 series processors
- 32 DIMM (Dual Xeon® 6700/6500), 24 DIMM (Dual Xeon® 6900), 16 DIMM (Single Xeon® 6700/6500), 12 DIMM (Single AMD EPYC™ 9005/9004)
- NVMe SSD support with up to 24 drives in 2U (optional SAS and SATA configurations)
- Up to 2 PCIe 5.0 x16 + 1 PCIe 5.0 x8 slots in 1U or 6 PCIe 5.0 x16 slots in 2U
- Up to 2 AIOM slots for flexible networking options



SYS-122C-TN / SYS-112C-TN / AS-1116CS-TN
SYS-222C-TN



WIO: Mainstream Rackmounts

Best General Purpose 1U & 2U Rackmount Servers, Packed with Features and Values

Built-in features that maximize the saving

Without Intel Chipset, many systems rely on additional SAS controller to support SATA/SAS drives. Luckily, SATA controllers are built in on all X14 WIO systems. There are also dual 1GbE LAN ports that come with the systems.

Toolless & Top Loading expansion slots for breathy maintenance Widely loved top-loading expansion slots continue in X14 Gen. Small updates that allow more GPUs/AOCs to be supported. Tool-less design also helps streamline any maintenance tasks.

Enhance the sustainability of server hardware

Common server building blocks allow for component upgrades in existing infrastructures, enhancing compute density, reducing carbon footprint, and minimizing waste.



Product Specs

- Single Intel® Xeon® 6 SP processor or AMD EPYC™ 8004 Series processor
- Free air cooling potent thermal solution
- 8 DIMM slots supporting fastest DDR5-6400 with 1DPC
- Hybrid Gen 5 NVMe drive options available in 1U and 2U
- 3 PCIe 5.0 slots in 1U and up to 5 PCIe 5.0 slots in 2U
- Onboard SATA controller and dual RJ45 1GbE ports



SYS-112B-WR
SYS-512B-WR
SYS-522B-WR

MegaDC: Modular Flexibility COTS Enterprise Rackmounts

Commercial Off The Shelf (COTS) Modular Flexibility

Cloud Native

Best performance per watt with predictable processor performance and more energy efficient

Modular Flexibility

Modular Design for multiple platforms with rear I/O and front I/O configurations

Commercial Off The Shelf (COTS)

Optimized component integration for increased cost-effectiveness and better operations efficiency

17" (430mm) Short Depth Optimize for Edge Compute and Telco

Wide operation temperature with Double-Width Accelerators for the Edge deployment



Product Specs

- Single AmpereOneX Cloud-Native Processor up to 192 Core 3.2Ghz
- Free air cooling potent thermal solution
- 16 DIMM slots up to 4TB capacity support
- Hybrid Gen 5 NVMe drive options available in 1U and 2U
- Modular design for PCIe expansion option:
 - Four DW FHFL GPU support
 - 4/8/16/24 NVMe device
 - Single Root PCIe Switch with Four GPU
- Short Depth 17" system for Edge Compute



ARS-111M-NR
ARS-211M-NR
ARS-211ME-FNR



4-Way (4 Sockets) Multi-Processor

Large in-Memory Computing and Mission Critical Applications

Quad socket servers can achieve high density in data centers by consolidating significant compute power, memory, and I/O capabilities into a single chassis

Maximizing performance per rack unit for demanding workloads like virtualization, large-scale databases, and AI/ML applications

Higher memory bandwidth and capacity

Intel quad socket servers support up to 64 DDR5 RDIMM (6400MT/s), delivering up to 35% higher memory bandwidth compared to the previous generation.

Large density of hybrid storage

Direct connect NVMe offering improved throughput and lower latency with PCIe Gen 5.

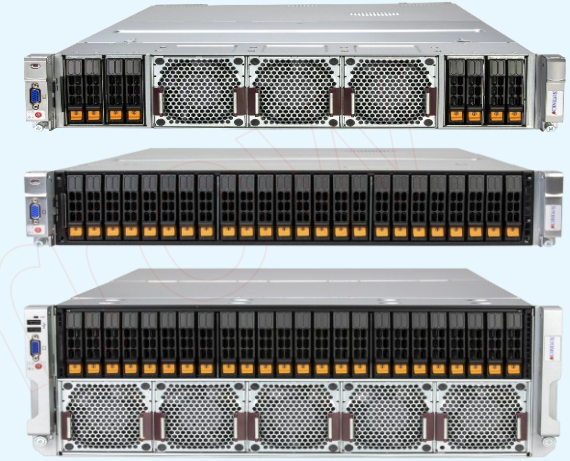
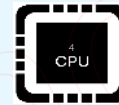
Supports SATA and SAS via an optional RAID controller

Flexible expansion slot and networking configuration

Configurable x16 and x8 slots supporting the latest GPU, FPGA, along with our AIOM (OCP3.0) for seamless network integration to provide the ultimate flexibility with tool-less design



SYS-242H-NR / SYS-242B-NR / SYS-442B-NR



Product Specs

- Quad Intel® Xeon® 6700P series processors, up to 86 cores and 350W TDP
- 64 DIMM of DDR5 6400MT/s, up to 16TB of total memory
- 8 to 24 U.2 drive bays, 2 onboard M.2 NVMe slots
 - SYS-242H-NR: 8x NVMe Gen 5.0
 - SYS-242B-NR / SYS-442B-NR: 24x NVMe Gen 5.0, Optional SATA/SAS support via optional RAID controller
- 6 to 18 PCIe Gen 5 Slots, 2 AIOM PCIe Gen 5 slots (OCP 3.0 Compatible)
 - SYS-242H-NR / SYS-242B-NR: 6 PCIe Gen 5 slots
 - SYS-442B-NR: 18 PCIe Gen 5 slots



8-Way (8 Sockets) Multi-Processor

Large in-Memory Computing and Mission Critical Applications

Octo socket servers can achieve high density in data centers by consolidating significant compute power, memory, and GPU capabilities into a single node

Maximizing performance per rack unit for demanding workloads like virtualization, large-scale databases, and AI/ML applications

Higher memory bandwidth and capacity

Intel Octo socket servers support up to 128 DDR5 RDIMM (4800MT/s), delivering up to 97% memory bandwidth.

Large density of hybrid storage

Direct connect NVMe offering improved throughput and lower latency with PCIe Gen 5.

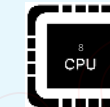
Supports SATA and SAS via optional RAID controller, up to 24 drives per system

Flexible expansion slot and networking configuration

Configurable x16 and x8 slots supporting up to 12 double-width L40S or H100 NVL GPU, FPGA, seamless network integration to provide the ultimate flexibility with tool-less design



SYS-681E-TR



Product Specs

- Octo Intel® Xeon® 8400H series processors, up to 60 cores and 350W TDP
- 128 DIMM of DDR5 4800MT/s, up to 32TB of total memory
- 24 2.5" NVMe/SATA/SAS hybrid drive bays, 2 onboard M.2 NVMe/SATA slots
 - Optional SAS support via an optional RAID controller
- 24 PCIe Gen 5 Slots, 2x Internal PCIe Gen 5 Slots (x16) (Optional)
 - Default 12 PCIe Gen 5 Slots and 2x Internal PCIe Gen 5 Slots (x16) (Optional)
 - 12 additional 12 PCIe Gen 5 Slots upgrade available

Multi-Node (Twin) System Product Families

BigTwin® Systems



Efficient Dual Processor System

- Dual Intel® Xeon® 6700/6500 with balance compute, storage, and network
- Air or liquid cooled
- Optimized for:
 - Hyper Converged Infrastructure
 - Virtualization
 - Software-defined Storage
 - Data Services

GrandTwin® Systems



Cost-Effective Single Processor System

- Single Intel® Xeon® 6700/6500/6900 or AMD EPYC™ 9005/9004 processor
- Rear or front I/O options. Air Cooled
- Optimized for:
 - Networking and Media
 - Cloud
 - Web Applications
 - Microservices

FlexTwin™ Systems



Purpose-Built HPC System

- Dual Intel® Xeon® 6900 or AMD EPYC™ 9005 processors
- High-density liquid-cooled only
- Optimized for:
 - HPC Data Center
 - Financial Services
 - Scientific Research
 - Oil and Gas

FlexTwin™: Multi-Node for HPC

Purpose-Built HPC Dual Processor Multi-Node Server

High-Performance Compute

Experience fast performance with support for the newest Intel and AMD processors, coupled with the power of 12-channel memory

Cost-Optimized Design

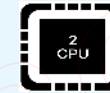
Invest wisely with a pay-for-what-you-need approach and enjoy the cost-effectiveness of liquid cooling and front-serviceable design

Increased Density and Energy Efficiency

Invest wisely with a pay-for-what-you-need approach and enjoy the cost-effectiveness of liquid cooling and front-serviceable design

Improved Cluster Availability

Ensure continuous operation with redundant power supplies and CDU pumps and the peace of mind that comes with single-vendor accountability



Product Specs

- Dual Intel® Xeon® 6900 series and AMD EPYC™ 9005 series processors
- Up to 24 DIMMs supporting up to 6TB 6400MT/s DDR5 RDIMM or 3TB 8800MT/s DDR5 MRDIMM
- Flexible networking with up to 1 OCP 3.0 compatible AIOM slots
- Default 1 PCIe 5.0 x16 LP slot + Optional 1 PCIe 5.0 x16 FHHL slot
- Optional 2 E1.S slots + Optional 2 M.2 NVMe drive slots
- 4 redundant 3200W Titanium Level Power Supplies



SYS-222FT-HEA-LCC and AS -2126FT-HE-LCC



BigTwin®: Multi-Purpose/Multi-Node

Balanced 2U Dual-Processor Multi-Node Server

Improve Energy Efficiency

Experience enhanced efficiency with shared power and cooling architecture and achieve up to 8.1% greater energy savings compared to traditional air cool 1U servers

Scalable Storage

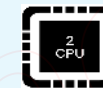
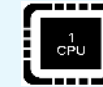
Optimize your software-defined storage deployments with our dense solution, supporting U.2, U.3, and E3.S for maximum capacity and performance

Dense Compute Performance

Performance density with up to 8 CPUs packed into a compact 2U form factor, delivering up to 1,152 efficient cores for demanding workloads

Customizable Designs

Density flexibility ranging from 2-node to 4-node configurations, support for various drive form factors, and GPU support for 2U 2-node



2U 4-Node



2U 2-Node



Product Specs

- Dual Intel® Xeon® 6700 series processors
- Up to 16 DIMMs supporting up to 2TB 6400MT/s DDR5 RDIMM
- Storage:
 - 2U 4-Node: 6 U.2 or 8 NVMe drives per node
 - 2U 2-Node: 12 U.2 NVMe per node
- Expansion Slots:
 - 2U 4-Node: 2 PCIe 5.0 x16 slots and 1 AIOM OCP 3.0
 - 2U 2-Node: 1 PCIe 5.0 x16 slot, 2 PCIe 5.0 x8 slots, and 1 AIOM OCP 3.0



2U4N: SYS-222BT-HNR/HNC8R/HNC9R/HER
2U2N: SYS-222BT-DNR/DNC8R



GrandTwin[®]: Single Processor Multi-Node

Cost-Effective Single-Processor Multi-Node Server

Efficient Performance

Support for single Intel or AMD processors with high-bandwidth DDR5 6400 MT/s memory

Cost-Effective Design

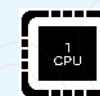
Cost-effective networking expansion enabled by the front I/O module while optimizing efficiency with shared infrastructure

Flexible Architecture

Accelerate parallel processing workloads with flexible rear and front I/O options and support for high-performance FHFL GPUs

Improved Availability

Front node accessing for better serviceability



2U 2-Node



2U 2-Node



Product Specs

- Single Intel® Xeon® 6900/6700 series and AMD EPYC™ 9005/9004 series
- Memory:
 - 2U 4-Node: Up to 16 DIMMs (2 DPC) supporting up to 2TB 6400MT/s DDR5 RDIMM
 - 2U 2-Node: Up to 12 DIMMs (1 DPC) supporting up to 4TB 6400MT/s DDR5 RDIMM or 2TB 8800MT/s MRDIMM
- Storage:
 - 2U 4-Node: 6 2.5" NVMe or 8 E1.S NVMe + 2 M.2 drives per node
 - 2U 2-Node: 4 2.5" NVMe + 2 M.2 drives per node
- Expansion Slots:
 - 2U 4-Node: Up to 1 PCIe LP x16 + up to 2 AIOM x16 per node
 - 2U 2-Node: Up to 1 PCIe FHFL x16 or 2 PCIe LP x16 + 1 AIOM x16 per node



2U4N: SYS-212GT-HNF/HNR and AS -2116GT-HNTF
2U2N: SYS-212GT-DNAF



Multi-Node (MicroCloud) System Product Families

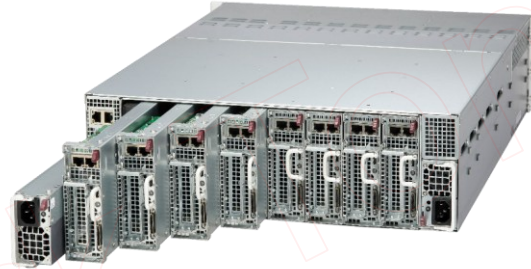
3U5N MicroCloud



3U5N MicroCloud GPU UP

- Double width GPU up to 400W supported
- AMD EPYC™ 4000 / AMD Ryzen™ processors
- 2x NVMe U.2 or 2x SATA per node
- 2x 1/10/25G MLP networking option port per node
- 1x PCIe Gen5 double width GPU per node or 2x PCIe FHFL x8 single width per node
- 1x Dedicated IPMI per node via Centralized Dedicated Port
- 80 nodes in 48U rack

3U8N MicroCloud



3U8N MicroCloud UP

- Rear IO with hot swap drive design
- Lowest cost and most energy-efficient server architecture
- Intel Xeon-E 6300/2400 and AMD EPYC™ 4000 / AMD Ryzen™ processors
- 1x Dedicated IPMI port per enclosure
- 2x 1/10/25G MLP networking option port per node
- LP GPU supported
- 128 nodes in 48U rack

3U10N MicroCloud



3U10N MicroCloud UP

- Highest density on the market
- AMD EPYC™ 4000 / AMD Ryzen™ processors
- 1x Standard LP PCI-E x8 / x16 slots for GPU/NIC/AOC supported per node
- LP GPU supported
- 1x Dedicated IPMI port per enclosure
- 2x 1/10/25G MLP networking option port per node
- 160 nodes in 48U rack

3U5N MicroCloud

Ideal for Enterprise for Inferencing, Edge AI and 3D graphics

Double width GPU up to 400W supported

Entry level GPU server solution for their core business applications at most 80 CPUs in a 48U rack.

Differ from traditional GPU server

Unlike traditional GPU servers that share GPUs or concentrate many GPUs under one OS, this design gives clean workload separation with dedicated acceleration per node.

The 3U5N MicroCloud is optimized for inference and edge AI

Best-fit workloads include:

AI inference

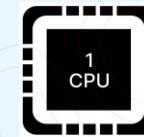
Video transcoding and analytics

Virtualized GPU workloads (per-tenant isolation)

Media streaming acceleration

AMD
EPYC

RYZEN
AMD



Product Specs

- Single Socket AM5 (LGA-1718) AMD EPYC™ 4005/4004 or Ryzen™ 7000 / 9000 Series Processor up to 170W with air cooling
- Support for up to 1 double-width PCIe GPU accelerator card
- Up to 192GB DDR5-5600MT/s (1DPC)/ 3600MT/s (2DPC), in 4 DIMM sockets
- Default Config: 1 PCIe 5.0 x16 FHFL + 1 PCIe 5.0 x4 MicroLP slots
- Optional Config: 2 PCIe 5.0 x8 (1 FHFL + 1 LP) + 1 PCIe 5.0 x4 MicroLP slots
- Up to 2 internal fixed 2.5" NVMe + 2 internal fixed 2.5" SATA drive bays
- 4 Redundant 2000W Titanium Level power supplies
- 3U Rackmount chassis with 30.51" (775mm) depth



AS -3015MR-H5TNR



3U8N MicroCloud

Balanced performance with cost effective per node

Most popular MicroCloud SKU

Max Cost-Performance aggregate solution in 3U 8Nodes. Support both Intel and AMD CPU platform.

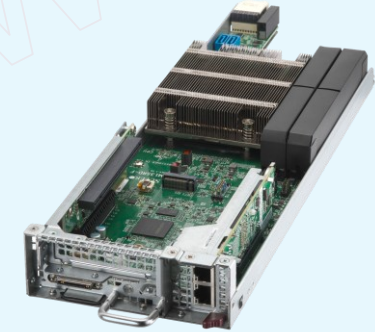
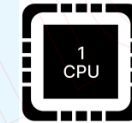
Value Proposition

High density without sacrificing reliability.

Flexibility to control cost per node.

Simpler scaling versus traditional servers.

Ideal for dedicated hosting, where noisy neighbors and fault isolation are critical.



Product Specs

Intel

- Intel® Xeon® 6300-Series/E-2400/Pentium® G7400 and G7400T Processor Single Socket LGA-1700 (Socket V0) supported, CPU TDP supports Up to 95W TDP

AMD

- Single Socket AM5 (LGA-1718) AMD EPYC™ 4005/4004 or Ryzen™ 9000/7000 Series Processor up to 170W with air cooling
- 8 sets of up to 192GB ECC/non ECC UDIMM; DDR5 5600MHz, in 4 DIMM sockets
- 8 sets of 1 PCIe 5.0 x8 (in x16) LP slots + 1 Micro-LP slot
- 8 sets of 2 front NVMe U.2/ SAS/ SATA3 drive bays with optional kits
- 2200W Redundant 80+ Titanium Level high-efficiency power supplies



SYS-531MC-H8TNR



AS-3015MR-H8TNR



3U10N MicroCloud

Highest Density Single Processor Multi-Node System for Cloud and Dedicated Hosting

Max Density

Highest density of nodes in an 3U form factor with up to 10 nodes. In a 48U rack, it will have up to 160 nodes.

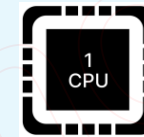
Space and Energy Efficiency

Maximize the productivity per square foot with higher server-node density. Shared power and cooling to save infrastructure investment.

MicroCloud offers up to 12.5% less power per node and 70% more in density improvements. MicroCloud 3U10N have the lowest cost and most energy-efficient server architecture.

AMD
EPYC

RYZEN
AMD



Product Specs

- Single Socket AM5 (LGA-1718) AMD EPYC™ 4005/4004 or Ryzen™ 7000 / 9000 Series Processor up to 170W with air cooling
- Support for up to 1 single-width PCIe GPU accelerator cards
- Up to 192GB DDR5-5600MT/s (1DPC)/ 3600MT/s (2DPC), in 4 DIMM sockets
- Up to 1 PCIe 5.0 x16 LP + 1 PCIe 5.0 x4 MicroLP slots
- Up to 4 internal 2.5" (2 NVMe and 2 SATA) drive bays
- 2000W Redundant (2+2 or 3+1) 80+ Titanium Level high-efficiency power supplies (2 PSUs by option)



AS -3015MR-H10TNR



Supermicro GPU Server Product Family

HGX™/OAM Platforms



CPU: INTEL AMD
GPU: NVIDIA INTEL AMD

- Maximum performance: 4-GPU and 8-GPU for training large AI models with pool of shared memory.
- Proven system architecture. Validated in real-world AI deployments.

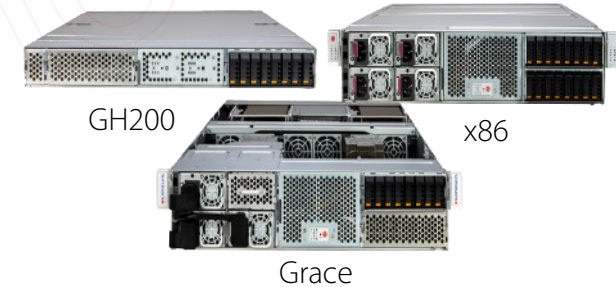
PCIe Gen5 Platforms



CPU: INTEL AMD
GPU: NVIDIA INTEL AMD

- Supports up to 10 PCIe GPUs, suitable for AI inference deployments, Cloud, and multi-tenant applications.
- Flexible GPU options like H100 PCIe for AI, L40S for video and MI210.

MGX™ Platforms



CPU: ARM INTEL
GPU: NVIDIA PCIe GH200

- Modular platform with both x86 and ARM CPUs.
- Grace Hopper Superchip systems and PCIe-based systems.
- Compact form factors for high computing density.

8U NVIDIA HGX™ H200 Server

Supercharging Generative AI and High-Performance Computing

Unmatched End-to-End Accelerated Computing Platform

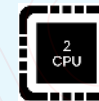
H200 GPUs feature the new Transformer Engine, with FP8 precision, that provides up to 5X faster training and 30X faster inference over A100 GPU for LLMs. This HGX system provides 32 petaFLOPS of AI performance (FP8)

Unmatched End-to-End Accelerated Computing Platform

The combination of faster (up to 4.8TB/s Bandwidth) and larger (1.1TB) HBM3e memory accelerates performance of computationally intensive generative AI and HPC applications leading to up to a 110X faster time to results than x86 CPU.

Reduced Energy and TCO

Hopper systems provide 5X more energy savings and 4X better cost of ownership savings over the NVIDIA Ampere generation



Product Specs

- Dual Intel® Xeon® 5th/4th Gen Scalable Processors, up to 350 TDP / AMD EPYC™ 9004/9005 Series Processors, up to 400W TDP
- 8 H200 SXM5 GPUs with NVSwitch, 900GB/s NVLink® GPU-GPU, up to 700W TDP (per GPU), air cooled
- 32 DIMM supporting up to 4TB 5600MT/s (1 DPC) or 8TB 4400MT (2 DPC) DDR5 RDIMM (Intel)/ 24 DIMM (1DPC only) supporting up to 6TB 6400MT/s DDR5 RDIMM (AMD)
- Up to 12 U.2 NVMe storage drives, 2 M.2 NVMe, 3 2.5" SATA
- 8 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHFL; 2 PCIe 5.0 x16 FHHL (optional)
- 2 USB 3.0 ports, 1 VGA port, 1 IPMI GbE LAN port, 2 10G NIC (X710)
- 6 redundant (4+2) 3000W Titanium level power supplies



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



4U (DLC) NVIDIA HGX™ H200 Server

Supercharging Generative AI and High-Performance Computing

Unmatched End-to-End Accelerated Computing Platform

H200 GPUs feature the new Transformer Engine, with FP8 precision, that provides up to 5X faster training and 30X faster inference over A100 GPU for LLMs. This HGX system provides 32 petaFLOPS of AI performance (FP8)

Latest Memory Technology: Supercharged HPC

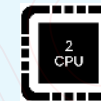
The combination of faster (up to 4.8TB/s Bandwidth) and larger (1.1TB) HBM3e memory accelerates performance of computationally intensive generative AI and HPC applications leading to up to a 110X faster time to results than x86 CPU

Increased Density (Double)

Liquid cooling enables more computing power to be packed into a smaller space compared to air-cooled systems, maximizing data center utilization

Reduced Energy and TCO

Hopper systems provide 5X more energy savings and 4X better cost of ownership savings over the NVIDIA Ampere generation



Product Specs

- Dual Intel® Xeon® 5th/4th Gen Scalable Processors, Direct Liquid Cooled/ AMD EPYC™ 9004/9005 Series Processors, up to 400W TDP
- 8 H200 SXM5 GPUs with NVSwitch, Direct Liquid cooled
- 32 DIMM supporting up to 4TB 5600MT/s (1 DPC) or 8TB 4400MT (2 DPC) DDR5 RDIMM (Intel) / 24 DIMM (1DPC only) supporting up to 6TB 6400MT/s DDR5 RDIMM (AMD)
- Up to 8 U.2 NVMe storage drives, 2 M.2 NVMe
- 8 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHHL; 2 PCIe 5.0 x16 FHHL (optional)
- 2 USB 3.0 ports, 1 VGA port, 1 IPMI GbE LAN port, 2 10G NIC (X710)
- 4 redundant (2+2) 5250W Titanium level power supplies



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



10U NVIDIA HGX™ B200 Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI

Scale-up x86 platform with up to 15X faster real-time inference performance and 3X faster training for LLMs than Hopper Generation (H100)

Advancing Data Analytics

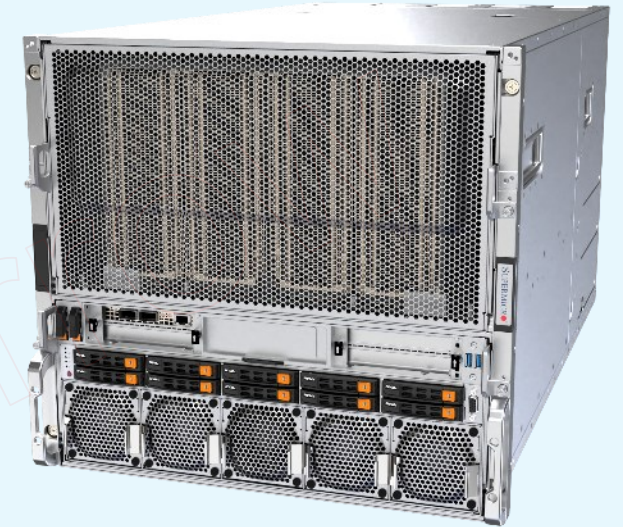
Blackwell's new dedicated Decompression Engine accelerates the full pipeline of database queries for the highest performance in data analytics and data science. HGX B200 perform 6X faster than CPUs and 2X faster than H100 for query benchmarks

Higher GPU Compute and Memory footprint

5th Gen NVLink® with 1.8TB/s of GPU-GPU Bandwidth, 1.4TB HBM3e (Up to 8TB/s Bandwidth) Memory, 144 petaFLOPS of AI performance (FP4)

Sustainable Computing

Sustainable computing is realized with efficiency gains as HGX B200 improves energy efficiency by 12X and lowers costs (TCO) by 12X compared to the Hopper



Product Specs

- Dual Intel® Xeon® 6900 Processors with P-cores, up to 500W TDP / AMD EPYC™ 9005/9004 Series Processors, up to 500W TDP
- 8 B200 SXM6 GPUs with NVSwitch, up to 1000W TDP per GPU, air cooled
- 24 DIMM (1DPC only) supporting up to 6TB 6400MT/s DDR5 RDIMM (Intel/AMD) or 6TB 8800MT/s DDR5 MRDIMM (Intel)
- Up to 10 U.2 NVMe storage drives, 2 M.2 NVMe
- 8 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHHL
- 2 USB 3.0 ports, 1 VGA port, 1 IPMI GbE LAN port, 2 10G NIC (X710)
- 6 redundant (3+3) 5250W Titanium level power supplies



SYS-A22GA-NBRT / AS -A126GS-TNBR



10U NVIDIA HGX™ B200 Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI

Scale-up x86 platform with up to 15X faster real-time inference performance and 3X faster training for LLMs than Hopper Generation (H100)

Advancing Data Analytics

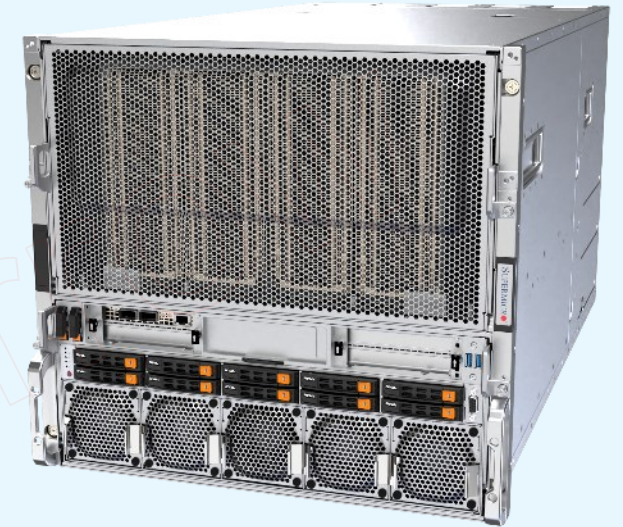
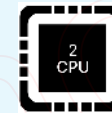
Blackwell's new dedicated Decompression Engine accelerates the full pipeline of database queries for the highest performance in data analytics and data science. HGX B200 perform 6X faster than CPUs and 2X faster than H100 for query benchmarks

Higher GPU Compute and Memory footprint

5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 1.4TB HBM3e (Up to 8TB/s Bandwidth) Memory, 144 petaFLOPS of AI performance (FP4)

Sustainable Computing

Sustainable computing is realized with efficiency gains as HGX B200 improves energy efficiency by 12X and lowers costs (TCO) by 12X compared to the Hopper



Product Specs

- Dual Intel® Xeon® 4/5th Gen Processors, up to 350W TDP
- 8 B200 SXM6 GPUs with NVSwitch, up to 1000W TDP per GPU, air cooled
- 32 DIMM (1DPC only) supporting up to 8TB 6400MT/s DDR5 RDIMM
- Up to 10 U.2/U.3 NVMe storage drives, 2 M.2 NVMe
- 8 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHHL
- 2 USB 3.0 ports, 1 VGA port, 1 IPMI GbE LAN port, 2 10G NIC (X710)
- 6 redundant (3+3) 5250W Titanium level power supplies



SYS-A21GE-NBRT



8U NVIDIA HGX™ B200 FIO Server

Purpose Built for Large Scale AI and HPC Workload

New Features and Benefits

GNR-SP 6700P CPU support with 32 DDR5 DIMM, FIO with 8 NICs

Accelerated Generative AI and Data Analytics

Scale-up x86 platform with up to 15X faster real-time inference performance and 3X faster training for LLMs than Hopper Generation (H100)

Advancing Data Analytics

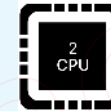
Blackwell's new dedicated Decompression Engine accelerates the full pipeline of database queries for the highest performance in data analytics and data science. HGX B200 perform 6X faster than CPUs and 2X faster than H100 for query benchmarks.

Higher GPU Compute and Memory Footprint

5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 1.4TB HBM3e (Up to 8TB/s Bandwidth) Memory, 144 petaFLOPS of AI performance (FP4)

Sustainable Computing

Sustainable computing is realized with efficiency gains as HGX B200 improves energy efficiency by 12x and lower costs (TCO) by 12X compared to the Hopper.



Product Specs

- Dual Intel® Xeon® GNR-SP 6700 Processors with up to 350W
- 8 B200 SXM6 GPUs with NVSwitch, up to 1000W TDP per GPU, air cooled
- 32 DIMM supporting up to 8TB of System Memory
- 8 E1.S NVMe storage drives, 2 M.2 NVMe
- 8 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHHL
- 2 USB Type-A ports, 1 IPMI GbE LAN port, 2 10G NIC (X710)
- 6 redundant (3+3) 6600W Titanium level power supplies



SYS-822GS-NBRT



Supermicro HGX™ B200 AC Systems with Intel CPU

A streamlined platform from infrastructure to solution lifecycle management



System	Form Factor	CPU	DIMM	Drive	AOCs
SYS-A21GE-NBRT (X13)	10U	Dual 5 th /4 th Xeon <i>Max 64C, 350W, 320MB per CPU</i>	32 slots (Max 5600MTs)	10x Gen5 U.2 (front) + 2x Gen5 M.2 (front)	8x PCIe x16 LP (rear) + 2x PCIe x16 FHHL (front) + 2x 10G RJ45 (rear) + BMC port (rear)
SYS-A22GA-NBRT (X14)	10U	Dual Xeon 6 6900 Series (AP) <i>Max 128C, 500W, 504MB per CPU</i>	24 slots (Max 6400MTs)	10x Gen5 U.2 (front) + 2x Gen5 M.2 (front)	10x PCIe x16 LP (rear) + 2x PCIe x16 FHHL (front) + 2x 10G RJ45 (rear) + BMC port (rear)
SYS-822GS-NBRT (X14) <i>Front IO</i>	8U	Dual Xeon 6 6700 /6500 Series (SP) <i>Max 86C, 350W, 336MB per CPU</i>	32 slots (Max 6400MTs)	8x Gen5 E1.S (front) + 2x Gen5 M.2 (front)	8x PCIe x16 LP (front) + 2x PCIe x16 FHHL (front) + 2x 10G RJ45 (front) + BMC port (front)

NEW



NVIDIA HGX™ B200 AI SuperCluster

256 Nvidia Blackwell GPUs Air-Cooled

Industry Leading AI Performance with Advanced Air Cooling

The SuperCluster provides the core infrastructure elements necessary to scale the NVIDIA Blackwell Platform and deploy the pinnacle of AI training and inference performance. SuperCluster simplifies the complexities of AI infrastructure by providing a fully validated AI cluster with a plug and-play deployment experience.

The SuperCluster creates a massive pool of GPU resources, acting as one AI supercomputer, featuring 1:1 GPU-to-NIC ratio supporting NVIDIA ConnectX®-7 NICs or BlueField®-3 SuperNICs for scaling across high-performance compute fabric.

Management

Supermicro's SuperCloud Composer software provides management tools for monitoring and optimizing air or liquid-cooled infrastructure, delivering a complete solution from proof of concept to full-scale deployment. Manage all data center racks, including compute, storage, networking in one unified dashboard.



Product Specs

- Fully integrated air-cooled 32-node cluster with 256 NVIDIA B200 GPUs
- 46 TB of HBM3e memory space
- Compute Fabric Leaf: 8x NVIDIA Quantum-2 400G InfiniBand Switch or 8x NVIDIA Spectrum-4 400GbE Ethernet Switch
- Compute Fabric Spine: 4x NVIDIA Quantum-2 400G InfiniBand Switch or 4x NVIDIA Spectrum-4 400GbE Ethernet Switch
- Rack: 9x 48U x 750mm x 1295mm
- PDU: 34x 208V 60A 3Ph

SRS-48UAC-10U4N-A1



4U (DLC) NVIDIA HGX™ B200 Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI and Data Analytics

Scale-up x86 platform with up to 15X faster real-time inference performance and 3X faster training for LLMs than Hopper Generation (H100)

Advancing Data Analytics

5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 1.4TB HBM3e Memory, 144 petaFLOPS of AI performance (FP4)

Increased Density (Double)

Liquid cooling enables more computing power to be packed into a smaller space compared to air-cooled systems, maximizing data center utilization

Sustainable Computing

Liquid cooling provides energy efficiency by utilizing coolant thus lowering overall operational cost and reduce the environmental impact of data centers by lowering overall power consumption up to 40% and improves PUE



Product Specs

- Dual Intel® Xeon® 6900 Processors with P-cores, Direct Liquid Cooled / AMD EPYC™ 9005/9004 Series Processors, up to 500W TDP
- 8 B200 SXM6 GPUs with NVSwitch, Direct Liquid Cooled
- 24 DIMM (1DPC Only) supporting up to 6TB 6400MT/s DDR5 RDIMM (Intel/AMD) or 6TB 8800MT/s DDR5 MRDIMM (Intel)
- 8 U.2 NVMe storage drives, 2 M.2 NVMe
- 8 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHHL
- 2 USB 3.0 ports, 1 VGA port, 1 IPMI GbE LAN port, 2 10G NIC (X710)
- 4 Redundant (2+2) 6600W Titanium level power supplies



SYS-422GA-NBRT-LCC / AS-4126GS-NBR-LCC



4U (DLC) NVIDIA HGX™ B200 Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI and Data Analytics

Scale-up x86 platform with up to 15X faster real-time inference performance and 3X faster training for LLMs than Hopper Generation (H100)

Advancing Data Analytics

5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 1.4TB HBM3e Memory, 144 petaFLOPS of AI performance (FP4)

Increased Density

Liquid cooling enables more computing power to be packed into a smaller space compared to air-cooled systems, maximizing data center utilization

Sustainable Computing

Liquid cooling provides energy efficiency by utilizing coolant thus lowering overall operational cost and reduce the environmental impact of data centers by lowering overall power consumption up to 40% and improves PUE



Product Specs

- Dual 5th /4th Gen Intel® Xeon® Scalable processors. Up to 350W TDP, Direct Liquid Cooled
- 8 B200 SXM6 GPUs with NVLink/NVSwitch, up to 1000W TDP per GPU, Direct Liquid Cooled
- 32 DIMM supporting up to 8TB DDR5 RDIMM
- Up to 8 U.2 NVMe storage drives, 2 x M.2 NVMe
- 10 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHHL
- 2 USB 3.0 ports, 1 VGA port, 1x IPMI GbE LAN port, 2x 10G NIC (X710)
- 4x 6600W Redundant (2+2) 6600W Titanium level power supplies



SYS-421GE-NBRT-LCC



4U (DLC) NVIDIA HGX™ B200 Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI & Data Analytics

Scale-up x86 platform with up to 15X faster real-time inference performance and 3X faster training for LLMs than Hopper Generation (H100)

Higher GPU Compute and Memory footprint

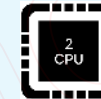
5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 1.4TB HBM3e Memory, 144 petaFLOPS of AI performance (FP4)

Increased Density (double)

Liquid cooling enables more computing power to be packed into a smaller space compared to air-cooled systems, maximizing data center utilization

Sustainable Computing

Liquid cooling provides energy efficiency by utilizing coolant thus lowering overall operational cost and reduce the environmental impact of data centers by lowering overall power consumption up to 30% with improved PUE



Product Specs

- Dual Intel® Xeon® 6700 Processors with P-cores
- 8x B200 SXM6 GPUs with NVSwitch, Direct Liquid Cooled
- 32x DIMM supporting up to 6TB 6400MT/s DDR5 RDIMM
- 8x E1.S front hot-swap drives, 2x M.2 NVMe
- 8x PCIe 5.0 x16 LP, 2x PCIe 5.0 x16 FHHL
- 2x USB ports, 1x VGA port, 1x IPMI GbE LAN port, 2x 10G NIC (X710)
- 4x Redundant (2+2) 6600W Titanium level power supplies



YS-422GS-NBRT-LCC



NVIDIA HGX™ (DLC) B200 AI SuperCluster

256 Nvidia Blackwell GPUs Direct-Liquid-Cooled

Industry Leading AI Performance with Advanced Liquid Cooling

These new SuperCluster offerings powered by the NVIDIA Blackwell Platform are available in 42U, 48U, or 52U configurations. The upgraded cold plates and 250kW coolant distribution unit (CDU) more than double the cooling capacity of the previous generation.

The new vertical coolant distribution manifold (CDM) means that horizontal manifolds no longer occupy valuable rack space. NVIDIA Quantum InfiniBand or NVIDIA Spectrum™ networking in a centralized rack enables a non-blocking, 256-GPU scalable unit in five racks, or an extended 768-GPU scalable unit in nine racks.

Management

Supermicro's SuperCloud Composer software provides management tools for monitoring and optimizing air or liquid-cooled infrastructure, delivering a complete solution from proof of concept to full-scale deployment. Manage all data center racks, including compute, storage, networking in one unified dashboard.



Product Specs

- Fully integrated liquid-cooled 32-node cluster with 256 NVIDIA B200 GPUs
- 46 TB of HBM3e memory space
- Compute Fabric Leaf: 8x NVIDIA Quantum-2 400G InfiniBand Switch or 8x NVIDIA Spectrum-4 400GbE Ethernet Switch
- Compute Fabric Spine: 4x NVIDIA Quantum-2 400G InfiniBand Switch or 4x NVIDIA Spectrum-4 400GbE Ethernet Switch
- Rack / PDU: 5x 48U x 800mm x 1400mm / 18x 415V 60A/100A 3Ph
- Liquid Cooling: 4x Supermicro 250kW CDU with redundant PSU and dual hot-swap pumps

SRS-48UDLC-4U8N-L1



8U NVIDIA HGX™ B300 Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI

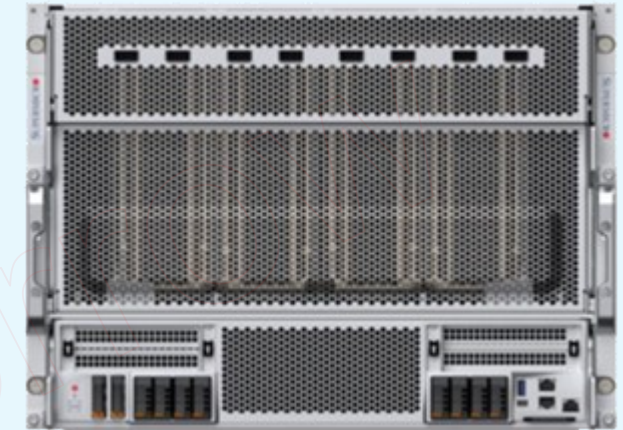
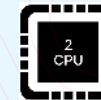
Scale-up x86 platform with up to 11X faster real-time inference performance and 4X faster training for LLMs than Hopper Generation (H100)

Higher GPU Compute and Memory footprint

5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 2.3TB HBM3e (Up to 8TB/s Bandwidth) Memory, 1.5X AI Compute (FP4) FLOPS and 2X attention acceleration than B200 which improves efficiency of reasoning models.

Sustainable Computing

Sustainable computing is realized with efficiency gains as HGX B300 improves energy efficiency by 12X and lowers costs (TCO) by 12X compared to Hopper



Product Specs

- Dual Intel® Xeon® GNR-SP 6700 Series Processors/ AMD EPYC™ 9005/9004 Series Processors, up to 500W TDP
- 8x B300 Blackwell Ultra GPUs with NVSwitch, up to 1100W TDP per SXM module, air cooled
- 32x DIMM (1DPC only) supporting up to 4TB 6400MT/s DDR5 RDIMM
- 8x E1.s NVMe storage drives, 2x M.2 NVMe
- 8x CX-8 Integrated Devices; 2x PCIe 5.0 x16 FHHL
- 2x USB 3.0 ports, 1x mini-DP, 1x VGA port, 1x IPMI GbE LAN port, 2x 10G NIC (x710)
- 6x redundant (3+3) 6.6kW Titanium level power supplies



SYS-822GS-NB3RT
AS-8126GS-NB3RT



4U NVIDIA HGX™ B300 DLC Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI

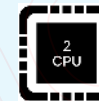
Scale-up x86 platform with up to 11X faster real-time inference performance and 4X faster training for LLMs than Hopper Generation (H100)

Higher GPU Compute and Memory footprint

5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 2.3TB HBM3e (Up to 8TB/s Bandwidth) Memory, 1.5X AI Inference Compute (FP4) and 2X attention acceleration for reasoning workload than B200

Sustainable Computing

Sustainable computing is realized with efficiency gains as HGX B300 improves energy efficiency by 12X and lowers costs (TCO) by 12X compared to Hopper



Product Specs

- Dual Intel® Xeon® GNR-SP 6700 Series Processors
- 8x B300 Blackwell Ultra GPUs with NVSwitch, up to 1100W TDP per SXM module, Liquid cooled
- 32x DIMM (1DPC only) supporting up to 4TB 6400MT/s DDR5 RDIMM
- 8x E1.s NVMe storage drives, 2x M.2 NVMe
- 8x CX-8 Onboard Devices; 2x PCIe 5.0 x16 FHHL Slots for N-S
- 2x USB 3.0 ports, 1x mini-DP, 1x VGA port, 1x IPMI GbE LAN port, 2x 10G NIC (x710)
- 4x redundant (2+2) 6.6kW Titanium level power supplies



SYS-422GS-NB3RT



2-OU NVIDIA HGX™ B300 DLS Server

Purpose Built for Large Scale AI and HPC Workload

Accelerated Generative AI

Scale-up x86 platform with up to 11X faster real-time inference performance and 4X faster training for LLMs than Hopper Generation (H100)

Higher GPU Compute and Memory footprint

5th Gen NVLink with 1.8TB/s of GPU-GPU Bandwidth, 2.3TB HBM3e (Up to 8TB/s Bandwidth) Memory, 1.5X AI Compute (FP4) FLOPS and 2X attention acceleration than B200 which improves efficiency of reasoning models

Sustainable Computing

Sustainable computing is realized with efficiency gains as HGX B300 improves energy efficiency by 12X and lowers costs (TCO) by 12X compared to Hopper

Rack Infrastructure

L11 with 9x B300 servers per ORV3 44-OU Rack with In-Rack-CDU
L11 with 18x B300 servers per ORV3 48-OU Rack with In-Row-CDU



Product Specs

- Dual Intel® Xeon® 6700 Series Processors, Direct Liquid Cooled
- 8x B300 Blackwell Ultra GPUs with NVSwitch, up to 1100W TDP per SXM module, Direct Liquid Cooled
- 32x DIMM (1DPC only) supporting up to 8TB 6400MT/s DDR5 RDIMM
- 8x E1.s NVMe storage drives, 2x M.2 NVMe
- 8x CX-8 Onboard Devices; 2x PCIe 5.0 x16 FHHL Slots for N-S
- 1x DCSCM, 1x USB 3.0, 1x USB 2.0, 1x VGA port, 1x IPMI GbE LAN port, 2x 10G NIC
- 21" OCP ORV3 1400A Busbar w/ power clip

SYS-222GS-NB30T-ALC



Supermicro's NVIDIA RTX Pro 6000 BSE

Solution Overview - PCIe Optimized Configuration

2U Air-Cooled / SYS-212GB-(F)NR 2/4-GPU	2U Air-Cooled / SYS-222C-TN 2-GPU	3U Air-Cooled / SYS-322GA-NR 8-GPU	4U Air-Cooled / SYS-422GA-NRT 8-GPU	4U Air-Cooled / MGX SYS-422GL-NR 4/8-GPU	5U Air-Cooled / SYS-522GA-NRT or AS -5126GS-TNRT2 8-GPU
1-4-3 1-2-3  	2-2-3  	2-4-3  	2-8-5  	2-8-5 2-4-3  	2-8-5   Intel AMD
2U edge reference architecture design for NVIDIA & perfect for RTX PRO 6000 BSE • 1 (Single) Intel Xeon 6700 series CPU with P-cores • 2 GPU PCIe (Up to 4-GPU) • 3 NICs: 1 BF3 and 2 CX-7 • CPU: Intel Xeon 6700 (6761P, 64C, 2.5 GHz) • Mem: 16 DIMM slots max up to 2TB 5200MT/s ECC DDR5 RDIMM • Storage: 4 front hot-swap E1.S NVMe drive bays and 2 M.2 PCIe 5.0 x2 NVMe slots (M-key 2280/22110)	2U CloudDC reference architecture design for NVIDIA and perfect for RTX PRO 6000 BSE • 2 (Dual) Intel Xeon 6700 series CPUs • 2 GPU PCIe per system (Up to 2-GPU) • 3 NICs: 1 BF3 and 2 CX-7 • CPU: Intel Xeon 6700P (2x 6737, 32C, 2.9 GHz) • Mem: 512GB • Storage: 2x1TB (M.2 boot), 4TB local SSD	3U edge reference architecture design for NVIDIA and perfect for RTX PRO 6000 BSE • 2 (Dual) Intel Xeon 6900 series CPUs with P-cores • 8 GPU PCIe 8 double-width or up to 19 single-width) • 3 NICs: 1 BF3 and 2 CX-7 • CPU: Intel Xeon 6900 (2x 6960P, 72C, 2.75 GHz) • Mem: 24 DIMM slots up to 6 TB 8800MT/s ECC DDR5 MRDIMM • Storage: Options for front hot-swap E1.S PCIe 5.0 x4 NVMe and hot-swap 2.5" PCIe 5.0 x4 NVMe. Also support for 2 M.2 PCIe 5.0 x2 NVMe (M-key)	4U edge reference architecture design for NVIDIA and perfect for RTX PRO 6000 BSE • 2 (Dual) Intel Xeon 6900 series CPUs with P-cores • 8 GPU PCIe 8 double-width or (2 quad-width, 2 triple-width or 10 single-width) • 5 NICs: 1 5 BF3 or 1 BF3 and 4 CX-7 • CPU: Intel Xeon 6900 (2x 6960P, 72C, 2.75 GHz) • Mem: 24 DIMM slots max up to 6TB 8800MT/s ECC DDR5 MRDIMM • Storage: 8 front hot-swap E3.S 1T PCIe 5.0 x4 NVMe and 2 M.2 NVMe slots (M-key)	4U MGX reference architecture design for RTX PRO 6000 BSE architecture with air-cooling. • 2 (Dual) Intel Xeon 6900 series CPUs • 4 or 8 GPU PCIe per system • 3 or 5 NICs: 1 BF3 and 2 or 4 CX-7 • CPU: Intel Xeon 6900 (2x 6960P, 72C, 2.75 GHz) • Mem: 1TB • Storage: 2x1TB (M.2 boot), 4TB local SSD	5U GPU Server reference architecture design for RTX PRO 6000 BSE architecture with air-cooling. • 2 (Dual) Intel Xeon 6th / AMD EPYC™ 4th CPUs • 8 GPU PCIe per system (Up to 8-GPU) • 5 NICs: 1 BF3 and 4 CX-7 • CPU: Intel Xeon 6900 (2x 6960P, 72C, 2.75 GHz) • Mem: 1TB • Storage: 2x1TB (M.2 boot), 4TB local SSD

5U 10 PCIe GPU Server

Built for a Variety of AI and HPC Workloads

Unparalleled GPU Performance

Supports up to 10 double-width GPUs, optimized for NVIDIA H100NVL, delivering exceptional computing power for AI, deep learning, and HPC applications

Cutting-Edge Processing Power

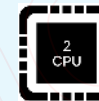
Equipped with dual 5th/4th Gen Intel® Xeon® Scalable processors, offering up to 64 cores per CPU for lightning-fast processing speeds and unmatched multi-threading capabilities

Massive Memory Capacity

Features 32 DIMM slots supporting up to 8TB of DDR5 ECC memory, enabling seamless management of complex computations and data-intensive workflows

Flexible Storage Options

Offers a variety of hot-swap drive configurations, including up to 24 2.5" NVMe/SAS/SATA bays, providing ample storage for data-intensive applications and future scalability



Product Specs

- Dual 5th/4th Gen Intel® Xeon® Scalable processors/ AMD EPYC™ 9005/9004 Series Processors, up to 500W TDP
- Up to 10 GPUs, air cooled
- 32 DDR5 DIMMs supporting up to 8TB (Intel)/ 24 DIMM (1DPC only) supporting up to 6TB 6400MT/s DDR5 RDIMM (AMD)
- 16 U.2 NVMe is optional. Default is 8 U.2 NVMe, 8 2.5" SATA and 2 M.2 NVMe
- 13 PCIe slots and 1 AIOM slot
- 4 redundant (N+N) 2700W Titanium level power supplies



SYS-521GE-TNRT / AS -5126GS-TNRT



5U 10 PCIe GPU Server

Built for a Variety of AI and HPC Workloads

Unparalleled GPU Performance

Supports up to 10 double-width GPUs, optimized for NVIDIA H200 NVL/L40S, delivering exceptional computing power for AI, deep learning, and HPC applications

Cutting-Edge Processing Power

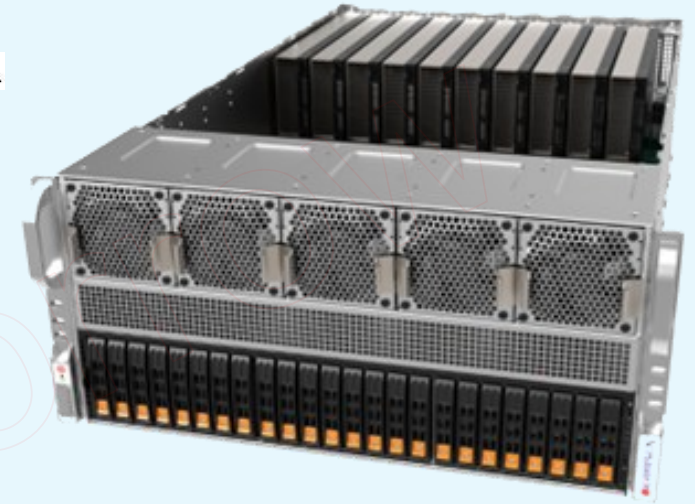
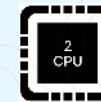
Equipped with dual Intel®6 Xeon® Scalable processors, offering up to 128 cores per CPU for lightning-fast processing speeds and unmatched multi-threading capabilities

Massive Memory Capacity

Features 24 DIMM slots supporting up to 6TB of DDR5 ECC memory, enabling seamless management of complex computations and data-intensive workflows

Flexible Storage Options

Offers a variety of hot-swap drive configurations, including up to 24 2.5" NVMe bays, providing ample storage for data-intensive applications and future scalability



Product Specs

- Dual Intel®6 Xeon® Scalable processors p-cores
- Up to 10 GPUs, air cooled
- 24 DDR5 DIMMs supporting up to 6TB or DDR5 MRDIMM
- Up to 24 U.2/U.3 NVMe and 2 M.2 NVMe
- 13 PCIe slots and 1 AIOM/OCP 3.0 slot
- 6 redundant (N+N) 2700W Titanium level power supplies



SYS-522GA-NRT



4U 13 PCIe GPU Server

4U height, Rear I/O, Optimized for Intensive Compute or AI Workloads

Dual processor 4U GPU Server

Designed for applications and use cases requiring high performance processing and high-density GPUs

CPU and Memory

Supports dual Intel® Xeon® 6700/6900 series processors. This system supports 24 DIMMs up to 6TB 6400MT/s DDR5 RDIMM or 6TB 8800MT/s DDR5 .

Multiple storage configuration options

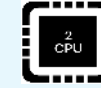
Has 8 drive bays. It also has 2 M.2 NVMe slots.

Multiple PCIe Expansion options

Supports up to 13 PCIe 5.0 x16 FHFL slots.

Multiple GPU configurations

Supports up to 2 quad-width or 2 triple-width or 8 double-width or 10 single-width GPUs
Support 8x RTX PRO 6000 BSE (600W) at 35C + 5 NICs



Product Specs

- Dual socket Intel® Xeon® 6900 /6700/6500 series processors
- 24 DIMMs supporting up to 6TB 6400MT/s DDR5 RDIMM or 6TB 8800MT/s DDR5
- Multiple storage options for front hot-swappable NVMe drive bays
 - Supports 3 options for up to 8 total bays
 - It also support up to 2 M.2 NVMe slots (M-key)
- High-capacity PCIe expansion slots
 - Supports up to 13 PCIe 5.0 x16 FHFL
- Multiple GPU configurations
 - 4U system supports up to 2 quad-width or 2 triple-width or 8 double-width or 10 single-width GPUs
- Multiple networking options



SYS-422GA-NRT



4U 10 PCIe GPU Server

Built for a Variety of AI and HPC Workloads

Unparalleled GPU Performance

Supports up to 10 double-width GPUs, optimized for NVIDIA L40S, delivering exceptional computing power for AI, deep learning, and HPC applications

Cutting-Edge Processing Power

Equipped with dual 5th/4th Gen Intel® Xeon® Scalable processors, offering up to 64 cores per CPU for lightning-fast processing speeds and unmatched multi-threading capabilities

Massive Memory Capacity

Features 32 DIMM slots supporting up to 8TB of DDR5 ECC memory, enabling seamless management of complex computations and data-intensive workflows

Flexible Storage Options

Offers a variety of hot-swap drive configurations, including up to 24 2.5" NVMe/SAS/SATA bays, providing ample storage for data-intensive applications and future scalability



Product Specs

- Dual 5th/4th Gen Intel® Xeon® Scalable processors / AMD EPYC™ 9004/9005 Series Processors, up to 400W TDP
- Up to 10 GPUs, air cooled
- 32 DDR5 DIMMs supporting up to 8TB(Intel)/ 24 DIMM (1DPC only) supporting up to 6TB 6400MT/s DDR5 RDIMM (AMD)
- 16 U.2 NVMe is optional. Default is 8 U.2 NVMe, 8 2.5" SATA and 2 M.2 NVMe
- 13 PCIe slots and 1 AIOM slot
- 4 redundant (N+N) 2700W Titanium level power supplies



SYS-421GE-TNRT / AS -4125GS-TNRT



4U 8 PCIe GPU Server

Built for a Variety of AI and HPC Workloads Needing Low Latency

Unparalleled GPU Performance

Supports up to 8 direct connect double-wide GPUs, optimized for NVIDIA L40S, delivering exceptional computing power for AI, deep learning, and HPC applications

Cutting-Edge Processing Power

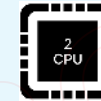
Equipped with dual 5th/4th Gen Intel® Xeon® Scalable processors, offering up to 64 cores per CPU for lightning-fast processing speeds and unmatched multi-threading capabilities

Massive Memory Capacity

Features 32 DIMM slots supporting up to 8TB of DDR5 ECC memory, enabling seamless management of complex computations and data-intensive workflows

Flexible Storage Options

Offers a variety of hot-swap drive configurations, including up to 12 2.5" NVMe/SAS/SATA bays, providing ample storage for data-intensive applications and future scalability



Product Specs

- Dual 5th/4th Gen Intel® Xeon® Scalable processors
- Up to 8 GPUs directly connected to CPU, air cooled
- 32 DDR5 DIMMs supporting up to 8TB
- 4 2.5" NVMe + 8 2.5" SATA + 2 M.2 NVMe
- 8 PCIe slots and 1 AIOM slot
- 4 redundant (N+N) 2700W Titanium level power supplies



SYS-421GE-TNRT3



3U 10 PCIe GPU Server

Dual Processor Edge Data Center System, Optimized for Intensive Compute or AI Workloads

Dual processor 3U GPU Server

Designed for applications and use cases requiring high performance processing and high-density GPUs

CPU and Memory

Supports dual socket Intel® Xeon® 6900/6700 series processors. This system supports 24 DIMMs up to 6TB 6400MT/s DDR5 RDIMM or 6TB 8800MT/s DDR5 .

Multiple storage configuration options

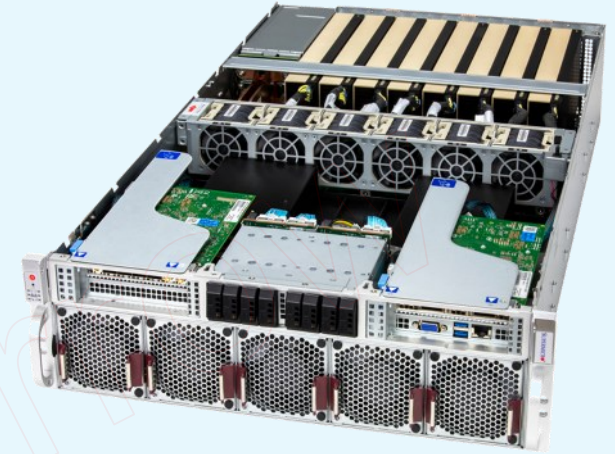
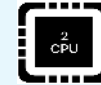
Supports configurations for 2, 4, 6, or 14 total bays. It also supports 2 M.2 NVMe slots.

Multiple PCIe Expansion options

Supports up to 10 PCIe 5.0 x16 FHFL or 20 PCIe 5.0 x8 FHFL slots

Multiple GPU configurations

Supports up to 8 double-width or 19 single-width GPUs



Product Specs

- Dual socket Intel® Xeon® 6900 /6700 series processors
- 24 DIMMs supporting up to 6TB 6400MT/s DDR5 RDIMM or 6TB 8800MT/s DDR5
- Multiple storage options for front hot-swappable NVMe drive bays
 - Supports configurations for 2, 4, 6, or 14 total bays
 - It also support up to 2 M.2 NVMe slots (M-key)
- High-capacity PCIe expansion slots
 - Supports up to 10 PCIe 5.0 x16 FHFL or 20 PCIe 5.0 x8 FHFL slots
- Multiple GPU configurations
 - Supports configurations up to 8 double-width or 19 single-width GPUs
- Multiple networking options



SYS-322GA-NR



2U 4 PCIe GPU Server

Single Processor Edge Data Center System

CPU

Single socket Intel® Xeon® 6700 series processor with P-cores

Memory

16 DIMMs supporting multiple memory configurations from 512GB up to 2TB

Storage

Supports 4 front hot-swap E1.S NVMe drive bays and up to 2 M.2 PCIe 5.0x2 NVMe slots

Multiple PCIe Expansion options

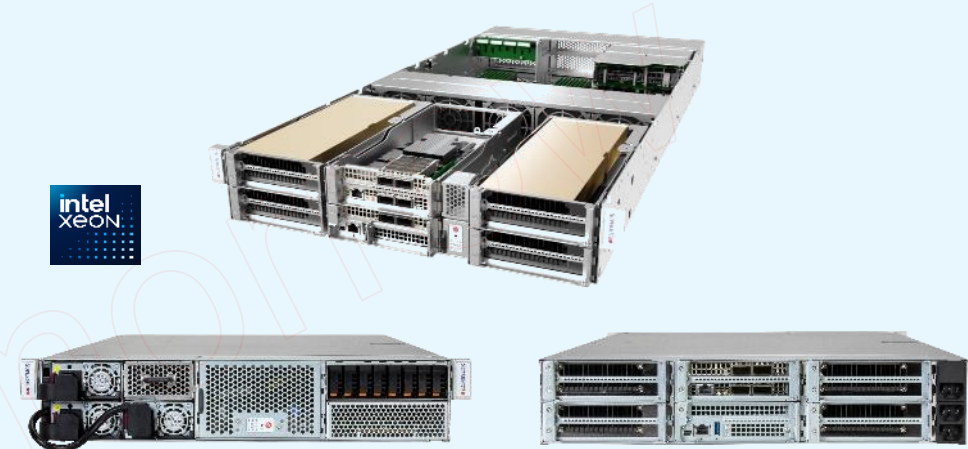
Supports up to 7 PCIe 5.0 x16 FHFL slots

I/O Access options

Rear I/O for SYS-212GB-NR. Front I/O for SYS-212GB-FNR

Multiple GPU configurations

Up to 4 double-width GPUs. With many options based on specific GPU combinations



Product Specs

- Single socket Intel® Xeon® 6700 series processor with P-cores
- 16 DIMMs supporting up to 1TB 6400 MT/s or 2TB 5200 ECC DDR5 RDIMM or up to 512GB 8000 MT/s ECC DDR5 MRDIMM
- Storage for 4 front hot-swap E1.S NVMe drive bays. Also supports up to 2 M.2 PCIe 5.0x2 NVMe slots (M-key 2280/22110)
- PCIe expansion slots, up to 7 PCIe 5.0 x16 FHFL slots
- Medium density GPU configurations
 - SYS-212GB-NR supports up to 2 H200 NVL / RTX Pro 6000 BSE or up to 4 RTX Pro 6000 Max-Q WS / RTX Pro 5000 WE or up to 5 RTX Pro 4500 WE / RTX Pro 2000 WE or up to 6 RTX Pro 4000 WE
 - SYS-212GB-FNR supports up to 4 H200 / RTX Pro 6000 BSE or up to 6 RTX Pro 4500 BSE



SYS-212GB-NR



SYS-212GB-FNR



4U 8 PCIe MGX GPU Server

AI Acceleration for Mainstream Enterprise Servers

Extreme Performance

Supports up to 8 double-width GPUs, including NVIDIA H200NVL, delivering exceptional computing power for AI, deep learning, and HPC applications

Cutting-Edge Processing Power

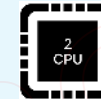
Equipped with latest generation Intel 6900 series processors, up to 500W TDP air cooled for lightning-fast processing speeds and unmatched multi-threading capabilities

Massive Memory Capacity

Features 24 DIMM slots supporting up to 3TB of DDR5 ECC memory, enabling seamless management of complex computations and data-intensive workflows

Latest Generation Storage

Offers the latest technology in storage with up to 8 E1.s drive bays + 2 M.2, providing ample storage for data-intensive applications and future scalability



Product Specs

- Dual Intel® Xeon® 6900 series processors
- Up to 8 GPUs, air cooled
- 24 DDR5 DIMMs supporting up to 3TB
- 8 hot-swap E1.S drive bays + 2 M.2 NVMe
- 13 PCIe slots; 8 double wide + 4 FHHL + 1 LP
- 4 redundant (2+2) 3200W Titanium level power supplies



SYS-422GL-NR



Diversity Products beyond AI GPUs



1U with Grace™ Hopper



1U with Grace Hopper LC



1U 2-Node with Grace Hopper



1U 2-Node with Grace CPU



2U with Grace CPU



2U with x86 DP

Model	ARS-111GL-NHR	ARS-111GL-NHR-LCC	ARS-111GL-DNHR-LCC	ARS-121L-DNR	ARS-221GL-NR	SYS-221GE-NR
CPU	72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip	72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip	72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip per node	144-core Grace Arm Neoverse V2 CPU in a single chip per node (total of 288 cores in one system)	144-core Grace Arm Neoverse V2 CPU in a single chip	4th or 5th Generation Intel® Xeon® Scalable processors
Cooling	Air-cooled	Liquid-cooled	Liquid-cooled	Air-cooled	Air-cooled	Air-cooled
GPU Support	NVIDIA H100 Tensor Core GPU with 96GB of HBM3	NVIDIA H100 Tensor Core GPU with 96GB of HBM3	NVIDIA H100 Tensor Core GPU with 96GB of HBM3 per node	Please contact for possible configurations	Up to 4 double-width GPUs including NVIDIA H100 PCIe, H100 NVL, L40S.	Up to 4 double-width GPUs including NVIDIA H100 PCIe, H100 NVL, L40S
Memory	CPU: 480G integrated LPDDR5X with ECC GPU: 96GB HBM3	CPU: 480G integrated LPDDR5X with ECC GPU: 96GB HBM3	CPU: 480G integrated LPDDR5X with ECC per node GPU: 96GB HBM3 per node	Up to 480GB of integrated LPDDR5X with ECC and up to 1TB/s of memory bandwidth per node	Up to 480GB of integrated LPDDR5X with ECC and up to 1TB/s of memory bandwidth per node	Up to 4TB, 32 DIMM slots, ECC DDR5-4800 DIMM
Networking	3 PCIe 5.0 x16 slots supporting NVIDIA BlueField®-3 or ConnectX®-7	3 PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7	2 PCIe 5.0 x16 slots per node, supporting NVIDIA BlueField-3 or ConnectX-7	3 PCIe 5.0 x16 slots per node, supporting NVIDIA BlueField-3 or ConnectX-7	3 PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7	3 PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7
Storage	8 Hot-swap E1.S drives and 2 M.2 NVMe drives	8 Hot-swap E1.S drives and 2 M.2 NVMe drives	4 Hot-swap E1.S drives and 2 M.2 NVMe drives per node	4 Hot-swap E1.S drives and 2 M.2 NVMe drives per node	8 Hot-swap E1.S drives and 2 M.2 NVMe drives	8 Hot-swap E1.S drives and 2 M.2 NVMe drives
Power Supplies	2 2000W Titanium Level	2 2000W Titanium Level	2 2700W Titanium Level	2 2000W Titanium Level	3 2000W Titanium Level	3 2000W Titanium Level

1U Grace™ Hopper Superchip System

NVIDIA GH200 Grace Hopper Superchip System Supporting NVIDIA BlueField®-3 or NVIDIA ConnectX®-7

Breakthrough Performance with no Bottle Necks

High speed NVL Interconnect between processor and GPU removes increases performance up to 55X versus x86 and H100 configurations.

Best Value for shared memory footprint

Configurations with up to 576GB of shared memory are a fraction of the cost of previous configurations.

Grace GPU Provides Power Efficiency and Performance

Grace low TDP and low latency GPUs excel at idle time energy savings delivering performance and large cost savings in data center deployments



ARS-111GL-NHR



Product Specs

- 1U systems with NVIDIA GH200 Grace Hopper Superchip
- 72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip
- NVLink chip to chip (C2C) high bandwidth, low latency interconnect between CPU and GPU at 900GB/s
- Onboard memory:
 - CPU: 480GB LPDDR5X
 - GPU: 96GB HBM3 for LLM
- 8 Hot-swap E1.S drives and 2 M.2NVMe drives
- Expansion:
 - Up to 2 PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7
- Power Supply
- Liquid-cooled and air-cooled systems available



MGX™ 1U 2Node Grace System Liquid Cooled

Dual NVIDIA Superchip on a Single Node for Inference Workloads

Increased density and sustainable computing

Liquid cooling enables more computing power to be packed into a smaller space maximizing data center utilization

Maximum Density Thread and Core Counts

288 cores and 576 threads in a 1U form factor at under 2000W TDP

Key Applications

Apache Spark, Web based low latency applications, HPC for NON-accelerated application performance and high-performance simulation of complex 3D graphics, and 1U Compact design with advanced liquid cooling solutions



ARS-111GL-DNHR-LCC



Product Specs

- 1U systems with NVIDIA GH200 Grace Hopper Superchip
- 72-core Grace™ Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip
- Onboard memory:
 - CPU: up to 960GB LPDDR5X
 - GPU: 96GB HBM3
 - 8 Hot-swap E1.S drives and 2 M.2
 - NVMe drives
- Expansion – 4 PCIe slots
 - Up to 3 PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7
 - 1 PCIe 5.0 x16 (physical slot x16, actual support Gen5 x8 signal only.)
- Power Supply – 2000W (2+2)
- Thermal and Cooling: Liquid-cooled and air-cooled systems available



Next Gen 2U 1Node GPU System Air Cooled

Dual NVIDIA GH200 Superchip on a Single Node for Inference Workloads

Ultimate inference workhorse in an Air-Cooled High density form factor

5th Generation NVLINK connecting 2 H200 GPUs and 2 Grace™ Processors in a shared memory footprint.

Higher GPU compute and memory footprint

Dual H200 superchips and Grace processors connected to provide maximum inference compute & 1104GB of shared memory with breakthrough performance on industry inference benchmarks

Key Applications

AI Compute/Model Training/Deep Learning, AI Inferencing for giant-scale DLRM (Deep Learning Recommendation Models), LLM



ARS-221GL-SR



Product Specs

- CPU – Dual socket dual GH200 Superchips (up to 2000W TDP)
- Drives – 8 Hot-Swap Bays
 - 3 E1.S
 - 2 M.2
- I/O ports (per Server)
 - 1 RJ45 1GbE IPMI
 - 1 mini-DP (needed converter to VGA), 1 USB 3.0 port
- Memory – On Board
 - up to 960 GB LPDDR5X + 288GB HBM3e (Depending on Dual GH200)
- Expansion – 4 PCIe Slot
 - 3 PCIe 5.0 x16
 - 1 PCIe 5.0 x16 (physical slot x16, actual support Gen 5 x8 signal only)
- Power Supply – 2000W (2+2)



1U Grace™ Blackwell 200 4-way NVLink

NVIDIA GB200 Grace Blackwell Superchip Supporting NVIDIA BlueField®-3 DPU and NVIDIA ConnectX®-8 SuperNIC

Purpose-built GPU-accelerated HPC machine

High speed NVLink5 interconnect among the 4 Blackwell 200 GPUs provides unparalleled price to performance for High Performance Computing workloads

Super dense DLC rack

A single 48U rack can hold up to 32 nodes (128 GPUs) with in-row CDU and 26 nodes (104 GPUs) with in-rack CDU

Superior Networking Flexibility

Options to maximize network bandwidth at 800G to each Blackwell GPU or optimize to any PCIe 5.0 x16 CEM NICs available in the market.



ARS-121GL-NB2B-LCC



Product Specs

- 1U NVIDIA GB200 NVL4
- Dual 72-core Grace Arm Neoverse V2 CPUs with CLINK CPU-CPU 300GB/s bidirectional
- Quad B200 1200W TDP with NVLink GPU-GPU 600GB/s bidirectional
- NVLink chip to chip (C2C) high bandwidth, low latency interconnect between CPU and GPU at 900GB/s
- Onboard memory:
 - CPU: 480GB LPDDR5X each
 - GPU: 186GB HBM3e each
- 4-8 Hot-swap E1.S drives and 1 M.2 NVMe drives
- Expansion:
 - 4 PCIe 5.0 x16 FHHL slots (OR) 4-port 800G ConnectX-8 + 2-port 200G BlueField-3
- Liquid-cooled with 48-54V DC busbar



Reference Rack GB200 NVL4

48U Rack

Management Networking

- 2x NVIDIA Management Switches

Power Shelves

- 5x Power Shelves

32 Compute Trays

- 4x NVIDIA Blackwell GPUs per tray

Power Shelves

- 5x Power Shelves

128 GPUs In-row CDU



~194kW, Liquid Cooled Busbar
SRS-NVL4-M0-LB-STD

128 x 800G



64 x 200G



Network-Switch Rack

104 x 800G



52 x 200G



104 GPUs In-rack CDU



~159kW, Liquid Cooled Busbar
SRS-NVL4-M1-LB-STD

Management Networking

- 2x NVIDIA Management Switches

Power Shelves

- 4x Power Shelves

26 Compute Trays

- 4x NVIDIA Blackwell GPUs per tray

Power Shelves

- 4x Power Shelves

Liquid Cooling

- Supermicro 250kW capacity coolant distribution unit (CDU)

NVIDIA GB200 NVL72 System

Compute for Trillion Parameter Scale AI Workload

SRS-GB200-NVL72-M1



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 interconnected at 1.8TB/s

8 Compute Trays

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

Power Shelves

- 8x 1U 33kW (6x 5.5kW PSUs)

Liquid Cooling

- Supermicro 250kW capacity cooling distribution unit (CDU) with redundant PSU and dual hot-swap pumps



Trillion Parameter AI

Scale-up ARM based platform with up to 30X faster real-time inference performance and 4X faster training for LLMs than Hopper Generation (H100)

Highest GPU Compute and Memory footprint

Rack-scale solution that boasts a 72-GPU NVLink domain that acts as a single massive GPU, 14TB HBM3e Memory, 1.4 exaFLOPS of AI performance (FP4)

Hyper Dense

72 Blackwell GPUs and 36 Grace CPUs in a liquid cooled rack (125kW per Rack/1200W per GPU) packs massive computing power and maximizes data center space utilization

Sustainable Computing

Sustainable computing is realized with efficiency gains as GB200 NVL72 improves energy efficiency by 25X and lowers costs (TCO) by 25X compared to the Hopper

Product Specs

- Per Compute Tray Specs:
 - CPU 2 72-core Grace Arm Neoverse V2
 - CPU Memory Up to 480GB LPDDR5X per Superchip (960GB per tray)
 - GPU 4 Blackwell B200
 - GPU Memory Up to 384GB HBM3e per Superchip (768GB per tray)
 - Up to 8 E1.S PCIe 5.0 drives
 - NVLink-C2C, 900 GB/s CPU-to-GPU interconnect
 - Fifth-Generation NVIDIA NVLink 1.8 TB/s GPU-to-GPU interconnect

NVIDIA GB300 NVL72 System

Node

The NVIDIA GB300 NVL72 delivers exascale computing capabilities in a single rack with fully integrated Liquid-Cooling. It incorporates 72 NVIDIA Blackwell GPUs and 36 Grace CPUs.



Rack

Configuration

72 NVIDIA Blackwell Ultra GPUs, 36 NVIDIA Grace CPUs

Fast Memory

Up to 40 TB

CPU and GPU

2x 72-core NVIDIA Grace Arm Neoverse V2 CPU
4x NVIDIA Blackwell B300 per Superchip

GPU Memory | Bandwidth

Up to 21 TB | Up to 576 TB/s

CPU Memory | Bandwidth

Up to 18 TB SOCAMM with LPDDR5X | Up to 14.3 TB/s

Networking

NVIDIA ConnectX®-8 NICs
Up to 2 NVIDIA BlueField®-3 DPUs

Storage

Up to 8x E1.S PCIe 5.0 drives per Compute tray

Power Supply

Shared power through 4+4 rack power shelves | Rack TDP: ~136 kW



Management Networking

- In-band management switch
- Out-of-band management switch

10 Compute Trays

- 4x NVIDIA Blackwell Ultra GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Compute Interconnect

- 9x NVLink Switches

8 Compute Trays

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

Liquid Cooling

- Supermicro 250kW capacity coolant distribution unit (CDU) with redundant PSU and dual hot-swap pumps

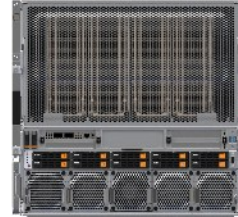
Supermicro's NVIDIA Blackwell Solutions Overview

4U Liquid-Cooled DLC-2 NVIDIA HGX B200 System



- Direct-to-Chip Liquid Cooling
- Up to 40% More AI per Watt
- While lowering TCO by up to 20%

10U Air-Cooled NVIDIA HGX B200 System



- FP4 Tensor Core 144 PFLOPS | 72 PFLOPS
- Total Memory Up to 1.4 TB
- Networking Bandwidth 0.8 TB/s

NVIDIA GB200 NVL72

- GPU Memory Size
186GB
- GPU Max TDP
1200 W
- Rack Power
125-132 kW
- E/W Networking
400G CX7 per GPU

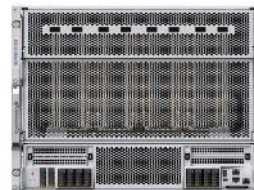


4U Liquid-Cooled DLC-2 NVIDIA HGX B300 System



- Direct-to-Chip Liquid Cooling
- Up to 40% More AI per Watt
- While lowering TCO by up to 20%

8U Air-Cooled NVIDIA HGX B300 System



- FP4 Tensor Core 144 PFLOPS | 72 PFLOPS
- Total Memory Up to 2.3 TB
- Networking Bandwidth 1.6 TB/s

NVIDIA GB300 NVL72

- GPU Memory Size
Up to 288GB
Offering 50% More Memory
- GPU Max TDP
1400 W
- Rack Power
132-140 kW
- E/W Networking
800G CX8 per GPU



Wide Range of kW Compatibility

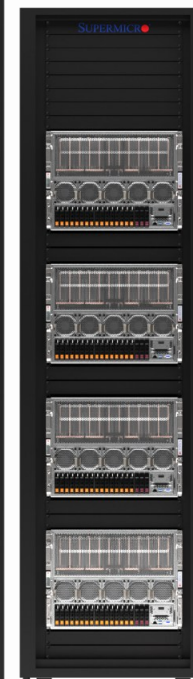
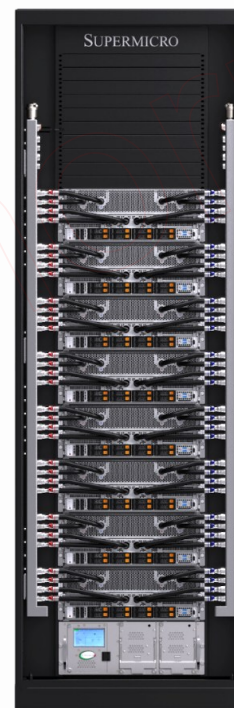
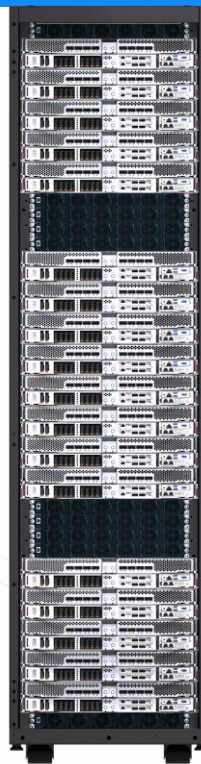
VR NVL72

GBx00

HGX R200

HGX B200/B300

HGX H200



kW Per Rack	~180-220 kW	125-136 kW	~192-384 kW	250 kW	170 kW	120 kW	60 kW	80 kW	50 kW
GPU Per Rack	72x R200	72x B200/300	64-128 R200	144x B300	96x B200/300	64x B200	32x B200	64x H200	32x H200
Rack Height	48U	48U	48U	48OU/52U	52U	48U	48U	48U	42U
Rack Weight	~4060 lbs	~3480 lbs	TBD	~4740 lbs	~5000 lbs	~3800 lbs	~1870 lbs	~2100 lbs	~1400 lbs

GPU AI Supercluster

HGX™ (Air) Supercluster



- 32 8U HGX™ Servers
- 256 H200 GPUs
- 60kW Compute Rack Power

HGX™ (LC) Supercluster



- 64 4U HGX™ Servers
- 512 H200 GPUs
- 100kW Compute Rack Power
- Up to 40% lower power consumption

GB200 Supercluster



- 36 Grace™ CPUs per Rack
- 72 Blackwell GPUs per Rack
- NVLink® Compute Fabric
- 120kW Rack Power
- NVIDIA branded AI LC Rack

AMD GPU Server Product Family

OAM Platforms (AIR-COOLED)



OAM Platforms (LIQUID-COOLED)



Unlock AI and HPC Excellence: Supermicro Servers with AMD Instinct MI355X/MI350X/MI325X - Accelerating Innovation at Scale

- 8U & 10U Systems supporting 8 MI325X or 8 MI350X or 8 MI355X Instinct GPUs (AIR-COOLED)
 - Dual EPYC™ Turin Socket SP5 processors, up to 500W
 - 8 NIC for GPU direct RDMA (1:1 GPU Ratio)
 - Redundant Titanium level power supplies
- 4U Systems supporting 8 MI325X or 8 MI355X Instinct GPUs (LIQUID-COOLED)
 - Dual EPYC™ Turin Socket SP5 processors, up to 500W
 - 8 NIC for GPU direct RDMA (1:1 GPU Ratio)
 - High-Density 4U system supporting 8 Instinct GPUs

8U AMD OAM GPU (MI325X/MI350X) Server

Purpose Built for Large Scale AI and HPC Workload

High performance AI/Deep Learning and HPC-optimized systems

Dual AMD EPYC™ 9005 Series Processors including CPUs with AMD 3D V-Cache™ technology

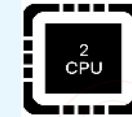
Double the CPU to GPU throughput with PCIe 5.0

Powered by AMD Instinct™ MI325X/MI350X 8-GPUs

The H14 GPU-optimized servers deliver unprecedented acceleration at every scale to power the world's highest performing data centers for AI, data analytics, and HPC applications

Air-cooled

AMD
EPYC



Product Specs

- Processors - Dual AMD EPYC™ 9005/9004 series processors, up to 500W
- GPU - High density 8U system with AMD Instinct™ MI325X/MI350X 8-GPU
- Memory - 24 DIMM slots, DDR5-6400 (1DPC); up to 6TB
- Storage drive - Up to Total 18 Drive Bays, 2 M.2
- PCIe Slots - 8 PCIe 5.0 x16 LP slots; Up to 4 PCIe 5.0 x16 FHFL slots
- I/O - 2 USB 3.0 ports, 1 VGA port, 1 BMC GbE LAN port
- Remote Management - ASPEED AST2600
- Power Supply - up to 8 redundant 5250W Titanium level power supplies



AS-8126GS-TNMR



4U AMD OAM GPU (MI325X/MI355X) Server

Purpose Built for Large Scale AI and HPC Workload

High performance AI/Deep Learning and HPC-optimized systems

Dual AMD EPYC™ 9005 Series Processors including CPUs with AMD 3D V-Cache™ technology

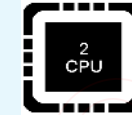
Double the CPU to GPU throughput with PCIe 5.0

Powered by AMD Instinct™ MI325X/MI355X 8-GPUs

The H14 GPU-optimized servers deliver unprecedented acceleration at every scale to power the world's highest performing data centers for AI, data analytics, and HPC applications

Liquid-cooled for highest density, rack solution

AMD
EPYC



Product Specs

- Processors - Dual AMD EPYC™ 9005/9004 series processors, up to 500W
- GPU - High density 8U system with AMD Instinct™ MI325X/MI355X 8-GPU
- Memory - 24 DIMM slots, DDR5-6400 (1DPC); up to 6TB
- Storage drive - Up to Total 8 Drive Bays, 1 M.2
- PCIe Slots - 8 PCIe 5.0 x16 LP slots; Up to 4 PCIe 5.0 x16 FHFL slots
- I/O - 2 USB 3.0 ports, 1 VGA port, 1 BMC GbE LAN port
- Remote Management - ASPEED AST2600
- Power Supply - up to 4 redundant 6000W Titanium level power supplies



AS-4126GS-TMR-LCC



10U AMD OAM GPU (MI355X) Server

Purpose Built for Large Scale AI and HPC Workload

High performance AI/Deep Learning and HPC-optimized systems

Dual AMD EPYC™ 9005 Series Processors including CPUs with AMD 3D V-Cache™ technology

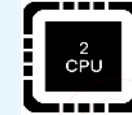
Double the CPU to GPU throughput with PCIe 5.0

Powered by AMD Instinct™ MI355X 8-GPUs

The H14 GPU-optimized servers deliver unprecedented acceleration at every scale to power the world's highest performing data centers for AI, data analytics, and HPC applications

Air-cooled

AMD
EPYC



Product Specs

- Processors - Dual AMD EPYC™ 9005/9004 series processors, up to 500W
- GPU - High density 10U system with AMD Instinct™ MI355X 8-GPU
- Memory - 24 DIMM slots, DDR5-6400 (1DPC); up to 6TB
- Up to 10 U.2/U.3 NVMe storage drives, 2 M.2 NVMe
- 8 PCIe 5.0 x16 LP; 2 PCIe 5.0 x16 FHHL
- 2 USB 3.0 ports, 1 VGA port, 1 IPMI GbE LAN port, 2 10G NIC (X710)
- Remote Management - ASPEED AST2600
- 6 redundant (4+2) 6600W Titanium level power supplies



AS-A126GS-TNMR



Supermicro AMD Instinct™ AI Clusters

MI325X(32)
Air Cooled Supercluster



- 32x GPUs per rack
- 4x Systems per rack
- 48U EIA Rack
- 50kW Compute Rack Power

MI325X(64)
Liquid Cooled Supercluster



- 64x GPUs per rack
- 8x Systems per rack
- 48U EIA Rack
- 106W Compute Rack Power

MI325X(96)
Liquid Cooled Supercluster



- 96x GPUs per rack
- 12x Systems per rack
- 52U EIA Rack
- 156kW Compute Rack Power

Multi-Processor System MI300A Product Family

AS-4145GH-TNMR



High-Capacity Air-Cooled Quad Socket Server

- Quad AMD MI300A Processors with 760W TDP
- 512GB HBM3 Unified Memory.
- Up to 16* 2.5" NVMe/24* SATA SAS drives.
- Expandable 8* PCIe Gen 5 Slots (x16).
- 2* AIOM Support (OCP 3.0).
- Optimized for additional IO/storage expansion.

AS -2145GH-TNMR-LCC



High Density Liquid-Cooled Quad Socket Server

- Quad AMD MI300A Processors with 760W TDP
- 512GB HBM3 Unified Memory.
- Up to 8* 2.5" NVMe drives.
- Expandable 6* PCIe Gen 5 Slots (x16).
- AIOM Support (OCP 3.0).
- Optimized compute density and energy saving.

4U AMD Air Cooled MI300A Server

High-Performance Interconnect with Robust Storage Capacity

Quad-socket maximizes data center performance and mixed workload.

Tailored for AI and HPC, supporting government labs, weather forecasting, energy research, genomics, drug design, finance, and risk analysis.

Single-chip solution with unified memory.

MI300A APU integrates GPU, CPU, and 128GB HBM3 into one unit. Unified memory boosts performance for high-demand workloads.

Superior storage and connectivity elevate performance.

Boost storage with up to 16* NVMe connections via PCIe Gen5, plus optional up to 24* SATA/SAS.

Enhance scalability meets diverse needs.

Enhances scalability with configurable 8* PCIe Gen5 slots (x16) or 16* PCIe Gen5 slots (x8), and 2* AIOM slots (OCP 3.0 compatible).



Product Specs

- APU: Quad AMD MI300A, 24 Zen4 CPU cores, 228 GPU compute unit with 760W
- Memory: 512G HBM3 memory
- Drives: 8 front 2.5" NVMe, up to 16 NVMe or 24 SATA/SAS drives
- Networking: 2 AIOM(OCP 3.0), 1X RJ45, support up to 400G connection
- Power: 4x 2700W hot swappable Titanium level
- Cooling: 10x heavy-duty 80mm hot-swappable fans
- Dimensions: 177x438.4x848.7 mm (7" x 17.3" x 33.4")



AS-4145GH-TNMR

2U AMD Liquid Cooled MI300A Server

High Performance Interconnect with Peak Compute Efficiency

Quad-socket maximizes data center performance and mixed workload.

Tailored for AI and HPC, supporting government labs, weather forecasting, energy research, genomics, drug design, finance, and risk analysis.

Single-chip solution with unified memory.

MI300A APU integrates GPU, CPU, and 128GB HBM3 into one unit. Unified memory boosts performance for high-demand workloads.

Optimize with power efficiency.

High performance with reduced 60% power consumption by Liquid-Cooled

High density 2U form factor maximizing scalability for rack level.

21x systems in a 48U rack. Achieve maximum compute density and space saving



Product Specs

- APU: Quad AMD MI300A, 24 Zen4 CPU cores, 228 GPU compute unit with 760W
- Memory: 512G HBM3 memory
- Drives: 8 front 2.5" NVMe
- Networking: 1 AIOM(OCP 3.0), 1X RJ45, support up to 400G connection
- Power: 4x2700W hot swappable Titanium level
- Cooling: 3x heavy-duty 80mm hot-swappable fans
- Dimensions: 88 x 438.4 x 848.7 mm (3.5" x 17.3" x 33.4")



AS-2145GH-TNMR-LCC

2U H13 Server

8U Intel® Gaudi®3 OAM GPU Server

Click to edit Subtitle

Accelerated Generative AI

Includes 128GB HBM2e, 64 Tensor Processor Cores, and 8 Matrix Math Engines, combining to drive greater efficiency and performance for accelerating deep neural network computations.

Offers seamless compatibility with the PyTorch framework, resulting in support for hundreds of Hugging Face models, frameworks, and libraries with Intel's open-source software stack

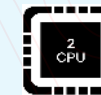
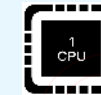
Ability to Scale

Readily supports demanding Gen AI workloads from 1 to 1000s of nodes.

Uses on-chip networking – 24 200Gb Ethernet ports - which allows for easy and cost-effective integration into Ethernet-based networks.

Reduced Energy and TCO

Intel Gaudi systems have a lower hardware cost, provide more energy savings, and have no software licensing costs - giving a combined better cost of ownership savings.



Product Specs

- Dual Intel® Xeon® 6900 series processors with P-cores
- Support for up to 24 DIMMs DDR5-6400 and 8800MT/s MRDIMMs
- Support for 8 Gaudi 3 OAM GPUs
- 6 On-board OSFP 800GbE ports
- Up to 2 PCIe 5.0 x16 FHFL and 2 PCIe 5.0 x8 FHFL slots and Default AIOM dual port 10GBase-T
- Up to 8 2.5" hot-swap NVMe PCIe 5.0 drive bays



SYS-822GA-NGR3



Rackmount Edge System Product Families

Ultra Short Depth



400 mm Depth

- 1U & 2U servers for space constrained environments with front or rear I/O and low-noise options
- Ideal for edge compute and Edge AI workloads
- Single and dual Intel® Xeon®, Core™, Core™ Ultra, AMD EPYC™, or NVIDIA Grace™ processors
- Multiple options for single-width and double-width GPU using PCIe expansion slots

1U Short Depth



1U single processor

- 1U servers with performance & efficiency for Edge AI
- Single & dual Intel® Xeon®, single Core™ or single AMD EPYC™ processor
- Front or rear I/O options
- Multiple options for single-width or double-width PCIe GPUs

2U AI Optimized



2U single or dual processor

- 2U servers with performance & efficiency for Edge AI
- Single & dual Intel® Xeon®, single Core™ or single AMD EPYC™ processor
- Front or rear I/O options
- Multiple options for single-width or double-width PCIe GPUs

Multi-node



Maximize Density and Efficiency

- 2U servers, including short-depth option
- Hot-pluggable nodes, each with single Intel® Xeon® processor and up to 8 DIMMs per node
- Supports half height, half length or full height, half length PCIe expansion cards
- AC or DC power configurations

Rackmount Edge: Ultra Short Depth Servers

Optimized for Space-constrained Edge Deployment and Edge AI workloads.

Multiple form factor choices,

Multiple ultra-short depth 1U and 2U systems that are ideal for network edge deployment with ability to fit into space-constrained environments

I/O flexibility for networking

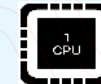
Select server with either Front I/O or Rear I/O

Choice of Intel CPU options

Support for Intel Xeon 6700/6500 series , Xeon® 6 SoC, Core™ i9/i7/i5/i3, Pentium® and Celeron

Flexible configurations

Variety of PCIe expansion card options, and multiple storage options enabling edge compute and Edge AI inferencing



Product Specs

- Single Intel® Xeon 6 SoC, Core™ i9/i7/i5/i3, Pentium® and Celeron(1U), Xeon® 6700/6500 (2U)
- Multiple memory options ranging from 4 – 8 DIMM slots depending on product
- Multiple storage options for hot-swappable or fixed 2.5" NVMe/SAS/SATA drive bays depending on product
- Variable expansion slots from 1 – 3 10 PCIe 5.0 x16 or PCIe 5.0 x 8 FHFL, FHHL, HHHL, and LP options.
- Flexible, multiple networking options
- Support from 1 - 3 single-width or 1 double-width GPU depending on product, Note: SYS-111AD-HN2 does not support GPU



SYS-111AD-HN2
SYS-112D-20C-FN3P
SYS-212B-FN4TP



Rackmount Edge: 1U Short Depth

Optimized for Edge Deployment and Edge AI workloads.

Multiple choices of 1U edge systems in Intel, AMD, or NVIDIA CPU

A comprehensive portfolio designed for network edge deployment, offering flexible configurations to meet diverse memory requirements and CPU compatibility

I/O flexibility for networking

Option to select the server form factor with Front I/O or Rear I/O configurations.

Range of CPU options

Supports Intel® Xeon®, Core™ i9/i7/i5/i3 , and AMD EPYC™, NVIDIA Grace C1

Variety of single-width GPU options

Supports a variety of PCIe-based single-width GPU, enabling Gen AI inferencing and rendering at the edge across Retail, Manufacturing, Healthcare, and Finance sectors



Product Specs

- Single Intel® Core i9/i7/i5/i3 Processors , Xeon 6700/6500 ,single AMD EPYC™ 8004 series processors, Nvidia Grace C1 Processor
- 4 DIMM (Core i), 8DIMM(Core I, Xeon 6700/6500) 6 DIMM (Single AMD EPYC™ 8004,
- Multiple storage options for 1U
- Variable and flexible expansion slots in 1U form factors
- Flexible, multiple networking options
- Variety of single-width GPUs



SYS-111AD-WRN2
SYS-111E-FWTR
SYS-112B-FWT
SYS-512AR-N4T
AS-1115S-FDWTRT
ARS-111L-FR



Rackmount Edge: 2U Single and Dual Processor Servers

Optimized for Edge Deployment and Edge AI workloads.

Multiple choices of 2U edge systems with Intel or AMD motherboard

A comprehensive portfolio designed for network edge deployment, offering flexible configurations to meet diverse requirements with options to balance PCIe expansion, GPU performance, and motherboard compatibility

I/O flexibility for networking

Option to select the server form factor with Front I/O or Rear I/O configurations.

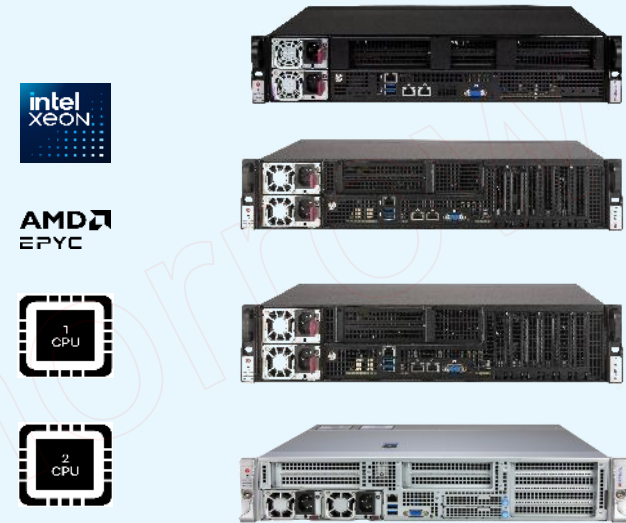
Range of CPU options

Supports single and dual processor Intel® Xeon® 6700/6500, AMD EPYC™ 9005/9004 2U dual processor systems support 32 DIMMS.

Range of GPU options

2U single processor servers support a variety of PCIe-based single-width and double-width GPUs

2U dual processor supports 3 double-width or 4 single-width GPUs, enabling Gen AI inferencing and rendering at the edge across Retail, Manufacturing, Healthcare, and Finance sectors



Product Specs

- Single or dual Intel® Xeon® 6700/6500 and single AMD EPYC™ 9005/9004 series processors
- Support 8 DIMM (Single Xeon® 6700/6500), 12 DIMM (Single AMD EPYC™ 9005/9004), 32 DIMM (Dual Xeon processor)
- Multiple storage options for each of the 2U servers. 2U (dual processor) supports 6 drive bays. It also supports 2 M.2 NVMe slots.
- Variable and flexible expansion slots
- Flexible, multiple networking options
- 2U Single processor supporting a variety of single-width and double-width GPUs
2U Dual processor supports 3 double-width or 4 single-width GPUs



SYS-212B-(F)N2T
SYS-212B-(F)LN2T
AS-2116S-FTNRT
SYS-222HE-FTN



Rackmount Edge: Multi-Node Servers

Maximize Density and Efficiency

Choice of 3-Node or 4-Node system in 2U form factor

3-Node system (3 hot-swappable nodes) is a short depth 2U form factor with front I/O

4-Node system (4 hot-swappable nodes) is a standard 2U form factor with rackmount depth with rear I/O

Common CPU (Per Node)

Both systems support single socket Intel® Xeon® 5th/4th Gen processors

Common Memory (Per node)

Both systems support 8 DIMMs with max memory up to 2TB 5600MT/s ECC DDR5 RDIMM

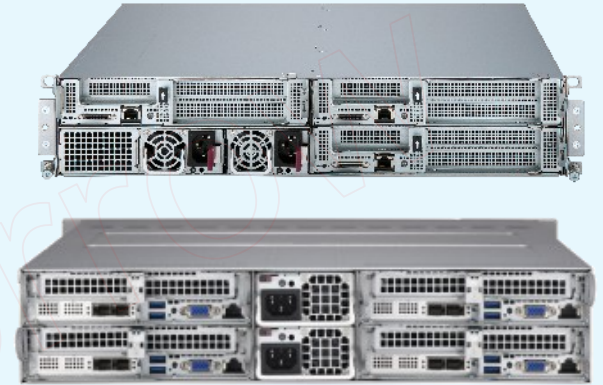
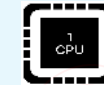
Half Length PCIe expansion slots (Per Node)

3 Node system supports 1 PCIe 5.0 x16 HHHL and 2 PCIe 5.0 x16 FHHL slots

4 Node system supports 2 PCIe 5.0 x16 HHHL slots

Common remote management

Both systems support dedicated IPMI LAN port



Product Specs

- Single socket Intel® Xeon® 5th/4th Gen processors *per node*
- 8 DIMMs *per node*, Up to 2TB 5600MT/s ECC DDR5 RDIMM
- 2 M.2 storage PCIe 5.0 x4 NVMe (M-key 2280/22110) in 2U system *per node*
- Flexible PCIe expansion slots in 2U form factor, 1 HHHL and 2 FHHL slots *per node*
- 1 RJ45 1 GbE LAN port *per node*



SYS-211SE-31A/D
SYS-211TP-HPTR



Compact Edge System Product Families

Fanless Gateway



Reliable Operations in Harsh Environments

- Fanless cooling eliminates moving parts, increasing resistance to vibrations, achieving quieter operation, and improved durability
- Range from palm-sized to ultra-compact and compact (mini-ITX) form factor
- Built for deployment in harsh environments and industrial settings

Performance Fan-Based Gateway



Balanced Performance & Efficiency

- Fan based, air-cooled systems with multiple form factor choices
- Many choices of CPUs and diverse I/O, storage, AI acceleration, and expansion options
- High performance compact edge servers running AI models outside of centralized data centers.

Compact Server



Compact Portable Server for AI Compute

- High performance Edge AI and edge compute in a compact form factor, easy to transport
- Versatile indoor and outdoor deployment options
- Easy Portable option using Pelican Case for secure, carry-on travel
- Front I/O provides ease of maintenance and I/O support for LAN, USB, Video, and Serial ports

Compact Edge: Fanless Systems

Embedded fanless systems for reliable operations in harsh environments

Embedded servers

Highly reliable fanless embedded servers with very small form factor ideal for deployment in applications such as airport security, industrial manufacturing, operating rooms. Very low TDP makes these servers very cost effective

Variety of processors

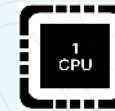
Offers Intel® Processor N97, Atom® x7433RE, x7835RE, Arm® Cortex®-A78AE, Core i3/i5/i7 or Celeron with 1 DIMM slot.

Wide operating temperature

Servers operate in a wide thermal range 20°C to 60°C (E101, E302). Jetson Orin operates at -25°C to 60°C

Small form factor, storage, mounting options

They have varying depth from 5" to 8" depth. The server with 8" depth offers 1 SATA drive. The Jetson Orin NX offers mounting options as wall, desktop, Din-Rail



Add Product Specs

- Multiple fanless products use one of the following single Intel® Atom®, 12th Gen Core™, or Intel® Processor N97. The ARS-103-JONX-H is built using a single NVIDIA Arm® Cortex®-A78AE v8.2 64-bit CPU
- Memory : Jetson Orin NX available at 8GB or 16GB. Other servers have onboard memory or 1 DIMM
- Storage : Support M.2 or 1 SATA Drive
- I/O networking options : Multiple options of RJ45 LAN ports.
- Operating temperature: Wide range from -20°C to 60°C. Jetson Orin operates from -25°C to 60°C

SYS-E50-14AM
SYS-E100-14AM-E
SYS-E101-14AM-H
SYS-E302-13-AD
ARS-E103-JONX-H2



Compact Edge: Fan-based Performance Systems

Fan-Based Systems for Balanced Performance & Efficiency

Fan based embedded servers

Balanced performance and efficiency in these fan-based embedded servers for verticals like Retail, Digital signage, Kiosks, Security, Industrial Automation
Very low TDP makes these servers very cost effective
Short depth of less than 9 inches and some servers only 1.7" height.

Variety of processors

Single processor support for Core™ Ultra 9/7/5, Xeon-D 1749NT, Atom® x7835RE, with up to 2 DIMM slots.

Wide operating temperature

Operates in a wide operating temperature range 0°C up to 40°C (32°F to 104°F). Extended range for SYS-E102-14AM-H up to 50°C (32°F to 122°F)

Flexible configurations

Variety of storage and expansion configurations available including PCIe expansion, M.2, SATA drive bay options, and multiple I/O including RJ45 LAN , USB, Video ports enabling edge compute and Edge AI inferencing.



Add Product Specs

- Single processor: Intel Xeon D (E102-14AM and E200-12D), Intel® Core™ Ultra 9/7/5 (Series 2) processors (E201-14AR and E300-14AR)
- Memory : 2 DIMM slot (E200, E300), 1 DIMM slot (E102)
- Storage : 1 Drive bay where applicable
- Expansion slots :1 PCIe 5.0 x16 LP (E300), multiple M.2 options
- I/O networking options multiple I/O including RJ45 LAN , USB, Video ports
- Operating temperature: 0°C up to 40°C (32°F to 104°F), extended to 50°C (32°F to 122°F) for SYS-E102-14AM-H



SYS-E102-14AM-L
SYS-E200-12D
SYS-E201-14AR
SYS-E300-14AR



Compact Edge: Compact Portable Server

Compact Portable Server for Edge AI and Edge Compute

High performance Edge AI and edge compute in a compact form factor

Supports Intel® Xeon® 6700/6500 series processors as well as up to 3 PCIe 5.0 x16 expansion slots

Deployment options both indoors and outdoors

Indoors – multiple options for wall mount or it can be set on a rackmount shelf

Outdoors – deploy within an IP65 compliant enclosure (SYS-403-12P-IPA2-C) for harsh environments with Operating Temp range of -40°C to 46°C + solar load (-40°F to 115°F + solar load)

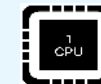
GPU – Supports either 1 double-width or 2 single-width GPUs

Easily portable option

System can be placed into a custom cut Pelican Case for secure, carry-on travel and transport

I/O flexibility, storage and PCIe expansion

Front I/O provides ease of maintenance and I/O support for LAN, USB, Video, and Serial ports, 3 PCIe 5.0 x16 FHFL slots, 4 drive bays.



SYS-E403-14B-FTN2T

- Single Intel® Xeon® 6700/6500 series processors with P-cores or 6700 series processors with E-cores up to 300W with air cooling
- Up to 8 DIMMs supporting up to 2TB DDR5-6400MT/s RDIMM
- 2 RJ45 10GBASE-T
- Up to 3 PCIe 5.0 x16 FHFL slots
- Up to 2 front hot-swap 2.5" NVMe + 2 internal fixed 2.5" SATA drive bay
- 1 double-width or 2 single-width GPUs

SYS-E403-12P-IPA2-C

- IP65 enclosure for harsh outdoor environments
- Operating Temperature -40°C - 46°C + Solar load
- Cabinet intrusion detection and self-diagnosis
- Single Socket Intel® Xeon-SP (Ice Lake) up to 32 Cores
- 3 Hot swap fans



SYS-E403-14B-FRN2T
SYS-E403-12PA2-C



Workstation Product Families

1 GPU/Integrated Graphics



Uncompromised single-GPU Experience

- Optimized for creative, engineering, and AI workloads that take full advantage of the latest single GPU architectures.
- Compact form factors ensure seamless integration in space-constrained environments without sacrificing thermal headroom.
- Designed for professionals prioritizing stable performance, lower power draw, and workstation-class reliability.

2 GPU



Balanced Performance & Scalable Acceleration

- Enables parallel computing for demanding AI training, simulation, and advanced visualization workloads.
- Flexible expandability to upgrade connectivity, storage, and memory for growing compute needs.
- Dual GPU configurations deliver enhanced throughput for users needing more cores and memory bandwidth, without oversizing.

4 GPU



Maximum compute density for extreme performance

- Harnesses the full power of four modern GPUs for high-end AI development, deep learning, and rendering applications.
- Engineered for robust cooling, power delivery, and high-speed I/O to meet the demands of intensive tasks.
- Ideal for researchers and studios requiring maximum scalability, resource pooling, and future-proofed compute density.

1 GPU / Integrated Graphics

Uncompromised Single-GPU Experience

Precision Performance

Delivers professional-grade performance for creative, engineering, and AI tasks on a single high-end GPU, ensuring optimal stability and efficiency.

Optimized Form Factor

Compact, quiet chassis engineered for demanding offices and labs, with robust thermal design and low power usage.

Versatile Expansion

Expandable storage and fast I/O for streamlined workflows, with flexible PCIe and NVMe options supporting modern SSDs and connectivity needs.



AS -3015A-I

SYS-532AW-C



AS -1015A-MT

Product Specs

- Intel Core Ultra 200S, AMD EPYC 4005, or AMD Ryzen 9000 series processors*
- Supports an extensive lineup of consumer and professional GPUs across NVIDIA, AMD, and Intel platforms
- Energy-efficient for small office and home office environments



SYS-532AW-C



AS -3015A-I



AS -1015A-MT



* Optional & model-specific

2 GPU

Balanced acceleration for advanced workloads

Dual-GPU Acceleration

Enables multi-threaded rendering, simulation, and AI compute with scalable processing and responsiveness.

Enhanced Memory & Bandwidth

Supports higher memory density and bandwidth for complex projects, including up to 32 DDR5 DIMMs.

Comprehensive Connectivity

Versatile expansion for storage and networking, leveraging PCIe Gen5 for lower latency, direct NVMe throughput, and multiple x16 slots for graphics and accelerators.



SYS-531A-I



AS-531AW-TC



AS-2015A-TR



SYS-741A-T

Product Specs

- Dual 5th/4th Gen Intel® Xeon® Scalable processors, Intel Xeon W2500, AMD Threadripper PRO 9000WX, AMD EPYC 4005, or AMD Ryzen 9000 series processors*
- Up to 512GB of 6400MT/s ECC DDR5 memory*
- Built-in IPMI supports secure remote management, health monitoring, and recovery, even when offline or powered off



SYS-531A-I



AS-531AW-TC



AS-2015A-TR



SYS-741A-T

* Optional & model-specific



4 GPU

Maximum compute density for extreme performance

Extreme Parallelism

Purpose-built for the most intensive AI, deep learning, and visualization workflows, featuring up to 4 double-width GPUs for parallel processing and extreme acceleration

Advanced Cooling & Power

Advanced cooling and power delivery, with support for high TDP processors and redundant PSUs to ensure consistent peak operation in demanding environments.

Massive Memory & Storage

Large capacity memory and storage, supporting up to 4TB ECC DDR5 and multiple hot-swap NVMe bays for massive datasets and rapid data access.



SYS-741GE-TNRT

SYS-551A-T
AS -5014A-TT



AS -2115HV-TNRT

Product Specs

- Intel Xeon W3500, or AMD Threadripper PRO 9000WX series processors*
- Supports up to 4 x Nvidia RTX PRO 6000 Workstation Edition Max-Q (300W) GPUs*
- Up to 4TB of 5600MT/s ECC DDR5 memory*



SYS-551A-T

AS -5014A-TT

AS -2115HV-TNRT

SYS-741GE-TNRT

* Optional & model-specific



Storage Product Families

ENTERPRISE



TOP LOADING



SIMPLY DOUBLE



HIGH AVAILABILITY DUAL-PORT



PETASCALE ALL-FLASH



Capacity Optimized - Cloud Density 3.5" Storage

- Optimized for enterprise data centers and edge deployments
- Front access, double-sided, product families
- 2U/3U/4U form factors
- 16-36 3.5" LFF Bays
- Maximum capacity for cloud-scale storage in 4U FF
- Highest storage capacity and \$/TB value
- Single node and dual-server systems
- 60, 90 drives in 4U
- Dual Intel CPUs
- High density LFF storage in 2U FF
- 2U Form Factor
- Up to 24x 3.5" bays and up to 4x Gen5 NVMe SSDs for caching
- Single Intel (X13) or AMD (H13) CPU

Hard Drive Products

TCO Optimized All-Flash Storage

- High Availability Dual Port (Active-Active or Active-Passive)
- Shared Everything Architecture (Scale-up Building Block) in 2U
- All-flash 24x U.2 NVMe drives
- Single Intel Xeon 5th Gen CPU, dual server configuration
- Revolutionary storage performance and capacity
- Highest IOPS and BW
- All-flash EDSFF NVMe in 1U and 2U
- Up to 32x E3.S NVMe SSDs
- Dual Intel and Single AMD CPUs

Flash Storage Servers



Enterprise Optimized Storage Highlight

Channel Friendly Barebone

Perfect for value-add solution with traditional enterprise storage applications

PCIe 5.0 architecture supports single and dual processor

Up to 36 3.5" HDD and 6 Gen 5 NVMe hybrid bays for caching

Built Upon Open Industry Standard Building Blocks

Simple ATX Motherboard architecture without complex risers

Chassis depths fit-in the conventional DC rack (less than 28" deep)

SAS3 -12G controller included in base configuration

L-Series feature ITmode/HBA

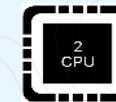
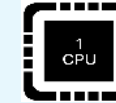
H-Series feature hardware RAID with Option for 8GB Flash backed Write cache

Target Solutions

General purpose file serving with Windows and Linux OS

Storage building blocks for software-defined storage

Cost effective scale-up solution with external JBOD expansion



Product Specs

- Dual 4th Gen Intel® Xeon® Scalable processors up to 270W
- Memory: Intel - 16 DIMM slots, supporting up to 4 TB DDR5
- Expansion: 6 PCIe 5 slots
- Storage: 12/16/24 front hot-swap 3.5" SAS/SATA and 2 rear hot-swap 2.5" SATA
- Storage Controller: Broadcom 3808 or 3908
- Redundant 1200/1600W Titanium level power supplies



Click to edit Model number



4U Cloud Scale Top Loading Highlight

Accessibility and Efficiency for Data Centers

Highest Density 3.5" Storage Servers in the Portfolio

Single, Twin, and JBOD configurations

Up to 2.16 PB capacity in 90 bay or 1.44 PB capacity in 60 bay system (with 24TB HDD)

Easy Tool-less Serviceability of all Active Components

Storage Drawer Design ensures user safety even in fully populated racks

Enterprise-grade design with cam-lever removal of drives, expanders, and nodes Best-in-class performance with bandwidth to support dual actuator media

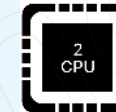
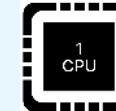
Target Solutions

Cloud scale object storage

Data-lake storage

Back-up target

Origen repository for media streaming



Single Node



Dual Node

Product Specs

- CPU: Single Intel® Xeon® 6700 Series processor
- Expansion: 4 PCIe Gen5 slots
- Storage: 60 hot-swap 3.5" SAS/SATA and 2 rear hot-swap 2.5" SATA
- Internal 2 Gen 5 NVMe SSD support
- Storage Controller: on board Broadcom RAID for boot device
- Power Supply: Redundant 2000/2600W Titanium level



Click to edit Model number



Simply Double Hybrid Storage Highlight

The New Standard for Cloud Analytics – Double the Storage in 2U

PCIe Gen 5 architecture with single Intel or AMD CPU

SAS3 -12G controller included in base configuration

L-Series feature ITmode/HBA

H-Series feature hardware RAID with Option for 8GB Flash backed Write cache

Hybrid Media Support to Optimize Performance and TCO Value

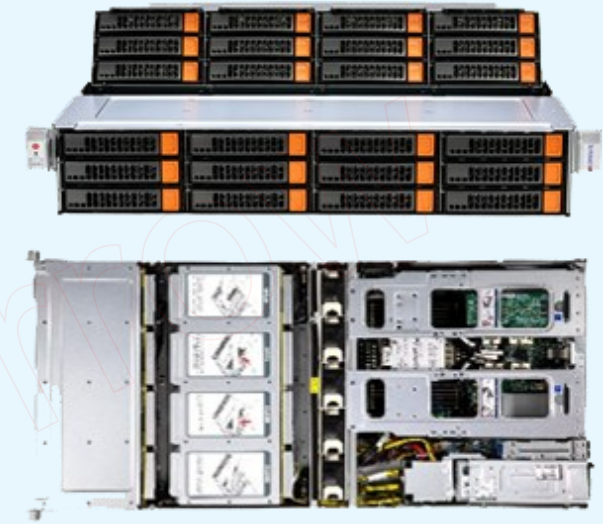
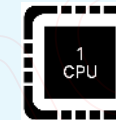
Up to 24 3.5" HDD (576TB) for data storage

Up to 4 Gen 5 NVMe in the rear for caching and performance tiering

Consolidate storage workloads and lower data center TCO

Target Solutions

Big Data, Data Lake Hadoop storage, SDS Cloud object storage, CDN, and video streaming



Product Specs

- Single Intel® Xeon® CPU up to 300W or Single AMD EPYC™ CPU up to 290W
- Memory: Intel - 16 DIMM slots, supporting up to 4TB DDR5 or AMD - 12 DIMM slots, supporting up to 3 TB DDR5
- Expansion: 6 PCIe 5 slots and 2 PCIe 5 AIOM slots
- Storage: 24 front hot-swap 3.5" SAS/SATA and 2 rear hot-swap 2.5" SATA
- Storage Controller: Broadcom 3808 or 3908
- Redundant 1600W Titanium level power supplies



Click to edit Model number



High Availability Dual-Port NVMe

Fully Redundant Dual-ported PCIe PCIe 5.0 Architecture

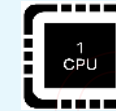
Software defined storage array for mission critical data
supports single Intel processor
Up to 24 U.2 dual port NVMe

High Performance Resiliency in 2U

Perfect for Linux + ZFS architectures
Commercial ISV Ecosystem allowing for NVMe acceleration
JBOD expansion option for tiered solutions using 90-Bay 3.5" JBOD(s)

Target Solutions

Scale-up storage, tiered storage
Parallel file systems, OSS/MDS
SAN/NAS Appliance, paired with SDS solution like OSNexus



Product Specs

- Single 5th Gen Intel® Xeon® processor up to 350W
- Memory: 8 DIMM slots, supporting up to 2TB DDR5
- Expansion: 4 PCIe 5 slots
- Storage: 24 hot-swap 15mm dual port U.2
- Networking: Dual 10GbE BaseT with Intel® X710
- Redundant 2000W Titanium level power supplies



Click to edit Model number

EDSFF Petascale All-Flash Highlights

Scale-out or Scale-up Storage for AI and CSP Workloads

TCO Optimized Design for Large-Scale Flash Applications

Optimized thermal design with EDSFF spec, balanced PCIe lanes for front SSD and rear I/O provides balanced non-blocking bandwidth for network-based clients

High Efficiency EDSFF Architecture Takes Performance to the Next Level

Allowing for more performance per watt than U.2 NVMe, as well as provides a new media bays standard for advanced technologies like FPGA and CXL devices

CXL Type-3 Expansion of Memory Bandwidth and Capacity

Up to 32 DIMMs plus up to 8 CXL expansion bays remove bottlenecks for memory constrained applications, like AI inferencing and in memory database



Click to edit Model number



Product Specs

Three motherboard platforms to choose from:

- Dual Intel® Xeon® 6700 series processor
- NVIDIA Grace Superchip (1U/16 E3.S)
- Single AMD EPYC™ 9004/9005 series

Optimized thermal design with EDSFF media support

- 1U E1.S and E3.S and 2U E3.S 1T and 2T (CXL)
- Up to 8 CXL modules in 2U

Balanced PCIe lanes for optimal I/O performance

- Up to 1.9 PB in 2U using E3.S media
- Up to 30M 4KB random read IOPS
- Up to 230 GB/s sequential read bandwidth



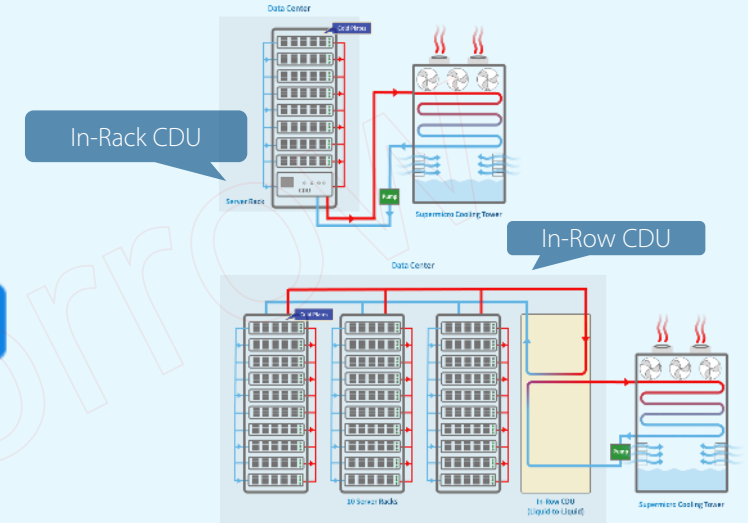
Supernano Liquid Cooling Solutions

Long history of liquid cooling solutions offerings

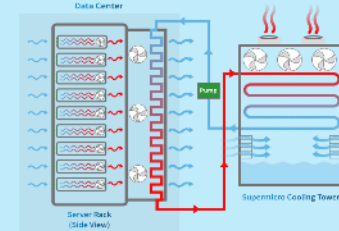


Reduction in Data Center Electricity Costs

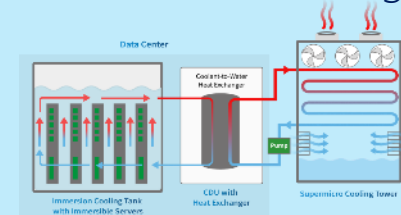
Direct-to-Chip



Air-to-Liquid



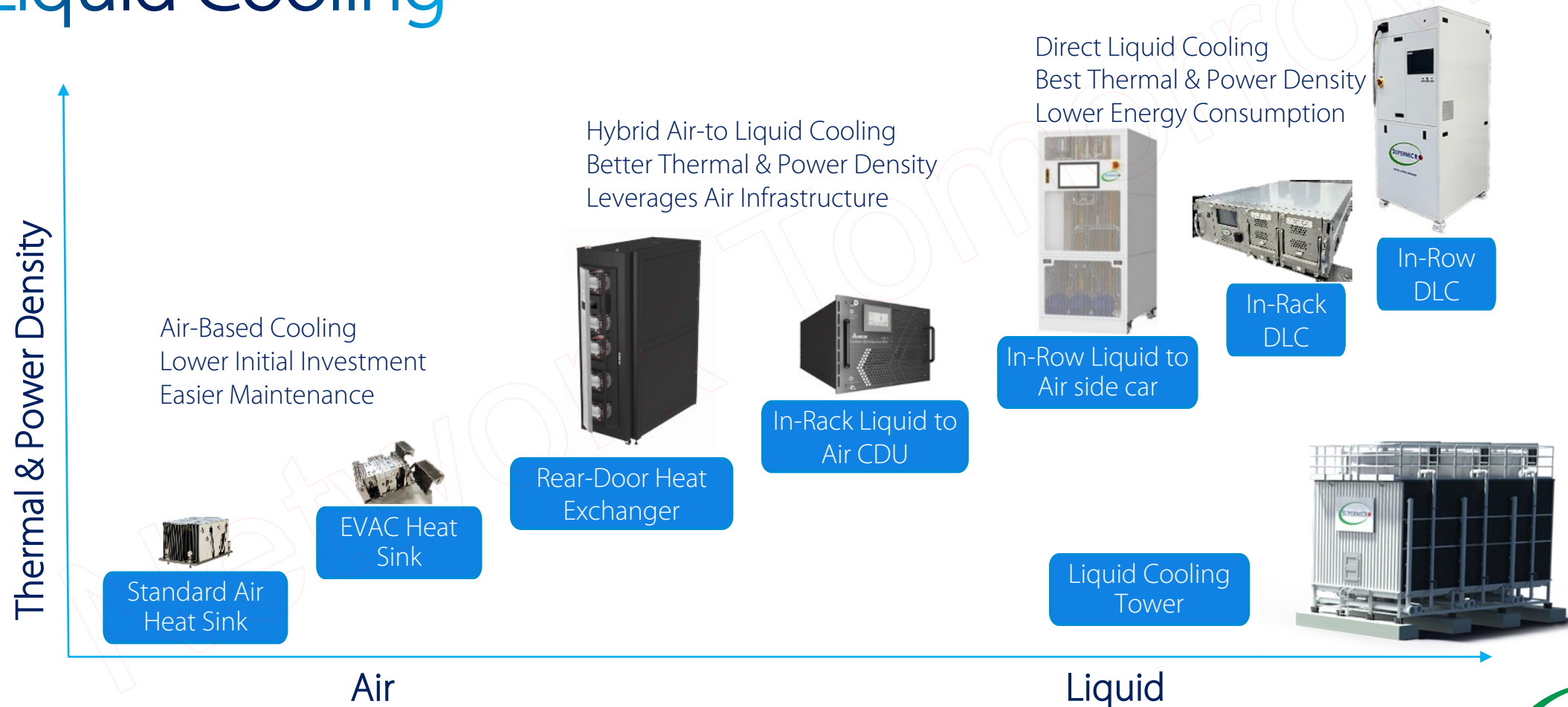
Immersion Cooling



Scan the QR Code to Learn More About Supernano Immersion and Rear Door Heat Exchanger Liquid-Cooling Solutions!



Journey to Direct Liquid Cooling



Supermicro CDU Portfolio



250kW 4U CDU



100kW 4U CDU

- N+1 Hot-Swappable Pumps & PSUs
- Pressure/Flow Rate/Leakage Monitor
- Automatic Anti-Condensation Control
 - Easy-to-Use Touch Screen
 - Remote Control Options
- Supports Up to 45°C Facility Water



1.8MW In-Row CDU



Reduction in
Server Electricity Cost



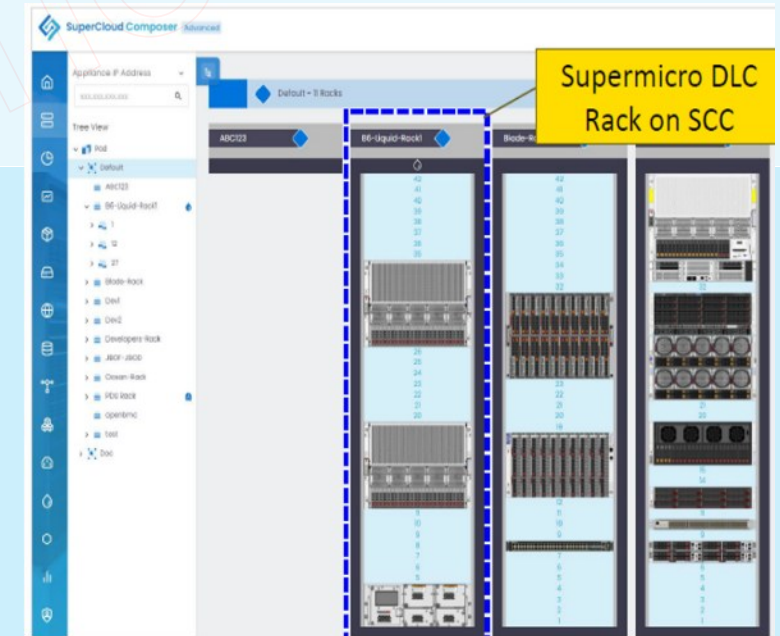
Reduction in
Data Center Server Noise

SuperCloud Composer® LCCM: Liquid Cooling Consult Module

Tool to collect and present in real time, data on physical assets and from sensors on Coolant Distribution Unit (CDU) and Cooling Tower

- Coolant/Facility Water Temperature
- Pump
- Valve
- Pressure
- Flow Rate
- Humidity
- Leakage Detection

Users can monitor operating efficiency of their liquid cooled racks, set up alerts, manage updates, and more.



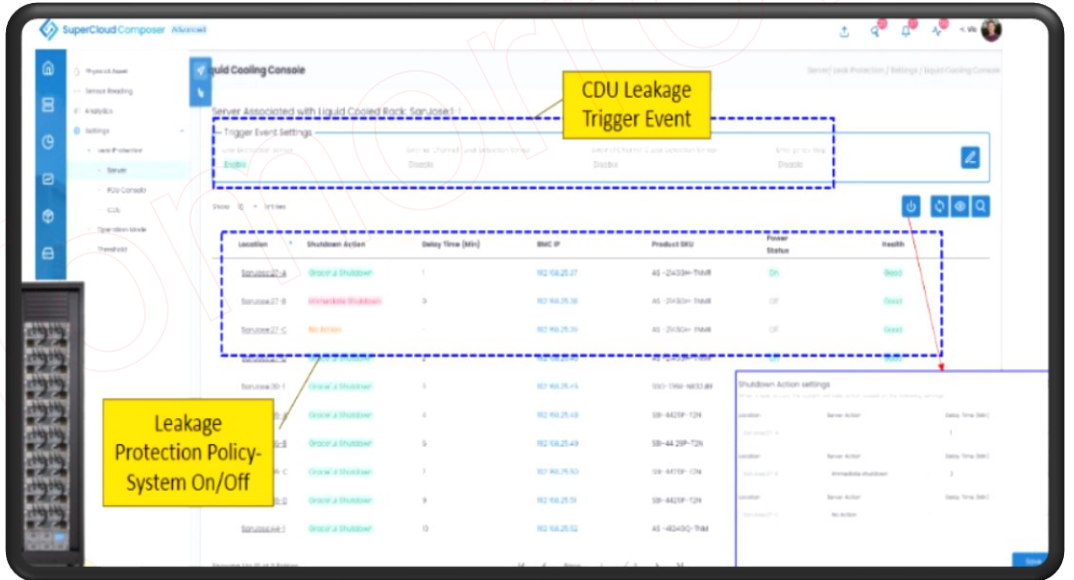
SSCC Liquid Cooling Management

Features

- Leak Detection & Alerting
- System Protection against Leaks
- Power Consumption Optimization
- Integrates with existing Alerting and Remediation workflows
- Instant Remediation w/automated system shutdown

Customer Benefit

- Protect expensive liquid cooling equipment.
- Save Power/OPEX \$\$ with power consumption optimization



Supermicro Liquid Cooling Solutions

Lawrence Livermore National Laboratory "Ruby"

- COVID 19 Research
- 1528 nodes (Intel Xeon with Intel Deep Learning Boost)
- Ranked #79 on November 2020 Top500 List
- Cooling Type - Liquid to Liquid (Direct Liquid Cooling)
- Press Release: [Supermicro Scalable Liquid-Cooled Supercomputing Cluster Deployed at Lawrence Livermore National Laboratory for COVID-19 Research](#)



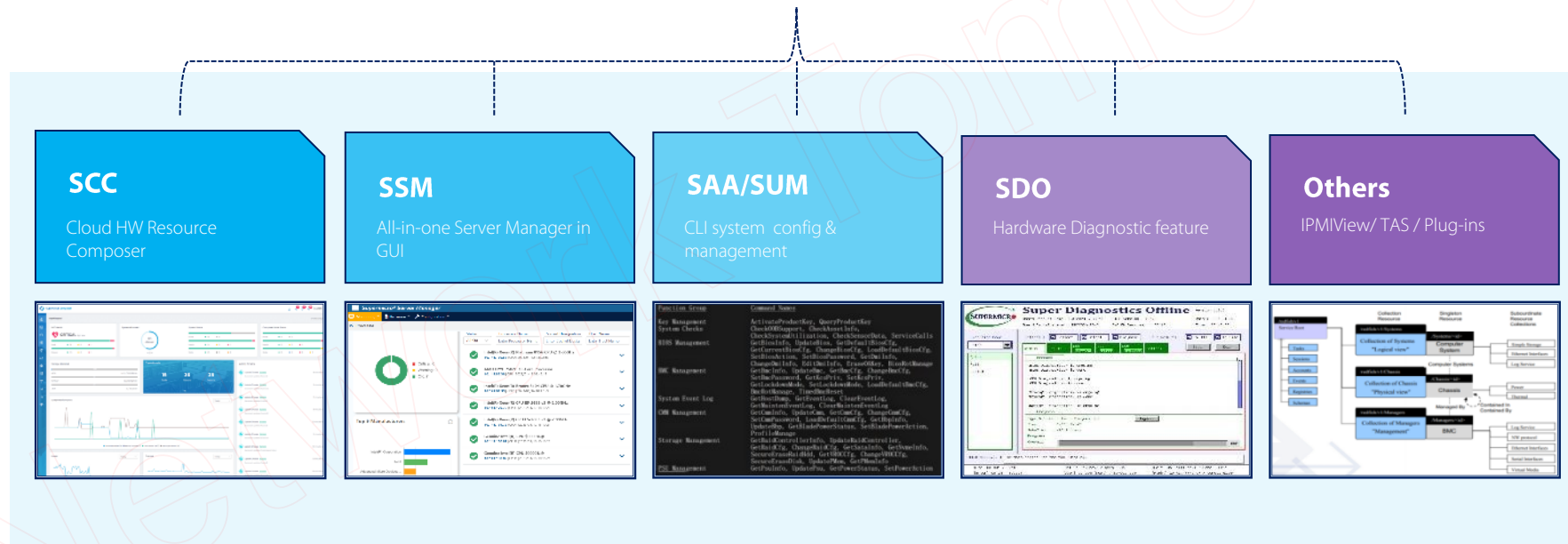
Osaka University - SQUID

- 1520 general-purpose CPU nodes
- 1528 nodes (Intel Xeon with Intel Deep Learning Boost)
- Ranked #79 on November 2020 Top500 List
- Cooling Type - Liquid to Liquid (Direct Liquid Cooling)
- Read More About [Osaka University and Supermicro Servers](#)

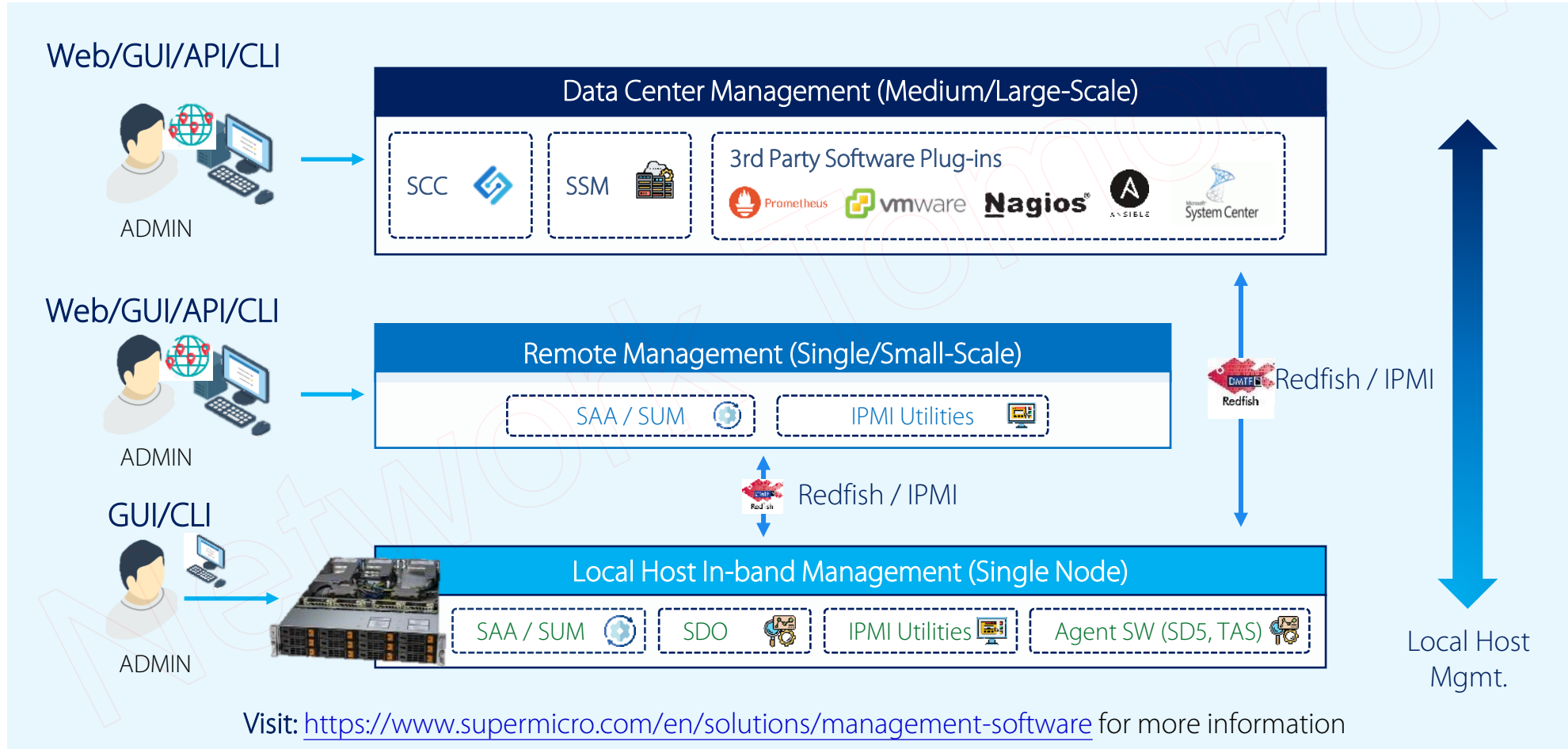


System Management Software Suite

Supermicro System Management SW Suite

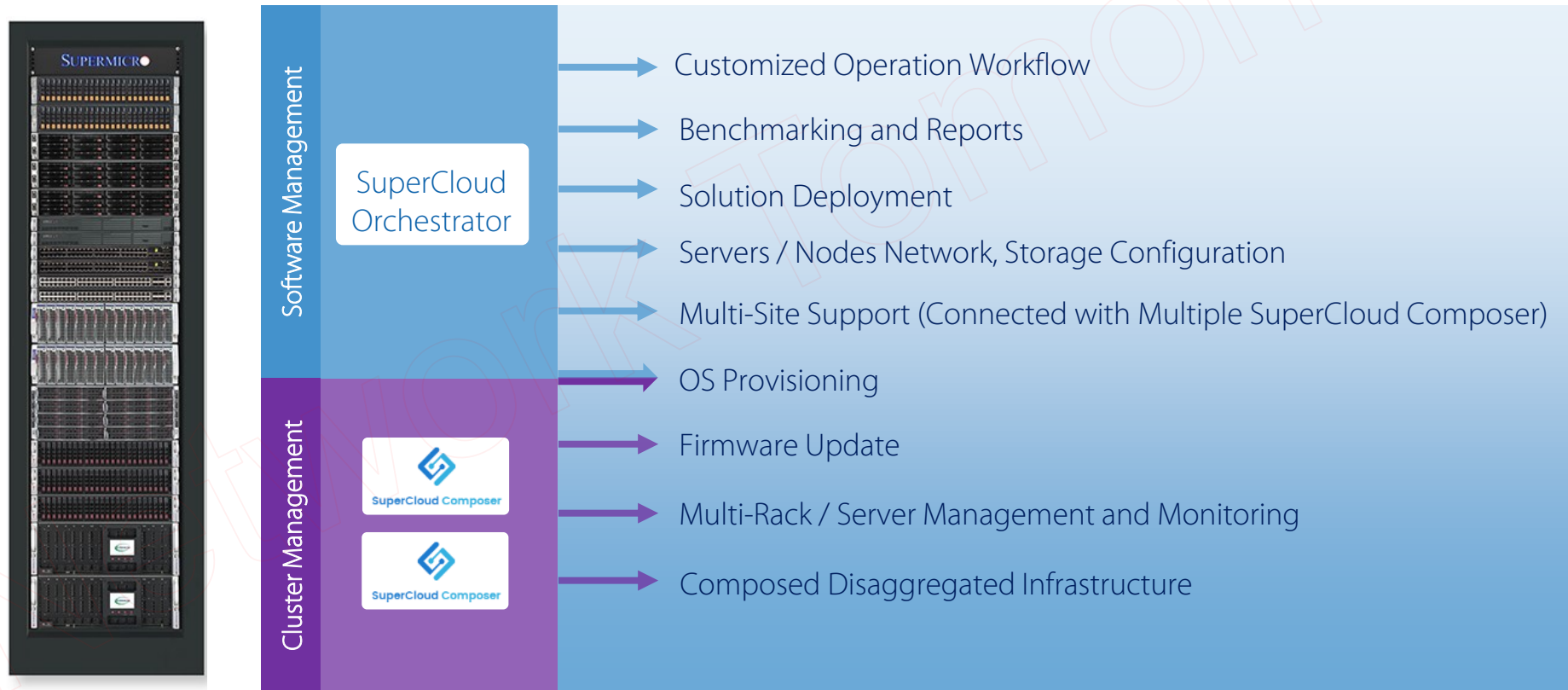


Supermicro System Management SW Portfolio



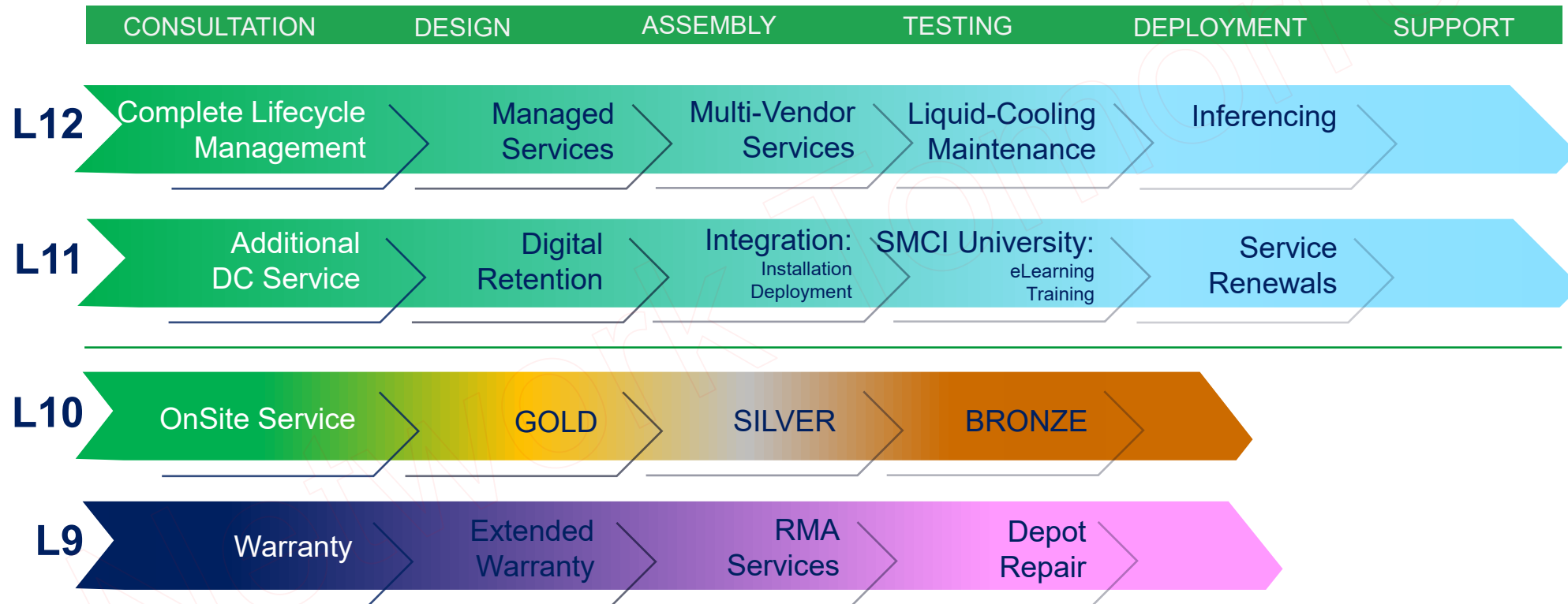
SuperCloud Platform Composer and Orchestrator

A streamlined platform from infrastructure to solution lifecycle management



Supermicro Services and Offerings

We offer Custom Service Solutions for your exact business needs!



Premium Service Features

- Standard Warranty (Included)
- Onsite Servicem (Service Code)

Access to Additional Service Options

- Custom Solutions to fit customer requirements
- Best-In-Class solutions design
- Installation and Integration Services

Supermicro Global Services

Warranty		Onsite Service		Bronze SMSADx	Silver OSNBDx	Gold OS4HRx
Technical support provided during business hours at Supermicro Regional HQ		Service Level Agreement		3 Business Day	Next Business Day	4-hour
Depot based Repair at Supermicro Regional HQ		Parts	Separate Stock to meet SLA	●	●	●
Standard Inventory Pool		Local Depot	In-Country or Metro to meet SLA	●	●	●
One-Way shipping In-Region		Advanced Logistics	2-Way Shipping Import/Export, and VAT.	●	●	●
No SLA		24x7 Service Desk	Entitlement checking, Phone Support, Partner escalation	●	●	●
Technical support CRM		Web Portal	Ticket submission, Case tracking and Routing	●	●	●
No Technical Account Manager		Field Technician	Global Availability Component replacement	●	●	●
		Technical Account Manager	Account Management Single point of contact, Escalation Support	●	●	●

● Included with Service

● leverages Service Desk Staff

● leverages Customer Staff

X describes service term in years

Summary

Supermicro – Better, Faster, Greener

- One-Stop Shop
- Fastest Time-to-Market
- Customized Service Solutions
- Industry's Broadest Portfolio
- Market Leader in AI & Liquid Cooling
- Engineering Focused...Customer Driven!

Better

Better Performance
Per Watt and Per Dollar

Faster

First-to-Market Innovation with the
Highest Performance Server Designs

Greener

Reduced Environmental Impact
and Lower TCO

SUPERMICR



Visit: [supermicro.com](https://www.supermicro.com)