Supermicro GPU Systems



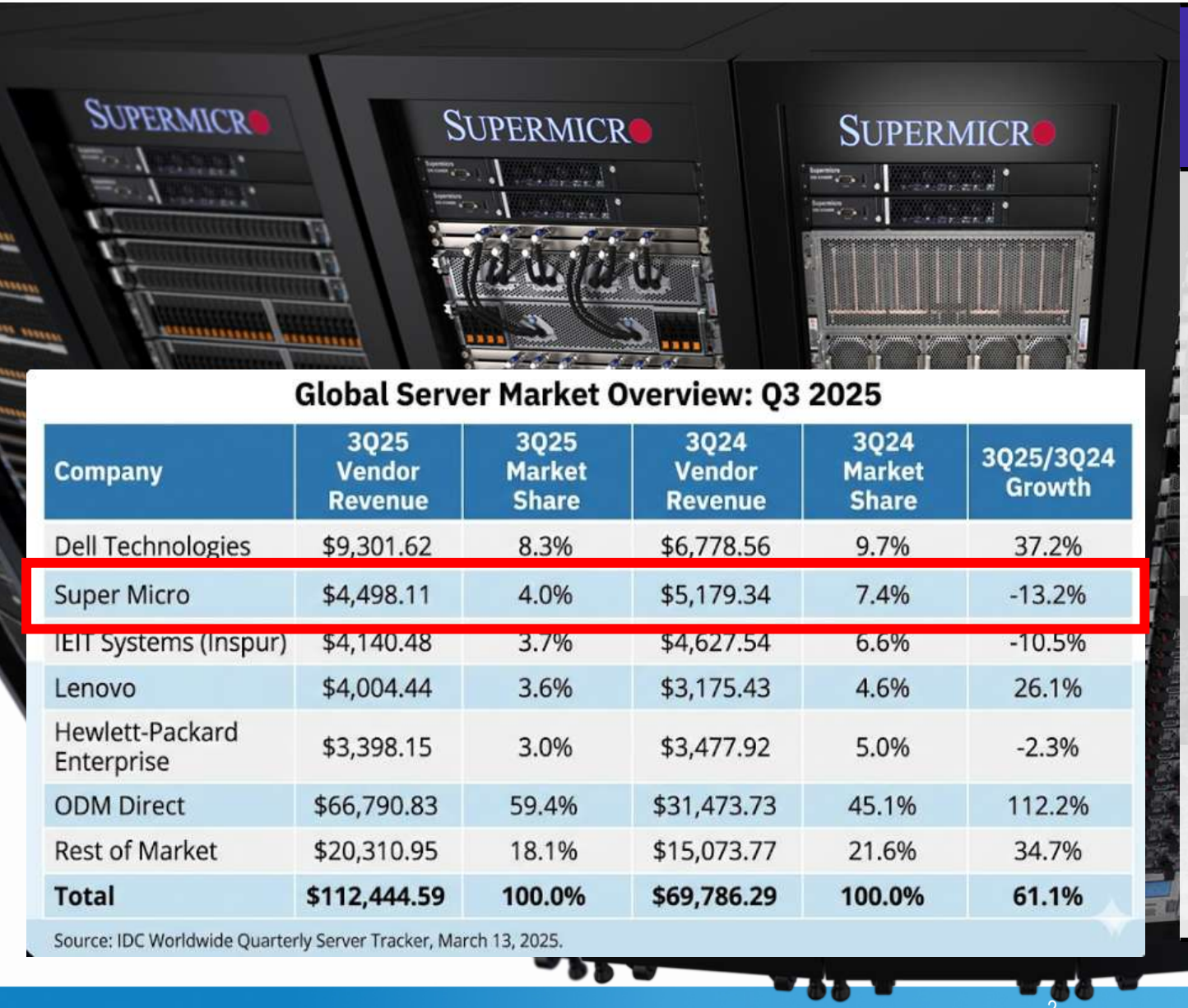
Q1, 2026

슈퍼마이크로코리아
FAE팀 김영남 책임
010-2908-1918
YnK@supermicro.com

Better Faster Greener™ © 2025 Supermicro

ABOUT SUPERMICRO

Confidential



Global Server Market Overview: Q3 2025

Company	3Q25 Vendor Revenue	3Q25 Market Share	3Q24 Vendor Revenue	3Q24 Market Share	3Q25/3Q24 Growth
Dell Technologies	$9,301.62	8.3%	$6,778.56	9.7%	37.2%
Super Micro	$4,498.11	4.0%	$5,179.34	7.4%	-13.2%
IEIT Systems (Inspur)	$4,140.48	3.7%	$4,627.54	6.6%	-10.5%
Lenovo	$4,004.44	3.6%	$3,175.43	4.6%	26.1%
Hewlett-Packard Enterprise	$3,398.15	3.0%	$3,477.92	5.0%	-2.3%
ODM Direct	$66,790.83	59.4%	$31,473.73	45.1%	112.2%
Rest of Market	$20,310.95	18.1%	$15,073.77	21.6%	34.7%
Total	**$112,444.59**	**100.0%**	**$69,786.29**	**100.0%**	**61.1%**

Source: IDC Worldwide Quarterly Server Tracker, March 13, 2025.

Revenue	**$33B+** (FY2026 Guidance) **$22B** (FY2025) $14.9B (FY2024)
Worldwide Presence	6M+ Sq ft. Facilities Worldwide 1. Silicon Valley (HQ), 2. Taiwan, 3. The Netherlands, 4. Malaysia and others
Production	$50B/yr Production Capacity ~5000 Air-cooled Server Racks/mo. ~3000 Liquid-cooled Server Rack/mo.
Human Resource in 4 Campuses	7000+ headcount Worldwide, ~50% Technical / R&D
Key Growth Matrix	**#1** in Generative AI and LLM Platforms 200%+ YoY Growth in Accel. Computing

Accelerate Everything © 2025 Supermicro

DATACENTER BUILDING BLOCK SOLUTIONS (DCBBS)

Confidential

인프라 솔루션

CDU, Water tower, Dry-towers, Chilled doors, A/C, piping, modular containers and more…

소프트웨어 솔루션

DC management, control disaggregated infrastructure, security, optimized TCO



랙 스케일 PnP 솔루션

application optimized system building blocks
GPU/CPU/Storage

네트워크 솔루션 Solutions

Switching, routing, cabling, virtualization, network management and security

Accelerate Everything © 2025 Supermicro

Supermicro GPU Systems



Better Faster Greener™ © 2025 Supermicro

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® 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

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

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+2) 6600W Titanium level power supplies

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

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 5th/4th Xeon *Max 64C, 350W, 320MB per CPU*	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) *Max 128C, 500W, 504MB per CPU*	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*) *Front IO*	8U	Dual Xeon 6 6700 /6500 Series (SP) *Max 86C, 350W, 336MB per CPU*	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-A

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 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

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

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 (1|100)

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



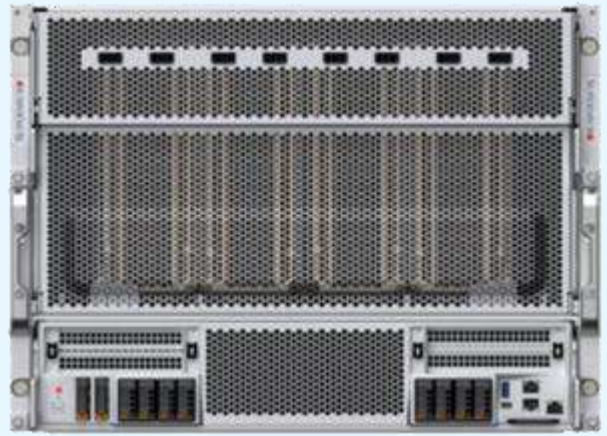
Product Specs

- 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

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 (1 1100)

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

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 (1 l100)

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-NB3OT-ALC

Supermicro's NVIDIA RTX Pro 6000 BSE

Solution Overview - PCIe Optimized Configuration

2U Air-Cooled / SYS-212GB-(F)NR | 2/4-GPU

1-4-3
1-2-3



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 Air-Cooled / SYS-222C-TN | 2-GPU

2-2-3



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 Air-Cooled / SYS-322GA-NR | 8-GPU

2-4-3



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 for2 M.2 PCIe 5.0 x2 NVMe (M-key)

4U Air-Cooled / SYS-422GA-NRT | 8-GPU

2-8-5



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 Air-Cooled / MGX SYS-422GL-NR | 4/8-GPU

2-8-5
2-4-3



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 Air-Cooled / SYS-522GA-NRT or AS -5126GS-TNRT2 | 8-GPU

2-8-5



Intel

AMD

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)
- 10 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 dua 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, or14 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, or14 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

A16-221GL-S3



Product Specs

- CPU – Dual socket dual GH200 Superchips (up to 2000W TDP)
- Drives – 8 Hot-Swap Bays
  - 6 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)

Diversity Products beyond AI GPUs



	1U with Grace Blackwell LC
Model	ARS-121GL-NB2B-LCC
CPU	Dual 72-core Grace Arm Neoverse V2 CPU on Grace Blackwell Superchip
Cooling	Liquid-cooled
GPU Support	Quad NVIDIA B200 GPU with 744GB of HBM3e total
Memory	CPU: 960GB total integrated LPDDR5X with ECC GPU: 744GB total HBM3e
Networking	4 PCIe 5.0 x16 slots or 4-port 800G ConnectX-8 & 2-port 200G BlueField-3
Storage	4-8 Hot-swap E1.S drives and 1 M.2 NVMe drives
Power Supplies	48-54V DC Busbar

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 unparallel 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 E1.S 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

NVIDIA GB200 NVL72 System

Compute for Trillion Parameter Scale AI Workload

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, 13.4TB HBM3e Memory, 1.4 exaFLOPS of AI performance (FP4)

Hyper Dense

72 Blackwell GPUs and 36 Grace™ CPUs in a liquid cooled rack (120kW 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



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

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

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

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

Power Shelves
- 8x 1U 3.3kW (6x 5.5kW PSUs)

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

Product Specs

Per Compute Tray Specs:

- 2x 72-core Grace Arm Neoverse V2
- CPU Memory 480GB LPDDR5X per Superchip (960GB per tray)
- 4x Blackwell B200 GPUs
- GPU Memory 384GB HBM3e per Superchip (768GB per tray)
- 8x 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

SRS-GB200-NVL72-M1

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.



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

Rack



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

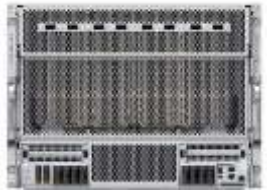


4U Liquid-Cooled DLC-2 NVIDIA HGX B300 System



- 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 NVL144	GBx00 NVL72	HGX B200/B300				NVIDIA HGX H200	
								
kW Per Rack	**~180-220** kW	**125-136** kW	**245** kW	**170** kW	**120** kW	**60** kW	**80** kW	**45** kW
GPU Per Rack	**72x** Rubin	**72x** B200/300	**144x** B300	**96x** B200/300	**64x** B200	**32**x B200	**64**x H200	**32x** H200
Rack Height	**48**U	**48**U	**48**OU/52U	**52**U	**48**U	**48**U	**48**U	**42**U
Rack Weight	**TBD**	**~3780 lbs**	**TBD**	**~3500 lbs**	**~2700 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 Systems supporting 8 MI325X or 8 MI350X 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 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

2 CPU



Product Specs

- Processors - Dual AMD EPYC™ 9005/9004 series processors, up to 500W
- GPU - High density 8U system with AMD Instinct™ MI325X 8-GPU
- Memory - 24 DIMM slots, DDR5-6000 (1DPC); up to 9TB
- 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

SUPERMICRO

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 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

2 CPU



Product Specs

- Processors - Dual AMD EPYC™ 9005/9004 series processors, up to 500W
- GPU - High density 8U system with AMD Instinct™ MI325X 8-GPU
- Memory - 24 DIMM slots, DDR5-6000 (1DPC); up to 9TB
- 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

SUPERMICRO

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

4U H13 Server

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

SUPERMICRO

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

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.

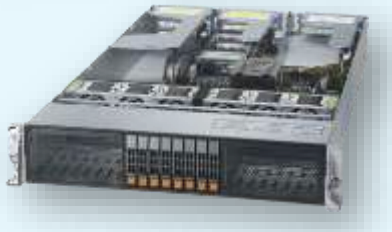


SYS-741GE-TNRT



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

SMC Super AI Station System GB300 Overview

Confidential

ARS-511GD-NB-LCC



HPM—Host Processor Motherboard

Specifications

Processor Support
- NVIDIA Grace™ CPU(72 Core Neoverse V2) with NVIDIA Blackwell™ Ultra GPU

Operating Systems
- DGX OS / Ubuntu 24.04+

Memory Capacity (784GB Coherence Memory)
- CPU:4x 128GB SOCAMM LPDD5X Modules (**Total 496GB System Memory, 396GB/s**)
- GPU:**288 GB HBM3e |8TB/s**

PCIe Expansion
- 1x PCIe 5.0 x16 Half Height Half Length (HHHL) PCIe slots
- 2x PCIe 5.0 x8 Full Height Half Length (FHHL) PCIe slots

Networking & I/O

Rear -
- 2x 400G QSFP Breakout from On-Board **ConnectX®-8 SuperNIC (up to 800Gb/s)**
- 1x 10GbE RJ45, 1x 1GbE IPMI
- 4x USB Type-A, 1x mini-DP , 1x USB Type-B
- 1x Audio Port

Front -
- 2x USB Type-A, 2x USB Type-C, 1x Front USB 2.0 Reserved for PCIE Wifi/BT Card
- 1x Audio Port

Drive Bays
- 2x M.2 2280 for OS RAID (DGX OS/Ubuntu 24.04+)
- 2x M.2 2280 for Training Data SSD

System Cooling
- Direct Liquid Cooling

Power Supply
- ATX PSU 1600W

Dimension
- W 17.93” (455.4 mm) x H 8.63” (219.22 mm) x D 27.57” (700.23 mm)

AI Models
- Support up to 1Trillion Parameter Models
- Multi-Instance GPU(MIG): Up to 7

FP4 Flops
- 20 Petaflop

NVIDIA RTX PRO 6000 Blackwell Series

Confidential

Max-Q Workstation Edition



Up to 300 W

Active

- 300 W blower shroud slots four GPUs per case, mirroring Ada cards to balance airflow, noise, and heat in tight builds.
- Optimized for multi-GPU workstations, trading wattage for cooler and quieter scaling across data-heavy pipelines.
- Excels at large-scale data modeling, distributed rendering, or multi-instance AI training and inference.

Workstation Edition

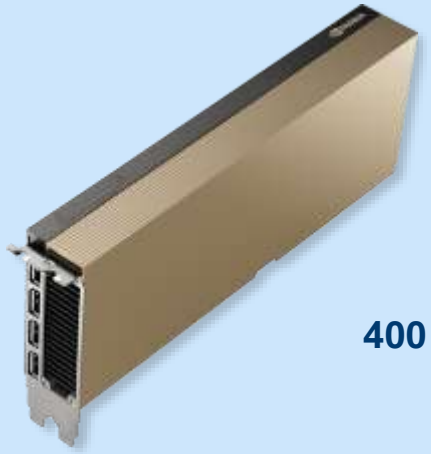


Up to 600 W

Double-flow-through

- Dual-fan flow-through cooler sustains the 600 W Blackwell GPU in spacious ATX towers during marathon workloads.
- Engineered for peak single-GPU throughput, delivering full compute for AI dev, ray-tracing, and CAD workflows.
- Ideal for 3D renders, AI prototyping, and real-time physics sims where stable boost clocks outweigh rack density.

Server Edition



400 to 600 W

Passive

- Passive heatsink taps chassis airflow so 400–600 W cards pack densely, boosting rack efficiency and lowering PUE.
- Most powerful PCIe Gen5 upgrade for L40/L40S/A40 slots—built for air-cooled racks, & up to 8-GPU scale-out servers.
- Powers enterprise inference, model fine-tuning, HPC, virtual desktops, and container-scale distributed graphics.

Accelerate Everything © 2025 Supermicro

Supermicro's NVIDIA RTX Pro 6000 BSE

Solution Overview - PCIe Optimized Configuration

2U Air-Cooled / SYS-212GB-(F)NR | 2/4-GPU

1-4-3
1-2-3



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 Air-Cooled / SYS-222C-TN | 2-GPU

2-2-3



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 Air-Cooled / SYS-322GA-NR | 8-GPU

2-4-3



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 for2 M.2 PCIe 5.0 x2 NVMe (M-key)

4U Air-Cooled / SYS-422GA-NRT | 8-GPU

2-8-5



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 Air-Cooled / MGX SYS-422GL-NR | 4/8-GPU

2-8-5
2-4-3



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 Air-Cooled / SYS-522GA-NRT or AS -5126GS-TNRT2 | 8-GPU

2-8-5



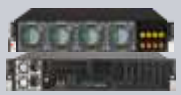
Intel

AMD

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

Edge & Enterprise AI Server Product Portfolio

Product	Model & Spec	Key Features	Max NVIDIA GPU Support
 Xeon 4U	**SYS-422GA-NRT** Sample: Now MP: Jan '26 - Dual Xeon 6900, 500W TDP - DDR5 6400 RDIMM or 8800 MCR x 24, max 6TB - 13 PCIe x16 + 1 AIOM - 8 E3.S/L 1T, 3.2KW PSU x 4, D = 737mm, up to 35°C op temp	- Supports eight 600W GPUs with NVLink - Dual-root structure to support up to 8 high end GPUs & 6 NICs - 4U short depth for rack space saving	x 8 RTX Pro 6000 BSE or H200
 Xeon 3U	**SYS-322GA-NR** Sample: Now MP: Already - Dual Xeon 6900, 500W TDP - DDR5 6400 RDIMM or 8800 MCR x 24, max 6TB - Max 8 PCIe x16 or 16 PCIe x8 (Rear) + 2 PCIe x16 or 4 PCIe x8 (Front) - Max 14 E1.S or 6 U.2, 3.2KW PSU x 3, D = 800mm, up to 35°C op temp	- Supports for eight 350W GPUs - Direct-attach structure brings the best C/P ratio - 3U height for rack space saving	x 8 RTX Pro 6000 Max-Q/L40S
 Xeon 2U	**SYS-212GB-(F)NR** Sample: Now MP: Jan '26 - Single Xeon 6700, 350W TDP - DDR5 5200/6400 RDIMM x 16, max 2TB/1TB - 7 PCIe x16 - 4 E1.S, 2KW PSU x 3 (None-F) or 4 (F SKU), D = 900mm, up to 35°C op temp	- Supports four (F SKU) 600W with NVLink or four 350W GPUs - One CPU supports four GPUs bringing the best C/P ratio - 2U height for rack space saving	x 4 or x 4 H200/RTX Pro 6000 BSE / RTX Pro 6000 Max-Q/L40S
 EPYC 2U	**AS -2115HE-(F)TNR** Sample: Now MP: Jan '26 - Single EPYC 9005, 400W TDP - DDR5 4400/5200 RDIMM x 24, max 6TB/3TB - Max 4 PCIe x16 or 8 PCIe x8 - Max 10 U.2, 2KW PSU x 2, D = 574mm, up to 35°C op temp	- Support for three 350W GPUs - 12-ch DDR5 up to 3TB 5200MT/s or 6TB 4400MT/s - 2U short depth for easy integration in racks/cabinets	x 3 RTX Pro 6000 Max-Q/L40S
 Xeon 2U	**SYS-222HE-(F)TN** Sample: Now MP: Already - Dual Xeon 6700, 350W TDP - DDR5 5200/6400 RDIMM x 24, max 8TB/4TB - Max 4 PCIe x16 or 8 PCIe x8 - Max 10 U.2, 2KW PSU x 2, D = 574mm, up to 35°C op temp	- Support for three 350W GPUs - 8-ch DDR5 up to 4TB 6400MT/s or 8TB 5200MT/s - 2U short depth for easy integration in racks/cabinets	x 3 RTX Pro 6000 Max-Q/L40S
 Xeon 2U	**SYS-212B-(F)N2T** Sample: Now MP: Dec '25 - Single Xeon 6700, 350W TDP - DDR5 6400 RDIMM or 8000 MCR x 8, max 2TB/512GB - Max 4 PCIe x16, 1 PCIe x8 - Max 4 U.2, 2KW PSU x 2, D = 450mm, up to 40°C op temp	- Support for two 350W GPUs - 2U ultra short for easy integration in racks/cabinets - Up to 40°C op temp for harsh environment deployment	x 2 RTX Pro 6000 Max-Q/L40S
 Xeon 2U	**SYS-212B-(F)LN2T** Sample: Now MP: Dec '25 - Single Xeon 6700, 350W TDP - DDR5 6400 RDIMM, max 2TB - Max 4 PCIe x16, 3 PCIe x8 - Max 4 U.2, 2KW PSU x 2, D = 450mm, up to 40°C op temp	- Support for one 350W or up to 4 L4 GPUs - 2U ultra short for easy integration in racks/cabinets - Up to 40°C op temp for harsh environment deployment	x 1 or x 4 RTX Pro 6000 Max-Q/L40S / L4
 EPYC 2U	**AS -2116S-(F)TNRT** Sample: Now MP: Jan '26 - Single EPYC 9005, 500W TDP - DDR5 6400, RDIMM x 12, max 3TB - Max 4 PCIe x16, 2 PCIe x8 - Max 4 U.2, 2KW PSU x 2, D = 450mm, up to 40°C op temp	- Support for one 350W or up to 4 L4 GPUs - 12-ch DDR5 up to 3TB 6400MT/s - 2U ultra short for easy integration in racks/cabinets - Up to 40°C op temp for harsh environment deployment	x 1 or x 4 RTX Pro 6000 Max-Q/L40S / L4
 Xeon 2U	**SYS-212B-FN4TP** Sample: Now MP: Already - Single Xeon 6700, 350W TDP - DDR5 6400 RDIMM or 8000 MCR x 8, max 2TB/512GB - Max 4 PCIe x16, 1 PCIe x8 - Max 2 U.2, 800W PSU x 2, D = 300mm, up to 50°C op temp	- Support for three L4 GPUs - Front I/O, ultra short for constrained space installation - Up to 50°C op temp for harsh environment deployment	x 3 L4
 Xeon Box	**SYS-E403-14B-FRN2T** Sample: Now MP: Already - Single Xeon 6700, 350W TDP - DDR5 6400, RDIMM x 8, max 2TB - 3 PCIe x16 - 800W PSU x 2, W x H x D = 267 x 117 x 406 mm^3, up to 45°C op temp	- Support for one 350W or up to three L4 GPUs - Industry unique box server for easy integration and flexible mounting - Up to 45°C op temp for harsh environment deployment	x 1 or x 3 RTX Pro 6000 Max-Q/L40S / L4

Confidential

Enterprise Oriented

Embedded/IoT Oriented

Data Center Accelerated Computing Portfolio

Note 1	GB300 NVL72	GB200 NLV72	HGX B300	HGX B200	GB200 NVL4	GH200	H200		H100		RTX Pro 6000 Blackwell Server Edition[1]	L40S	L4
Value Proposition	Highest Perf AI Reasoning, Training and Inf	High Perf LLM Training and Inf, Data Processing	Highest AI Perf HGX	High Perf AI, HPC HGX	Single Server AI + HPC	Versatile AI and HPC Self-Hosted	Mainstream LLM Training, Inference, and HPC	Mid-Size LLM Inference, Fine-Tuning, HPC	Mainstream LLM Training, Inference, and HPC	Mid-Size LLM Inference, Fine-Tuning, HPC	Highest Perf Universal	High Perf Universal	Universal AI, Video, and Graphics
Form Factor	MGX Rack 72-GPU	MGX Rack 72-GPU	HGX 8-SXM	HGX 8-SXM	MGX Server	MGX Server	HGX 8 or 4-SXM	NVL PCIe Gen5 x16 2 Slot FHFL NVLink Bridged	HGX 8 or 4-SXM	NVL PCIe Gen5 x16 2 Slot FHFL NVLink Bridged	PCI Gen5 x16 2 Slot FHFL	PCIe Gen4 x16 2 Slot FHFL	PCIe Gen4 x16 1 slot LP
Single GPU TDP	Configurable up to 1400W	Configurable up to 1200W	Configurable up to 1,100W	Configurable up to 1000W	Configurable up to 1200W	Configurable up to 1000W (GPU+CPU+Memory)	Configurable up to 700W	Configurable up to 600W	Configurable up to 700W	Configurable up to 400W	Configurable up to 600W	Configurable up to 350W	Configurable up to 72W
FP4 TC TFLOPS[2]	15,000 | 20,000	10,000 | 20,000	14,000 | 18,000	9,000 | 18,000	10,000 | 20,000	-	-	-	-	-	1,900 | 3,753	-	-
FP8 TC TFLOPS[2]	10,000	10,000	9,000	9,000	10,000	3,958	3,958	3958	3,958	3,341	1,900	1,466	485
INT8 TC TOPS[2]	330	10,000	307	9,000	10,000	3,958	3,958	3,958	3,958	3,341	NA	1,466	485
FP16 TC TFLOPS[2]	5,000	5,000	4,500	4,500	5000	1979	1979	1979	1979	1,671	950	733	242
TF32 TC TFLOPS[2]	2,500	2,500	2,200	2,200	2,500	989	989	989	989	835	475	366	120
FP32 TFLOPS	80	80	75	75	80	67	67	67	67	60	117	92	30
FP64 TC | FP64 TFLOPS	1.3 | 1.3	40 | 40	1.2 | 1.2	37 | 37	40 | 40	67 | 34	67 | 34	67 | 34	67 | 34	60 | 30	NA	NA	NA
GPU Memory | Bandwidth	279 GB | 8 TB/s	186 GB HBM3e | 8 TB/s	270GB | 7.7 TB/s	186 GB HBM3e | 8 TB/s	180 GB HBM3e | 7.7 TB/s	Up to 144GB HBM3e | Up to 4.9 TB/s	141GB HBM3e | 4.8 TB/s	141GB HBM3e | 4.8 TB/s	80GB HBM3 | 3.35 TB/s	94GB HBM3 | 3.9 TB/s	96GB GDDR7 | 1.6 TB/s	48GB GDDR6 | 864 GB/s	24GB GDDR6 | 300 GB/s
Multi-Instance GPU (MIG)	Up to 7	Up to 7	Up to 7	Up to 7	Up to 7	Up to 7	Up to 7	Up to 7	Up to 7	Up to 7	Up to 4	-	-
NVLink Connectivity	72	72	8	8	4	2	4 or 8	4 cards bridged	4 or 8	2 cards bridged	-	-	-
Media Acceleration	7 JPEG Decoder 7 Video Decoder4	7 JPEG Decoder 7 Video Decoder4	7 JPEG Decoder 7 Video Decoder4	7 JPEG Decoder 7 Video Decoder4	7 JPEG Decoder 7 Video Decoder4	7 JPEG Decoder 7 Video Decoder	7 JPEG Decoder 7 Video Decoder	7 JPEG Decoder 7 Video Decoder	7 JPEG Decoder 7 Video Decoder	7 JPEG Decoder 7 Video Decoder	4 Video Encoder 4 Video Decoder 4 JPEG Decoder	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	2 Video Encoder3 4 Video Decoder3 4 JPEG Decode
Ray Tracing	-					-					4th Gen | 352	3rd Gen | 142	3rd Gen | 60
Transformer Engine	2nd gen					1st Gen					2nd Gen	1st Gen	
Graphics	For in-situ visualization (no NVIDIA vPC or RTX vWS)					For in-situ visualization (no NVIDIA vPC or RTX vWS)					Best	Better	Good
vGPU	Yes					Yes						Yes	
NVIDIA AI Enterprise	Add-on					Add-on	Add-on	Included	Add-on	Included		Add-on	

1. *Specifications are per GPU.*
2. *Specification in dense \| sparse*
3. *Specification in sparse only. Dense is ½ sparse spec shown.*
4. *Supported formats provide these speed-ups over H100 Tensor Core GPUs – 2X H.264, 1.25X HEVC, 1.25X VP9. AV1 support is new to Blackwell GPUs.*

NVIDIA

Nvidia HGX™ Rubin NVL8 and B300 NVL8 Comparison

SPEC	HGX Rubin NVL8	HGX B300 NVL8	Performance Gain
NVFP4 Inference	400 PFLOPS		
NVFP4 Training	280 PFLOPS	144 PFLOPS	⬆
FP8/FP6 Training	140 PFLOPS	72 PFLOPS	
INT8 Tensor Core	2 PFLOPS	2 PFLOPS	-
FP16/BF16 Tensor Core	32 PFLOPS	36 PFLOPS	-
TF32 Tensor Core	16 PFLOPS	18 PFLOPS	-
FP32	1040 TFLOPS	600 TFLOPS	
FP64/ TF64 Tensor Core	264 TFLOPS	10 TFLOPS	⬆
Nvlink GPU-to-GPU Bandwidth	3.6 TB/s	1.8 TB/s	
Total GPU Memory	2.3 TB	2.3 TB	-

© 2025 Supermicro

Confidential

B300 vs Rubin NVL16 Server Reference Design

Parameters	B300 Server Reference	Rubin NVL16 Server Reference
GPU Baseboard	HGX B300 NVL8 SXM6	HGX Rubin NVL16 SXM8
Power Input	DC Busbar / AC C19	DC Busbar
Average Power	DC SKU 14.5KW / AC SKU 15.1KVA [cite: <IMAGE 1>]	24KW
System cooling	Air flow 1350CFM. Inlet 10-35C. Outlet 50C at 35C inlet.	34LPM T.Inlet 45C PG-25
System Memory	2TB (32x 64GB DDR5 6400), configurable to 4TB	2TB (32x 64GB DDR5 6400), configurable to 4TB
Storage	OS 2x M.2 (2TB), Data 8x E1.S (4TB)	OS 2x M.2 (2TB), Data 8x E1.S (4TB)
N/S Networking	2x BF3 (NDR, 400G)	2x BF3 (NDR, 400G)
E/W Networking	8x OSFP (IB XDR, ETH 2x400G) from CX-8 on HGX B300 NVL8 cable	8x OSFP (IB XDR, ETH 1x80G) from CX-9 on HGX Rubin NVL16 onboard
Management	Pluggable Nvidia DC-SCM	Pluggable Nvidia DC-SCM
Cable	Cold aisle	Cold aisle
Chassis	10RU MGX Rack compatible	2RU MGX Rack compatible
System Dimensions	441.75 x 448 x 800 mm (H x W x L)	87 x 448 x 800 mm (Hx W x L)
System weight	DC SKU 123KG / AC SKU 168KG	55KG

NVIDIA CONFIDENTIAL

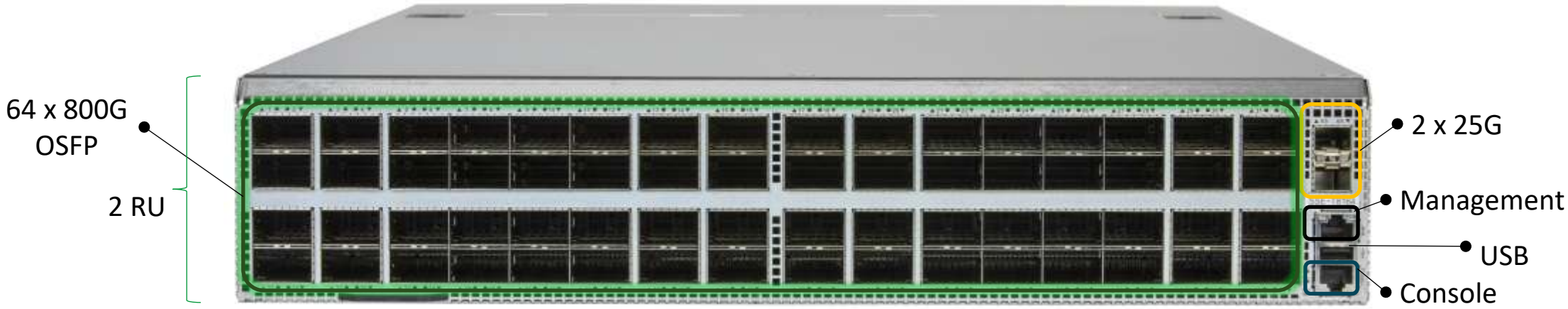
Supermicro Network



Better Faster Greener™ © 2025 Supermicro

Key Use Cases: GPU Cluster Spine and Leaf, 400G/800G leaf for CX7/CX8/BF3/Gaudi3/ Vulcano/Thor Ultra, Storage at 400G/800G, High Density 800G/400G/200G

Supermicro 800G Switch - SSE-T8164



Hardware Specifications

Ideal for AI Fabric 800G Spine (or Leaf via Breakouts)

- ✓ 64x800G OSFP & 2 x 25G SFP+ Ports
- ✓ 51.2T Switch Fabric
- ✓ 1x1G dedicated Management Port
- ✓ 1+1 hot swap AC / DC power options
- ✓ N+1 hot swap fans
- ✓ Standard & Reverse Airflow Models
- ✓ DAC 1:2 cable at 4m to AOC-S400G-B1C
- ✓ Direct Liquid Cooling option (Q2CY26)

Supermicro Enterprise Advanced SONiC OS

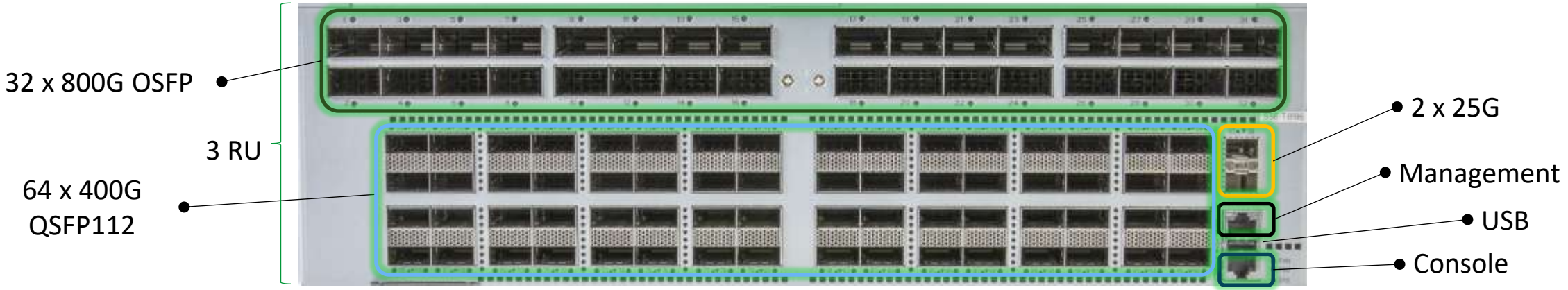
Distributed & Supported by Supermicro

- Hardened, Optimized, Cable/TRX assuranceRoCEv2, Advanced QoS, PFC, DCQCN, L3 VXLAN, MP-BGP EVPN, ZTP, VRF, many others in Industry Standard CLI
- Enhanced Ethernet Features for AI with switch fabric (DLB and GLB, Enhanced Telemetry, Mod16 Hashing, others)

Better Faster Greener™ © 2026 Supermicro

Key Use Cases: GPU Cluster Leaf, 400G leaf for CX7/BF3/Pollara/Thor2, Reduced/Eliminated Cable breakouts, Storage at 400G, High Density 800G/400G/200G

Supermicro 400/800G Switch - SSE-T8196



Hardware Specifications

Ideal for AI 400G Leaf Use Cases to 800G Spines

- ✓ 32x800G OSFP Ports & 64x400G QSFP112 & 2 x 25G SFP+ Ports
- ✓ 51.2T Switch Fabric
- ✓ 1x1G dedicated Management Port
- ✓ 1+1 hot swap AC / DC power options
- ✓ Standard & Reverse Airflow Models
- ✓ DAC cable at 4m to AOC-S400G-B1C
- ✓ Direct Liquid Cooling option (FCS+6)

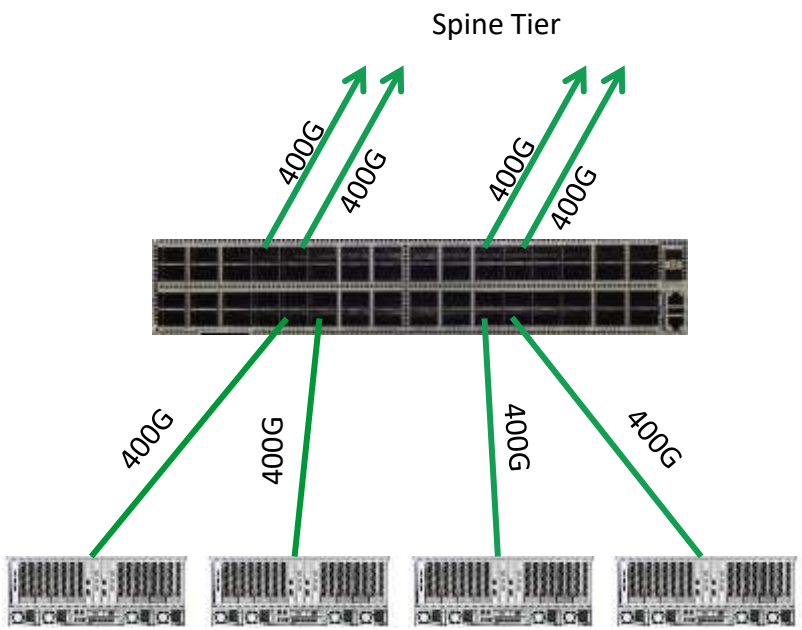
Supermicro Enterprise Advanced SONiC OS

Distributed & Supported by Supermicro

- Hardened, Optimized, Cable/TRX assurance
- RoCEv2, Advanced QoS, PFC, DCQCN, L3 VXLAN, MP-BGP EVPN, ZTP, VRF, many others in Industry Standard CLI
- Enhanced Ethernet Features for AI with switch fabric (DLB and GLB, Fast Link FO, Reactive Path Rebalancing, Enhanced Telemetry, Mod16 Hashing, others)

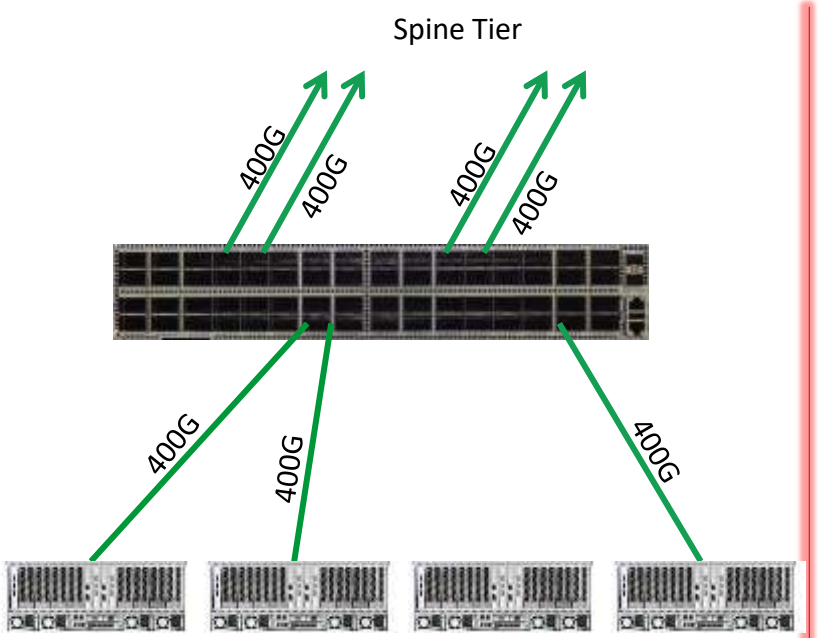
Better Faster Greener™ © 2026 Supermicro

Handling Cluster Hashing and DLB Transients



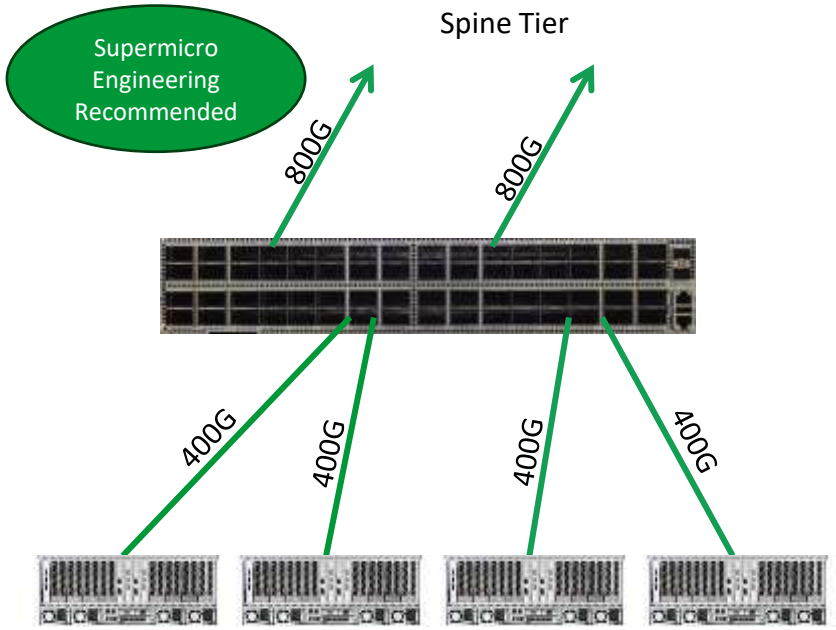
Traditional Non-Blocking Flows

- Steady-state flow multipathing spreads flows on spine links
- New flow creation and transient situations can place multiple 400G flows onto single uplink temporarily
- DCQCN and/or PFC mechanisms used to back pressure for no packet loss



Undersubscribing Flows

- Steady-state flow multipathing spreads flows on spine links
- New flow creation and transient situations can place multiple node 400G onto single uplink temporarily – but some are lightly loaded
- Delayed or avoided congestion impacts depending on transient before DCQCN and/or PFC employed
- Cost is empty leaf ports and fewer nodes under leafs and need for 3rd tier sooner



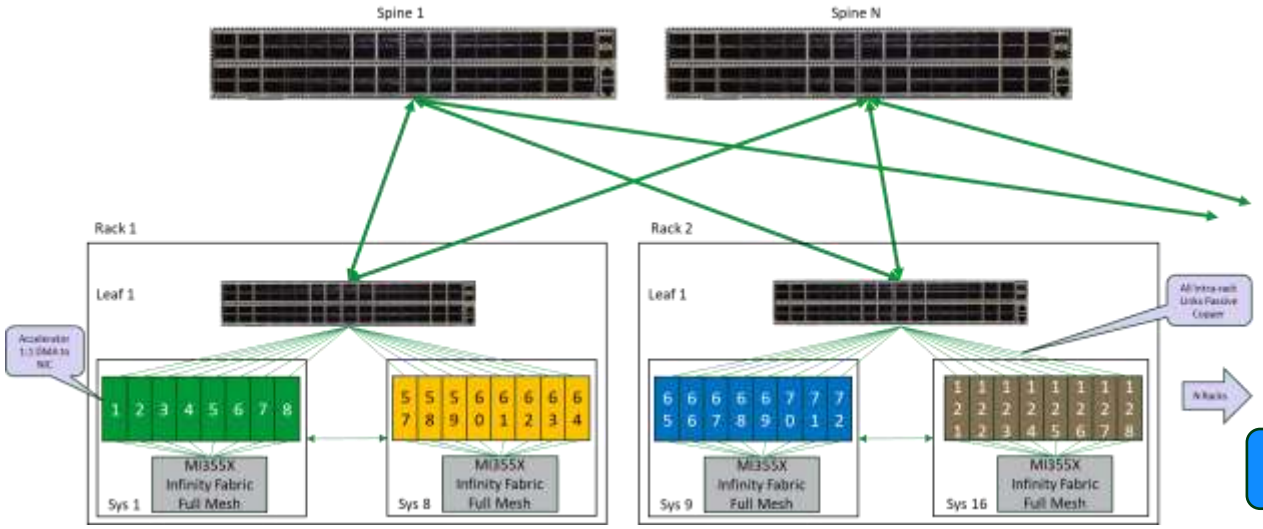
Alternative for Maximum Efficiency

- Steady-state flow multipathing spreads flows on spine links
- New flow creation and transient situations can place multiple node 400G onto single 800G uplink temporarily
- Fewer larger links means lower probability of any transient overloading given link
- Delayed or avoided congestion impacts depending on transient before DCQCN and/or PFC employed

Configuring Slurm for Optimal Scheduling

An efficient cabling alternative to a rail-optimized design

- Scheduling on local high bandwidth domain always most performant (i.e. Infinity Fabric), often need to scale higher
- Uncoordinated network will lead to any to any elephant flows over various leafs/spines/superspines with widely different performance
- Slurm 24.05 with the topology/block plugin allows for hierarchical scheduling in these jobs
- Keeps the advantages of scheduler knowing resource adjacency to minimize elephant flow scale-out switching hops
  - Same goal as rail-optimized designs – but alternative provides direct control vs. static local rank
  - Example shows 2 DLC racks, simple scale up to 8 DLC racks (64 systems / 512 MI325X) with 8 leafs/4 spines
- Slurm sbatch command arguments to control:
  - --exclusive=topo Means no other jobs are place on the allocated block
  - --segment=YY YY is a number less than planning block size, to parse over many blocks (so flows could need to cross higher level network tiers)
  - --NXX XX is the number of workers collectively executing this step of the job



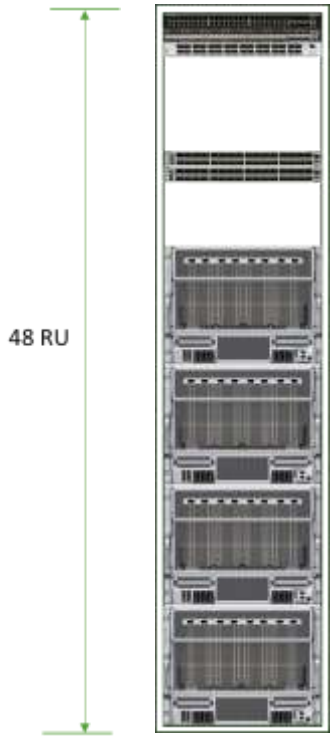
Example Slurm Topology File for topology/block plugin:

```
# Sample Network Topology File for Block Plugin
BlockName=System1 Nodes=node[01-08]
BlockName=System2 Nodes=node[09-16]
…
BlockName=System16 Nodes=node[121-128]
BlockSizes=8,64,8192,65536
```

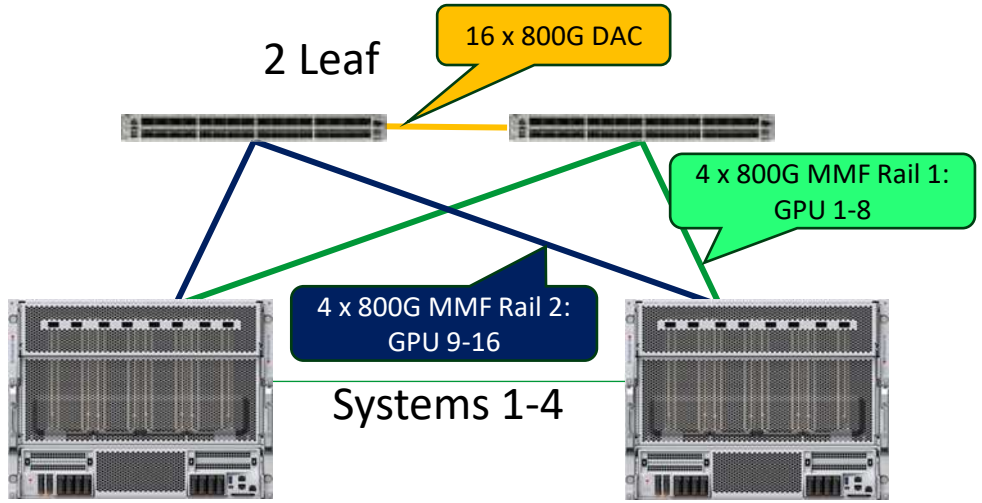
- HBD in System = planning block size
- 1st Network Tier (Leaf)
- Max 2 tier spine-leaf (if switch radix=64)
- Example 3 tier size (possible to go higher)

Physical POD Spacing and Cabling – 64 Blackwell

SMC Reference Architecture for 4 B300 NVL8 in ~60kW Racks



Cold Aisle View

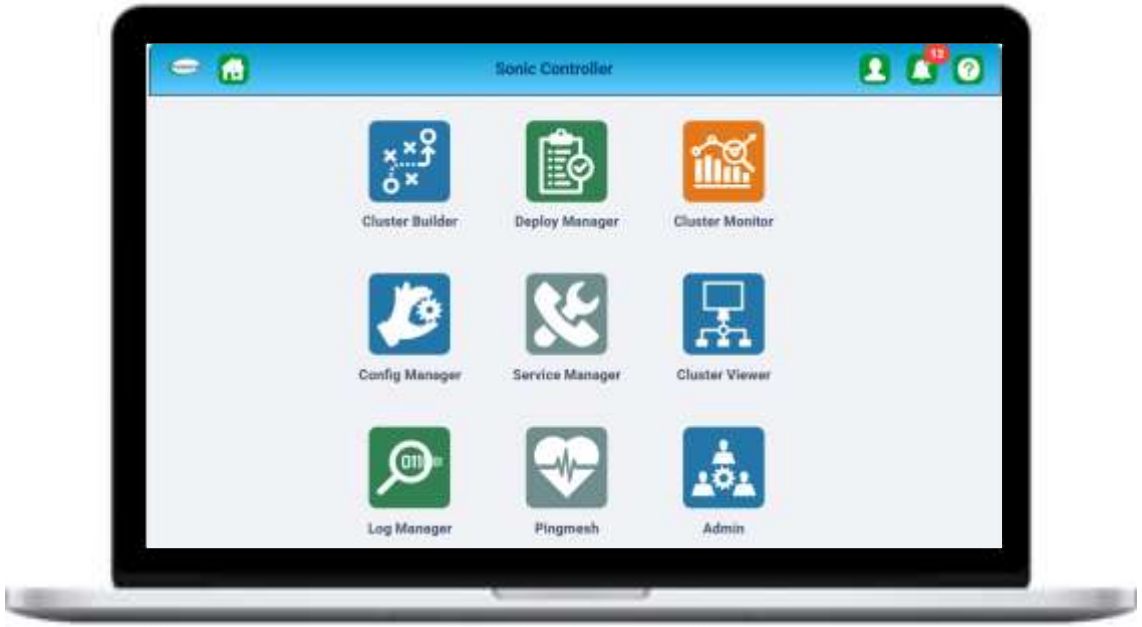


- POD is 1 rack
- 4 Systems per rack
  - 8 x 800G CX8 integrated in each system
- Each rack has 2 leafs (SSE-T8032S) for rails 1-2

- Rack also has 1 48x1G mgmt switch (SSE-C3748S), 1 48x25G inband switch (SSE-F3548S), 1 32x100G storage switch (SSE-C4632S)
- 5m MMF Fiber to nodes from 2 leafs
- 2m DAC 800G Cables leaf to leaf

Simplifying Cluster Architecture and Operations

Supermicro Cluster Orchestration and Validated Designs



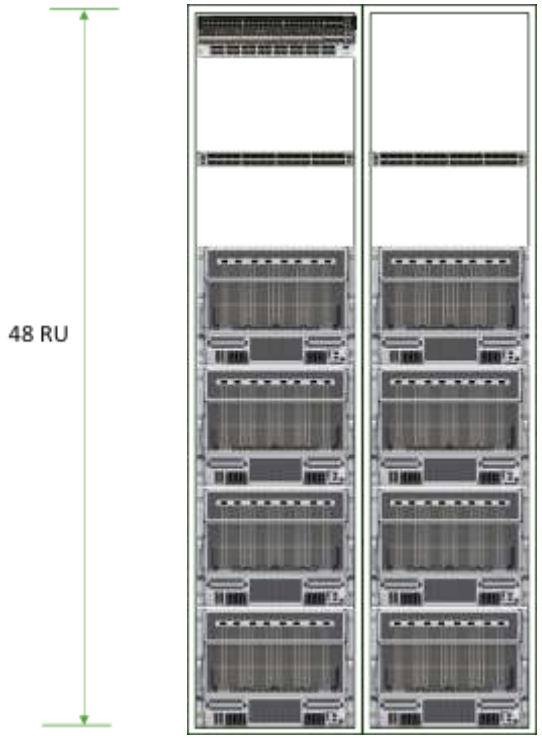
- Cluster Design and BOM Building
- Implement Validated Performant Designs
- All GPU and Accelerator Vendors Supported
- Zero Touch deployment, connectivity validation
- Operations, Telemetry, Performance, and Multi-Tenant Management


SUPERMICRO
BROADCOM
AMD
BUILDING TENANT AWARE MI325X ACCELERATOR CLUSTER WITH AN INTEGRATED SUPERMICRO SOLUTION
April 2025
Table of Contents
Executive Summary 2
Glossary of Terms 2
Foundations of AI Fabrics: RDMA, PCIe Switching, Ethernet, IP, and BGP 3
Validated Design Equipment and Configuration 8
Scaling out the Accelerators with an optimized Ethernet Fabric – Components and Configurations 10
Design of the Scale Unit - Scaling out the cluster 11
Supermicro Validated MI325 Design: 4 Scale Units for 256 MI325X 15
How to Minimize Deployment Time – L12 Rack Pre-Built Solution from Supermicro 22
Summary 23
Appendix A: Accelerator Fabric Detailed Leaf and Spine Configuration Steps 24

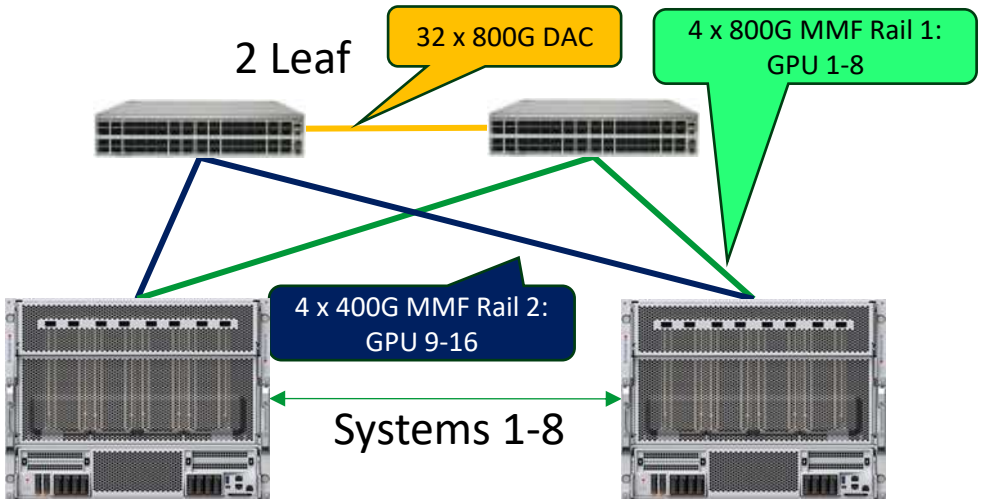
- Joint partner work on validating and templating performant clusters (GPU, Server, NIC, Switch, Storage)
- Simpler support and operation structures
- Multiple Supermicro Validated Designs available

Physical POD Spacing and Cabling – 128 Blackwell

SMC Reference Architecture for 8 B300 NVL8 in ~60kW Racks



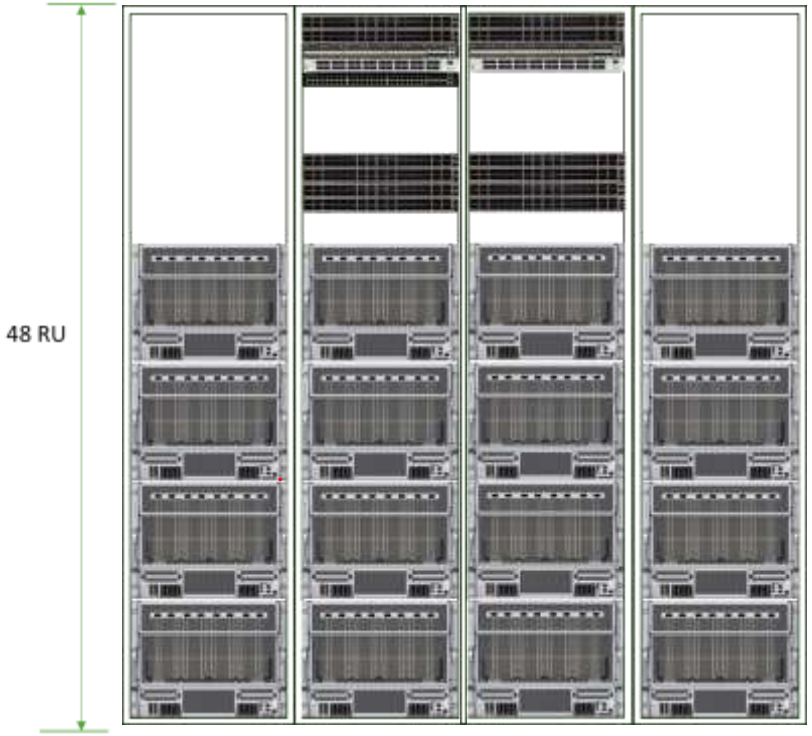
Cold Aisle View



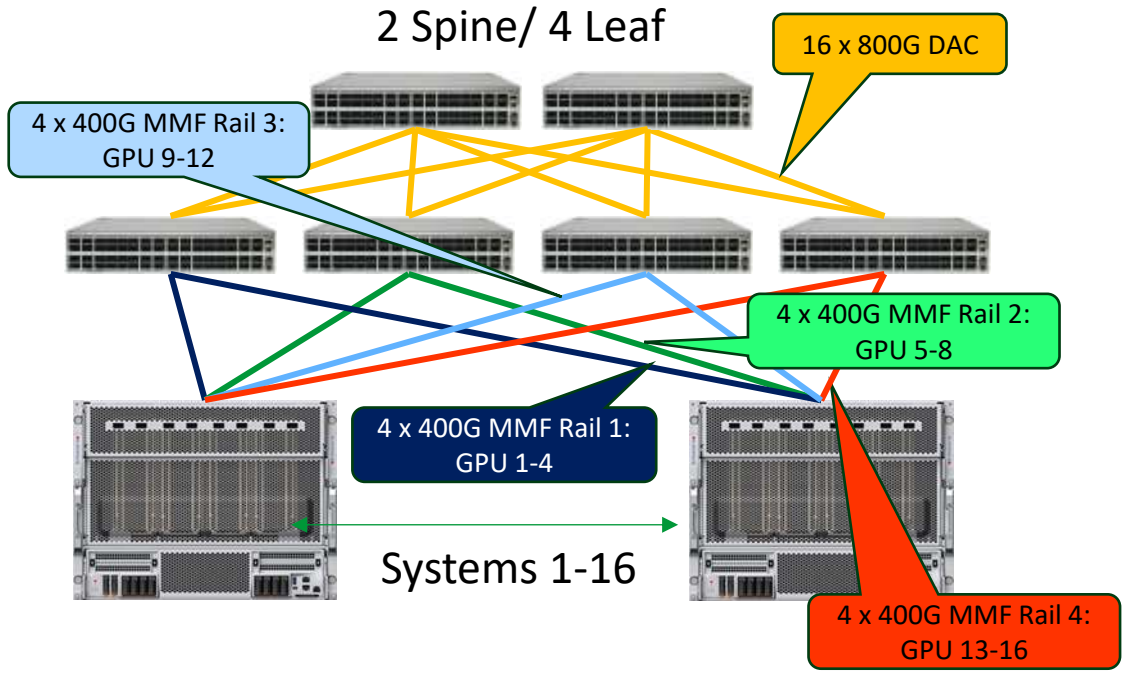
- POD is 2 racks
- 4 Systems per rack
  - 8 x 800G CX8 integrated in each system
- Each rack has 1 leaf (SSE-T8164S with SFT-BCM800G-5YR) for rails 1-2
- Racks share 1 48x1G mgmt switch (SSE-C3748S), 1 48x25G inband switch (SSE-F3548S), 1 32x100G storage switch (SSE-C4632S)
- 5m MMF Fiber to nodes from 2 leafs
- 2m DAC 800G Cables leaf to leaf

Physical POD Spacing and Cabling – 256 Blackwell

SMC Reference Architecture for 16 B300 NVL8 in ~60kW Racks

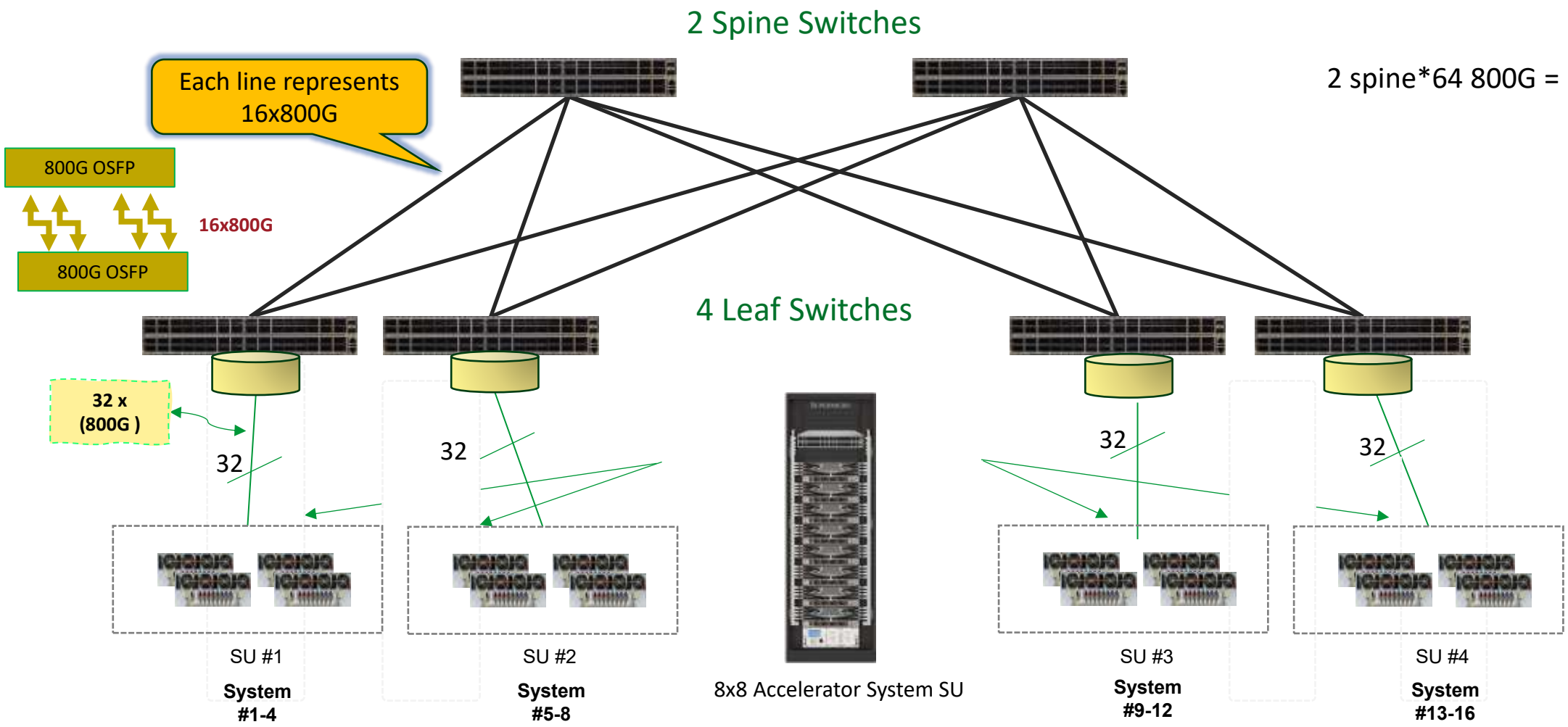


Cold Aisle View



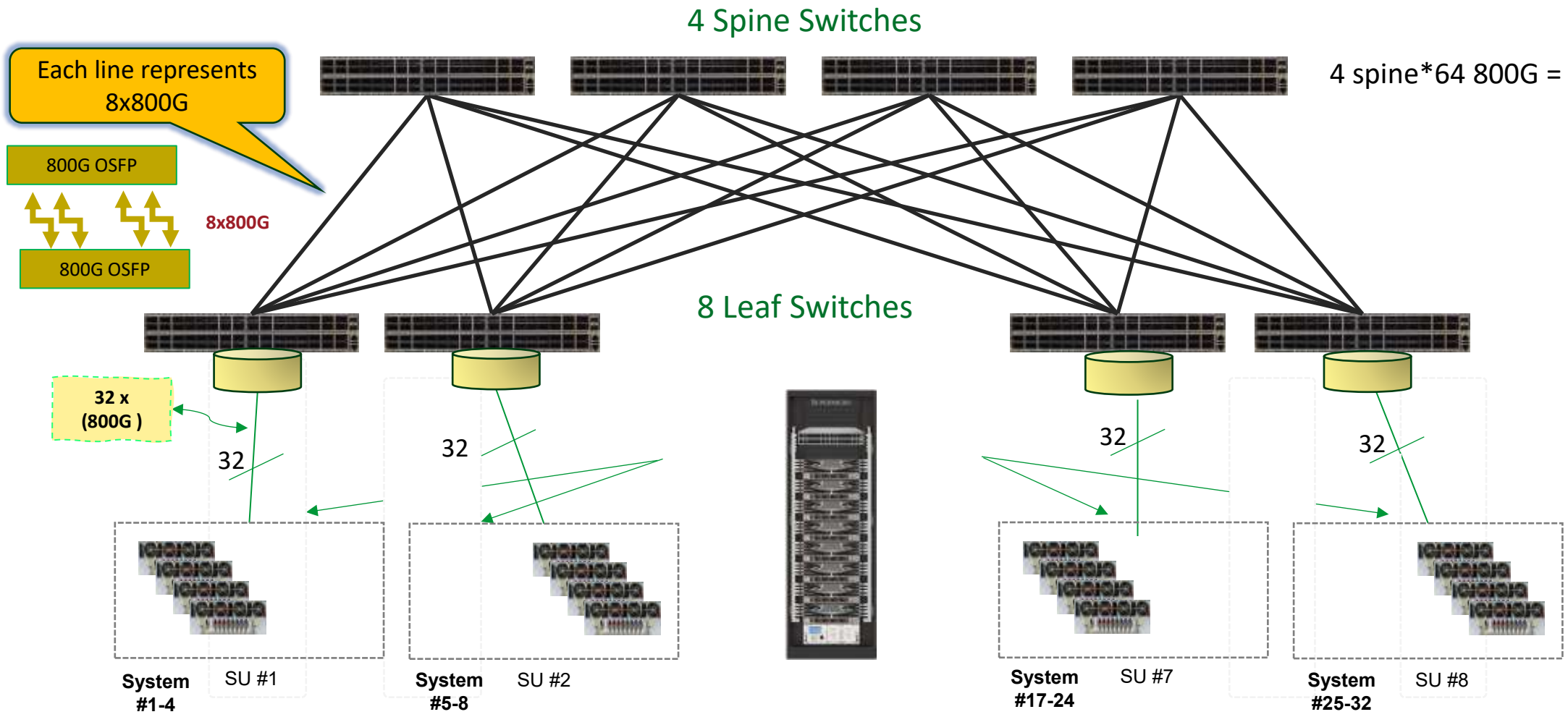
- POD is 4 racks
- 4 Systems per rack
  - 8 x 800G CX8 integrated in each system
- End Racks have only B300 NVL8 Systems
- Middle 4 racks have 2 leafs (SSE-T8164S with SFT-BCM800G-5YR) each for rails 1-4
- Middle 4 racks also have 1 48x1G mgmt switch (SSE-C3748S), 2 48x25G inband switch (SSE-F3548S), 2 32x100G storage switch (SSE-C4632S)
- 2 Spines in center racks
- 5m MMF Fiber to nodes from 4 leafs
- 2m DAC 800G Cables leaf to spine

16 B300 Systems (128 -16x8 B300 GPU)



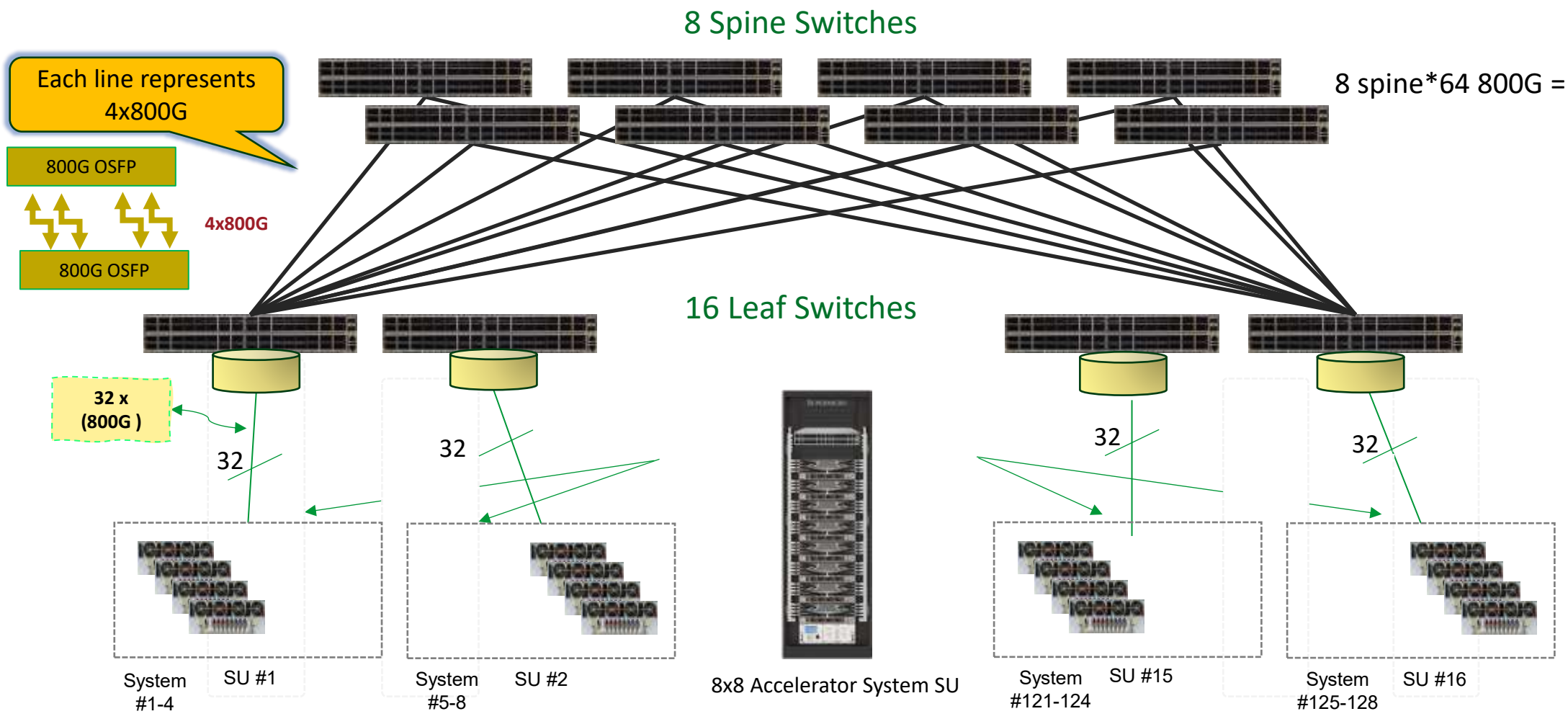
32 800G ports towards systems/leaf * 4 leafs = 128 800G to OSFP on CX8 NIC

32 B300 Systems (256 -32x8 B300 GPU)



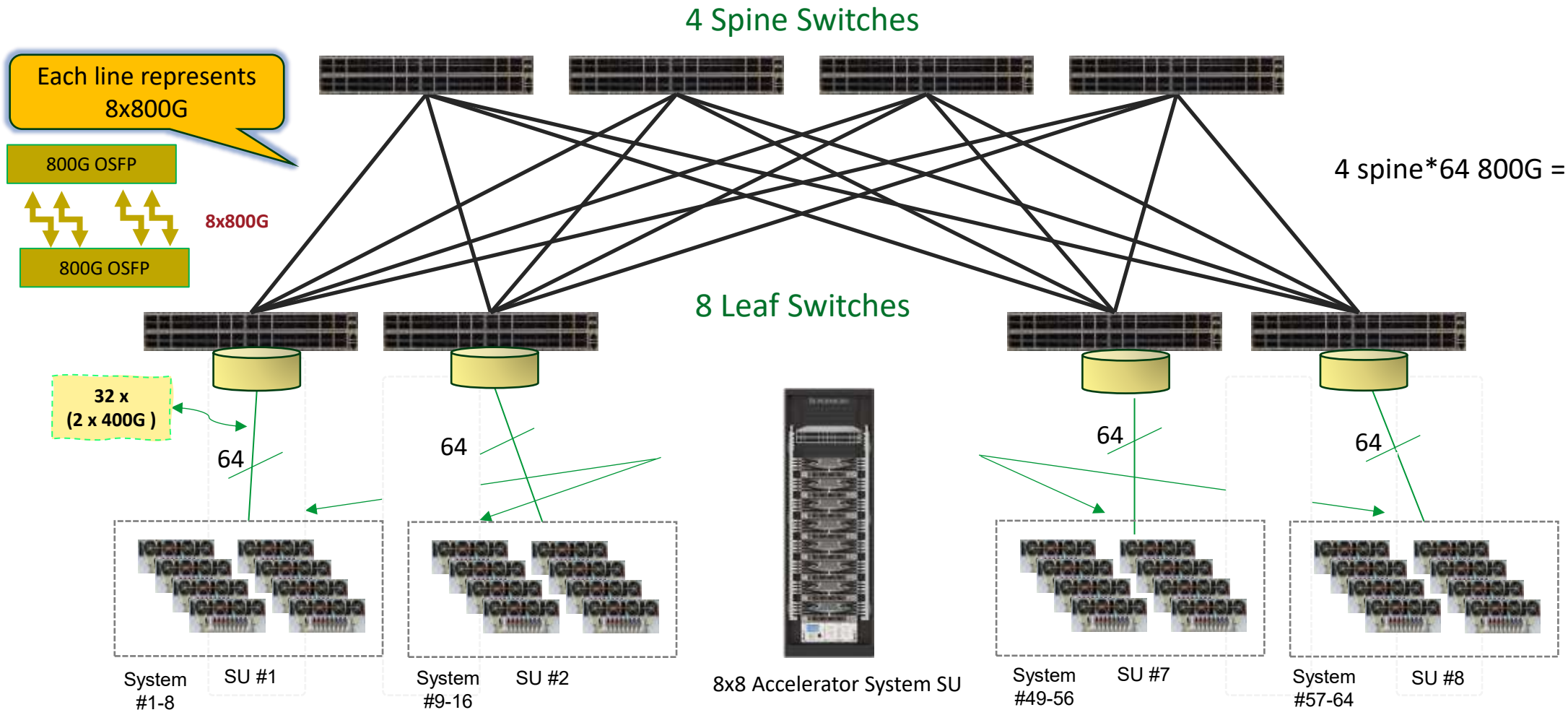
32 800G ports towards systems/leaf * 8 leafs = 256 800G to OSFP on CX8 NIC

64 B300 Systems (512 -64x8 B300 GPU)



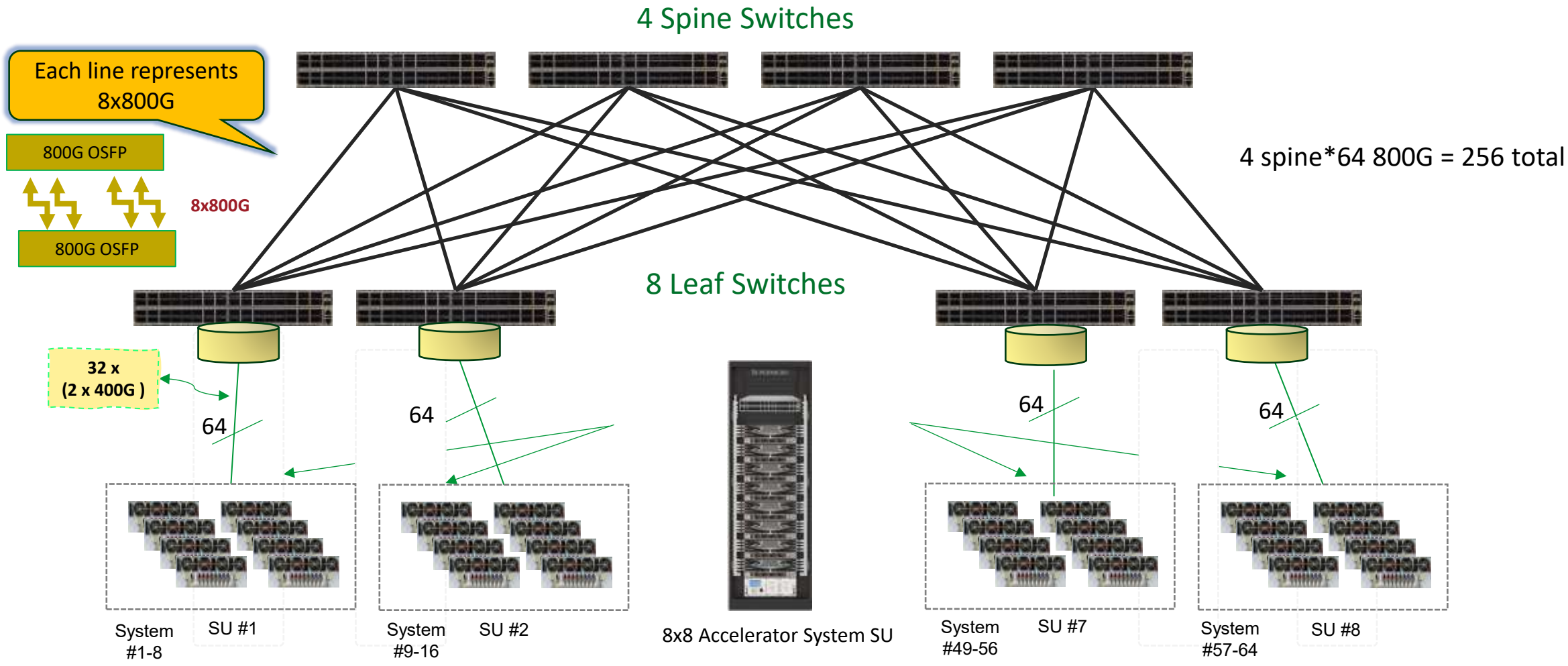
64 800G ports towards systems/leaf * 16 leafs = 512 800G to OSFP on CX8 NIC

64 AMD instinct MI355X Systems (512 Accelerators)



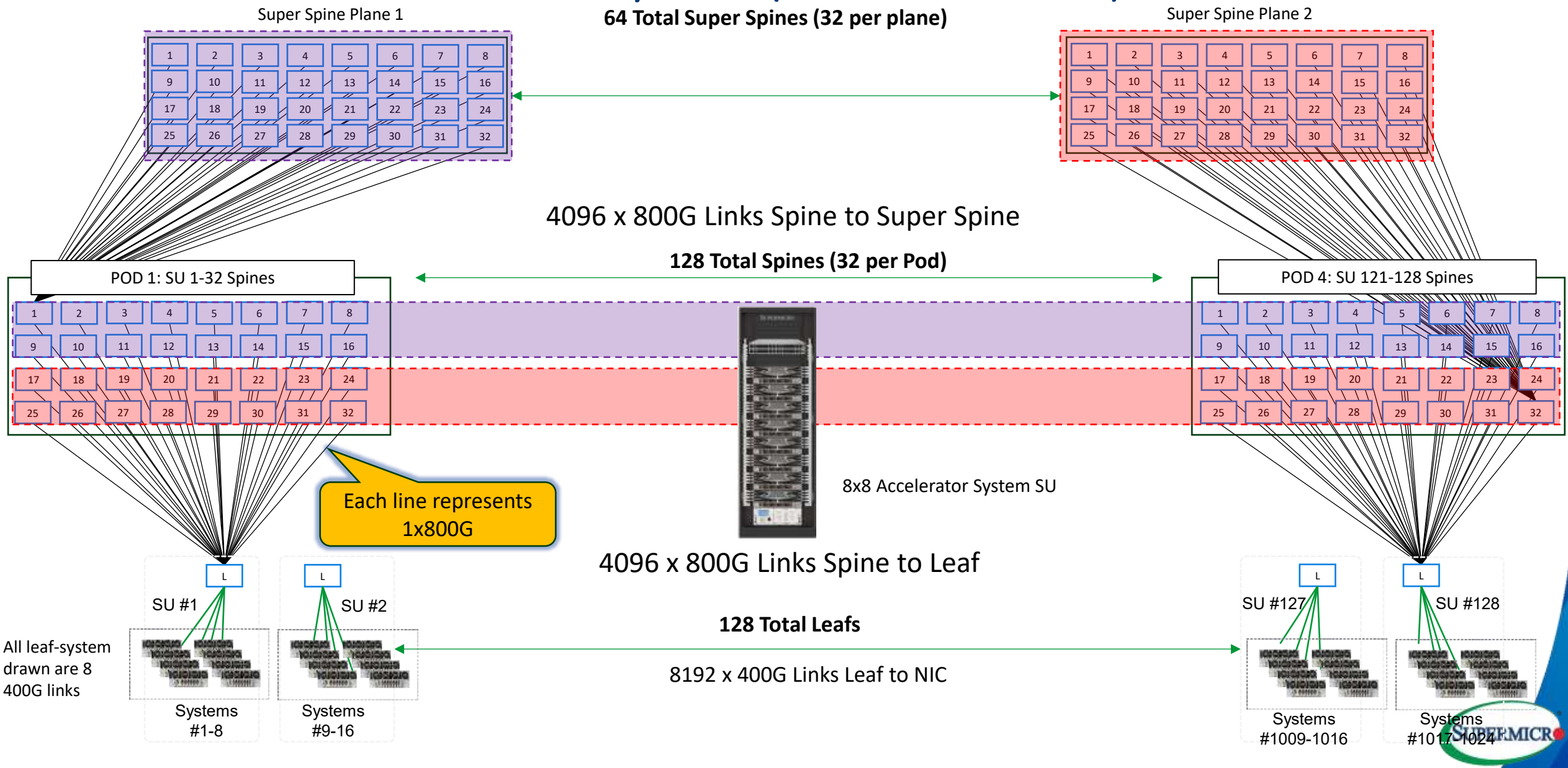
64 400G ports towards systems/leaf * 8 leafs = 512 400G to QSFP112 on NIC

2x 200G - MI355X Systems (2560 Accelerators)

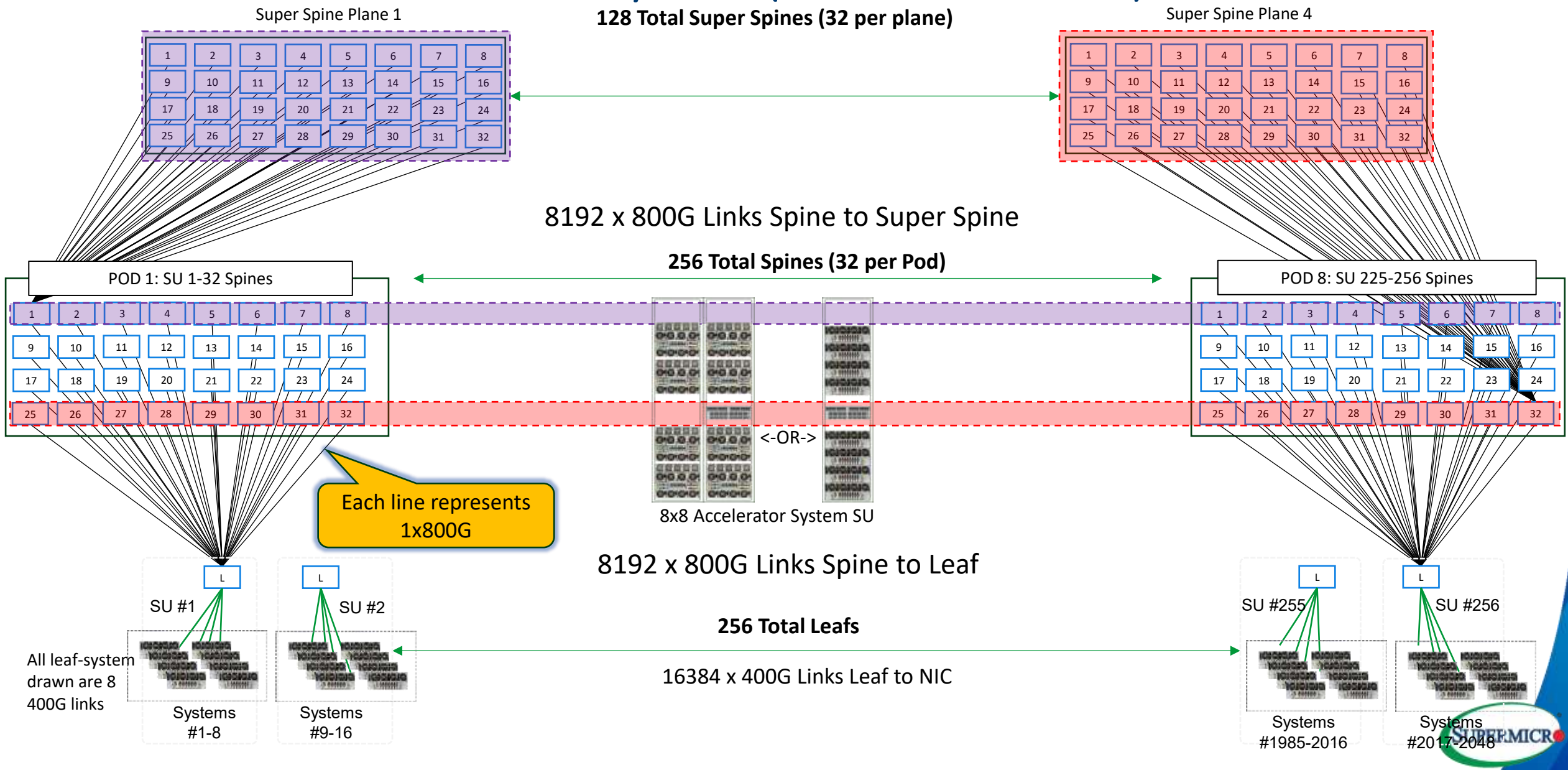


64 400G ports towards systems/leaf * 8 leafs = 512 400G to QSFP112 on NIC

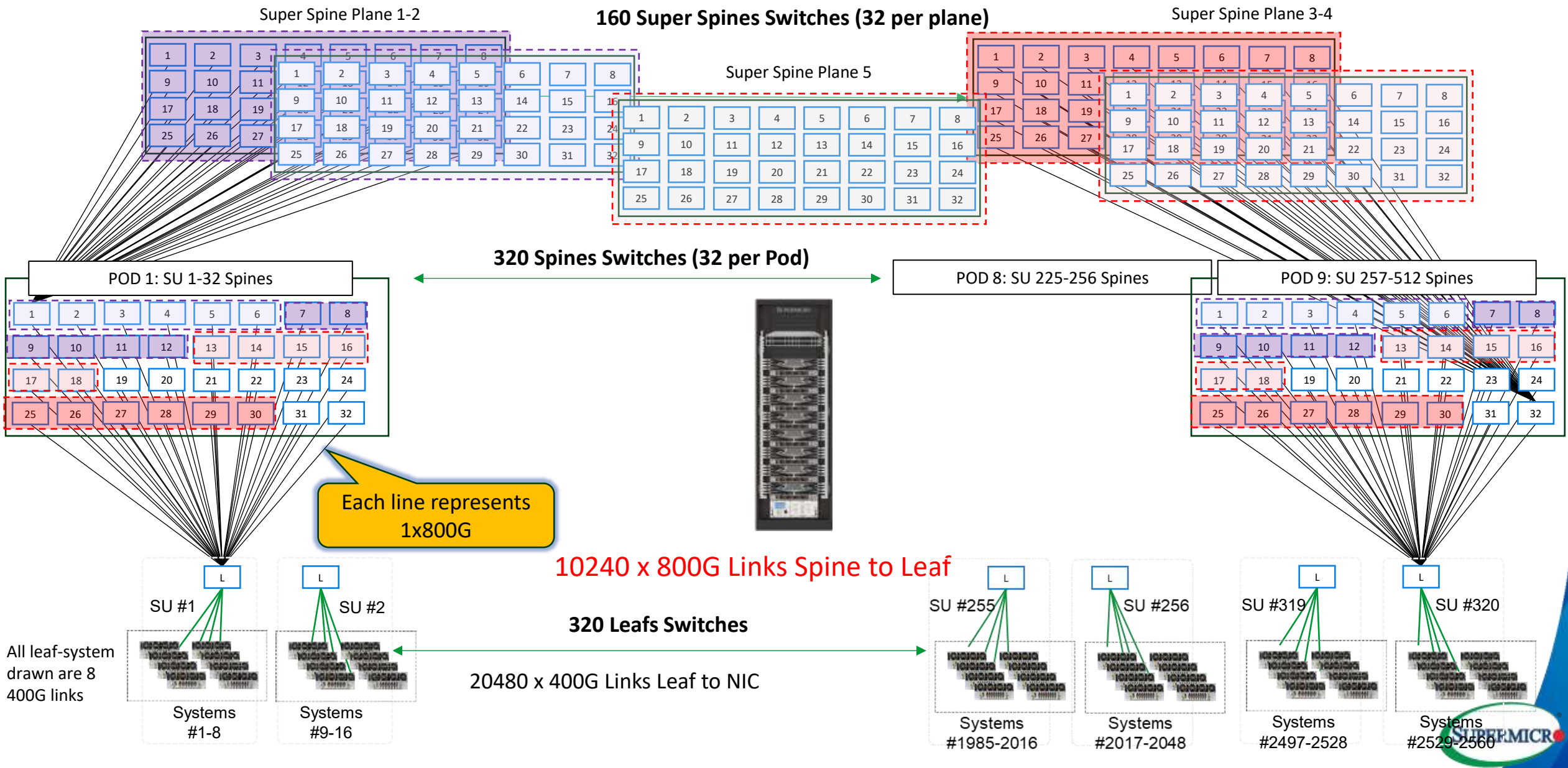
1024 AMD instinct MI355X Systems (8192 Accelerators)



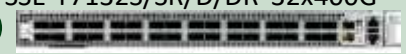
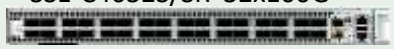
2048 AMD instinct MI355X Systems (16384 Accelerators)



2560 AMD MI355X Systems (20480 GPU)



2026 Supermicro Switching Hardware Roadmap

Switch Type	Shipping	2026	2027
1600GB AI Leaf/Spine		SSE-T8264S/SR/D/DR 64x1600 OSFP (CQ2/3 26) 	SSE-T8264D-CPO-ALC 64x1600 2U DLC with Co-Packaged Optics MPO-16  (NC: CQ1/27)
800GB AI/DC/Multi-tenant/ Storage Leaf/Spine	SSE-T8164S/SR/D/*S-LCC/SR-LCC/D-LCC/S27 64x800 OSFP  SSE-T8196S/SR/D/*S-LCC/SR-LCC/D-LCC/S27 32x800 OSFP & 64x400 QSFP112 	SSE-T8116S/SR/D/DR 16x800 OSFP  (NC: CQ3/26) SSE-T8132S/SR/D/DR 24x400 QSFP112 & 8x800 OSFP  (NC: CQ3/26)	
400GB AI Leaf/Multi-tenant/Storage	SSE-T8032S 64x400G 2Port OSFP  SSE-T7132S/SR/D/DR 32x400G 		
10/25/100GB DC Leaf/Spine	SSE-C4632S/SR 32x100G  SSE-F3648S/SR 48x10/25G + 8x100G 		
1GB OOB	SSE-G3748/R/D/DR 48x1G, 6x10/25G 	SSE-G3848/R/D/DR 48x1G, 4x10/25G  (CQ2/26)	NC: Not Committed

* LCC Versions in Q2CY25

Community SONiC

Enterprise Advanced SONiC

Better Faster Greener™ © 2026 Supermicro

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

© 2025 Super Micro Computer, Inc. All rights reserved.



www.supermicro.com