Use Cases

- Game development
- Product design
- City planning/architectural
- Digital twins (manufacturing, assembly lines, logistics)

Opportunities and Challenges

- AI-aided game development and asset generation
- Closer to real world scenarios
- Integrated engineering
- Enterprise-scale simulations
- Lower latencies
- Cloud collaboration opportunities

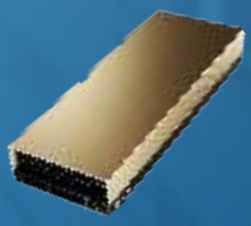
Key Technologies

- NVIDIA OVX™ certified architecture
- NVIDIA L40S, L40, and RTX 6000 Ada GPUs
- NVIDIA BlueField®-2, or BlueField®-3 (DPU)
- NVIDIA RTX GPUs with ray tracing
- Rack-scale integration

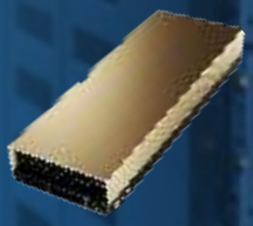
Solution Stack

- Universal Scene Description Connectors
- NVIDIA Omniverse™ Enterprise

L40S
FHFL DW
PCIe 4.0 x16
350W
48GB GDDR6



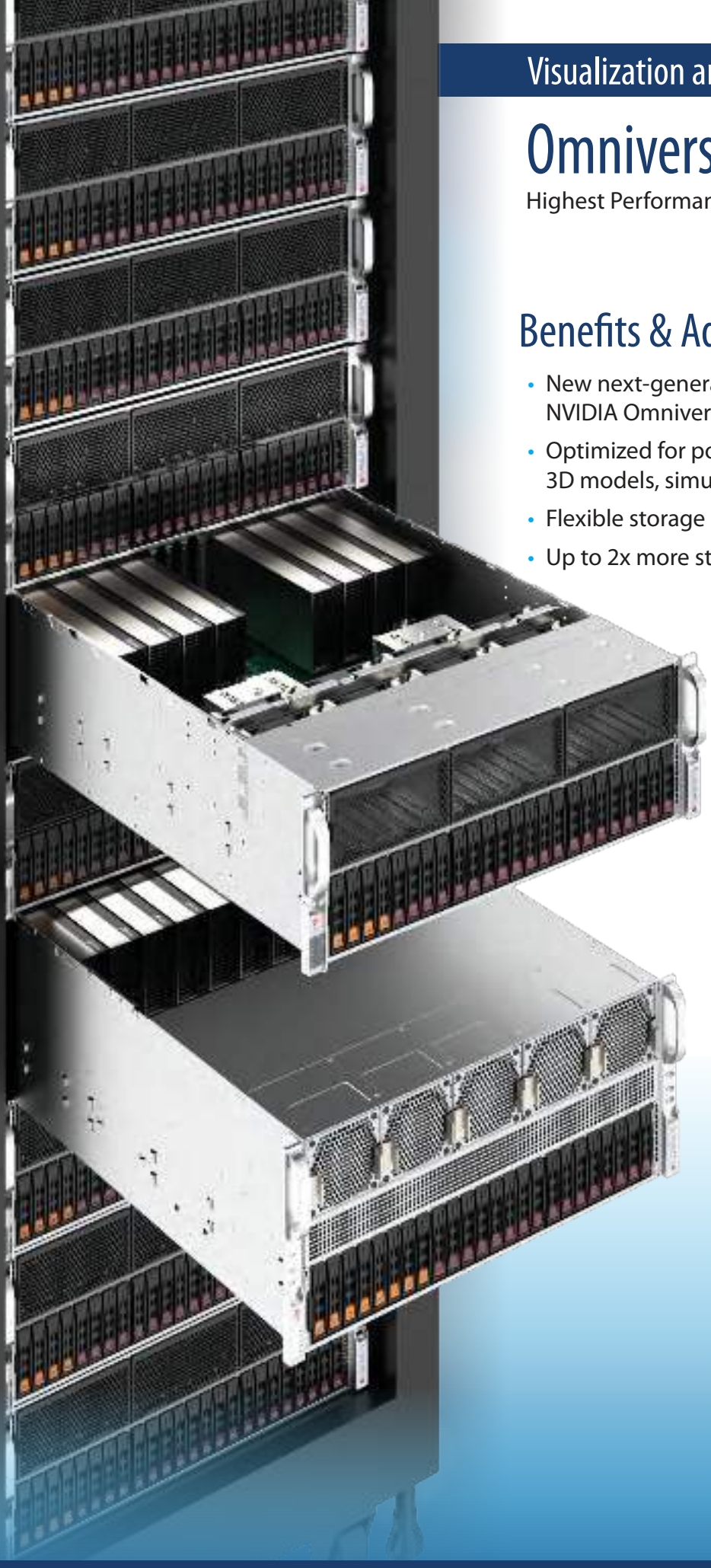
L40
FHFL DW
PCIe 4.0 x16
300W
48GB GDDR6



RTX 6000 ADA
Graphics, Ray Tracing
FHFL DW
PCIe 4.0 x16
300W
48GB GDDR6

Omniverse Optimized Systems

Highest Performance, Tailored for NVIDIA Omniverse



Benefits & Advantages

- New next-generation purpose-built system for NVIDIA Omniverse™ Enterprise
- Optimized for power immersive, photorealistic 3D models, simulations, and digital twins
- Flexible storage configurations
- Up to 2x more storage and I/O flexibility

4U/5U 8 GPU (PCIe)
8 NVIDIA L40S/L40 PCIe
3 NVIDIA ConnectX-7
16 U.2 NVMe drives
SYS-421GE-TNRT /
AS -4125GS-TNRT /
SYS-521GE-TNRT

Key Features

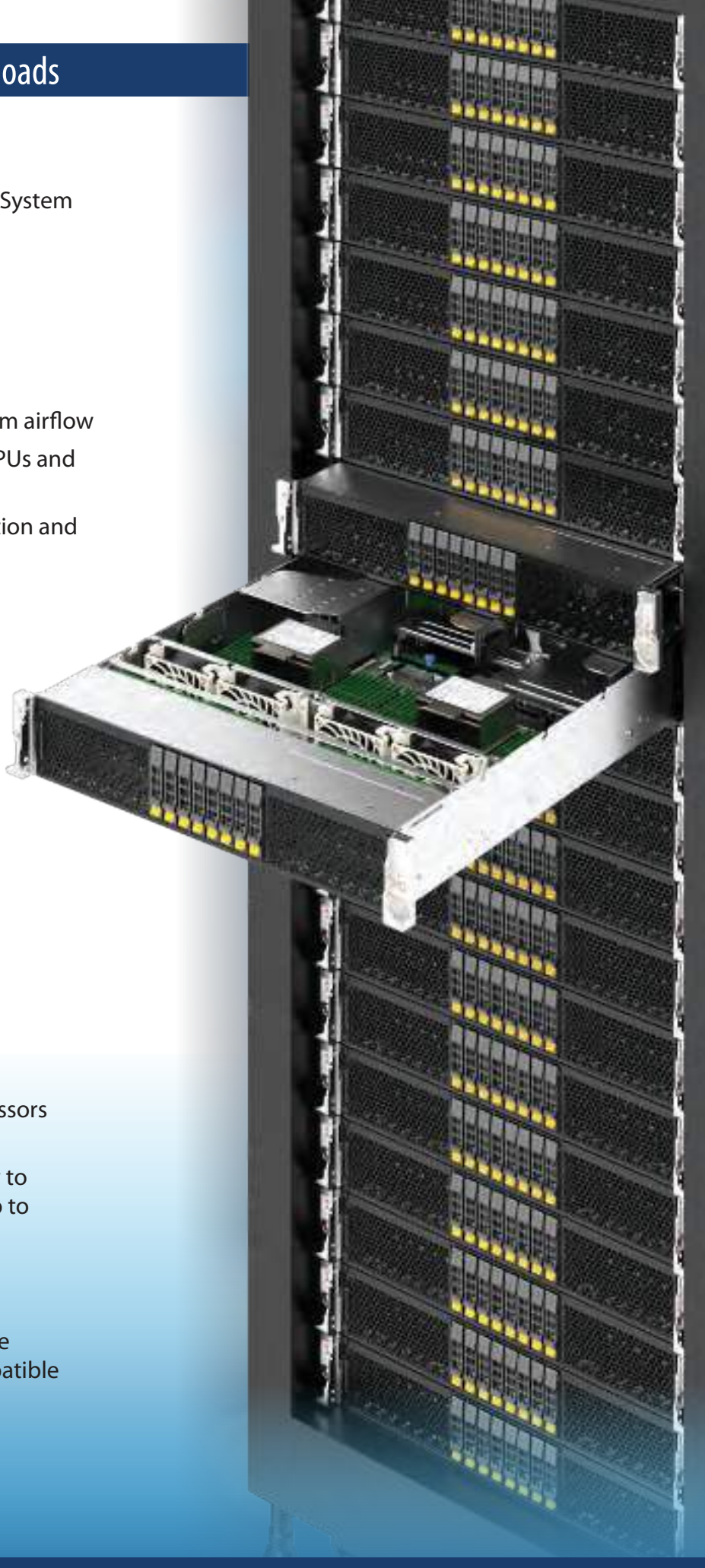
- 8 NVIDIA L40S/L40 PCIe GPUs
- Dual 4th Gen Intel® Xeon® Scalable processors or AMD EPYC™ 9004 Series processors
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- 3 NVIDIA ConnectX-7
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling.
- 16 U.2 NVMe drive bays

2U Hyper Systems

Hyper - Flagship Performance Rackmount System
Designed for Ulimate Flexibility

Benefits & Advantages

- Highly flexible modular architecture
- Compute optimized design for maximum airflow
- Maximum availbility of PCIe lanes for GPUs and networking
- Tool-less platform for ease of configuration and servicing



2U Hyper
4 NVIDIA L40 PCIe
8 NVMe drives
32 DIMMs DDR5-4800
SYS-221H-TNR / AS -2115HS-TNR

Key Features

- Up to 4 NVIDIA L40S/L40 GPUs
- Dual 4th Gen Intel® Xeon® Scalable processors or AMD EPYC™ 9004 Series processors
- Optimized thermal capacity and airflow to support CPUs up to 350W with GPUs up to 350W with air cooling
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- Advanced I/O Module (AIOM) for flexible networking options - OCP 3.0 SFF compatible

AI Workstations

4-GPU 5U Full-Tower Rackmount Workstation

Benefits & Advantages

- Powerful, compact configuration optimized for Omniverse and AI development
- Rackmount data center server performance in portable tower form factor
- Ideal for office, school, lab or field deployment
- NVIDIA qualified system

5U Full-Tower Workstation

4 NVIDIA L40S PCIe
Dual 4th Gen Intel® Xeon® Scalable
16 DIMM slots DDR5-4800

SYS-741GE-TNRT

Key Features

- 4 NVIDIA L40S/L40 PCIe GPUs
- Dual 4th Gen Intel Xeon Scalable processorsSupports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- 8 3.5" hot-swap NVMe/SATA/SAS and 2 M.2 slots
- 4 PCIe 5.0 x16 double-width slots (for GPUs) and 3x PCIe 5.0 x16 single-width slots for maximum flexibility
- On-board 10GbE LAN

Graphic Workstations

4-GPU 5U Full-Tower Rackmount Workstation

Benefits & Advantages

- Versatile and flexible configuration for a range of media, visualization and AI workloads
- High core count to support maximum I/O for PCIe expansion, M.2 storage and SATA drive bays
- NVIDIA Certified platform

Full Tower Workstation

4 NVIDIA RTX A6000 or 3 RTX 6000 ADA
AMD Ryzen™ Threadripper™ PRO
8 DIMM Slots DDR4-3200
AS -5014A-TT

Key Features

- 4 NVIDIA RTX™ 6000 Ada or A6000 GPUs
- Single AMD Ryzen Threadripper PRO processor up to 64 cores
- 4 PCIe 4.0 x4 M.2 slots + 6 SATA drive bays
- Onboard 10GbE LAN
- Optional CPU liquid cooling

5

Video Delivery Workloads

Content Delivery Networks (CDNs), Transcoding, Compression, Cloud Gaming/Streaming

Workload Sizes

Large

BigTwin® 2U 4-Node
Content Delivery Networks

Medium

CloudDC 2U UP
Streaming and Transcoding

Small

Hyper-E 2U DP
Edge Video

Use Cases

- Content delivery networks
- 8K, 4K streaming, livebroadcast
- High resolution, high framerate cloud gaming and streaming

Opportunities and Challenges

- Save data bandwidth and reduce delivery delays
- Faster, more efficient transcoding and compression
- Reduce power consumption and infrastructure cost

Key Technologies

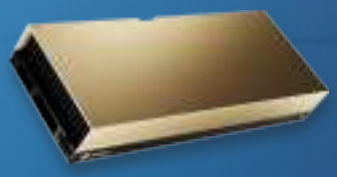
- GPU media engines with transcoding acceleration including AV1 encoding and decoding
- NVIDIA L40, L4, and RTX GPUs
- NVIDIA BlueField®-2 or BlueField-3 (DPU)
- Dense, resource-saving multi-node, multi-GPU systems for space and power efficiency
- High-capacity, high-throughput hot-swap storage

Solution Stack

- Red Hat, VMWare
- Container orchestration and management
- SDKs to accelerate and optimize decoding, encoding and transcoding workloads

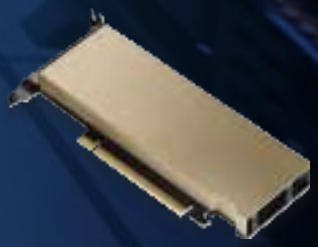
L40

FHFL DW
PCIe 4.0 x16
300W
48GB GDDR6



L4

HHHL SW
PCIe 4.0 x16
72W
24GB GDDR6



BigTwin® 2U 4-Node

BigTwin – Award Winning Multi-Node System with Resource Saving Architecture

Benefits & Advantages

- Multi-node form factors optimized for compute or storage density
- Dual processors per node
- Free-air cooling and liquid cooling options
- Front hot-swap storage drives and rear hot-swap server nodes

BigTwin 2U 4-Node
1 NVIDIA L4 PCIe per node
6 2.5" NVMe drives per node
16 DIMMs DDR5-4800 per node
SYS-221BT-HNTR / SYS-621BT-HNTR

Key Features

- Up to 1 GPUs per node
- Dual 4th Gen Intel® Xeon® Scalable processors per node
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- 2 PCIe 5.0 x16 (LP) slots
- 6 NVMe drives per node (2U4N) or 12 NVMe drives per node (2U2N)
- Networking via AIOM (OCP 3.0 compatible) per node

2U CloudDC UP

CloudDC - All-in-one Platform for Cloud Data Centers

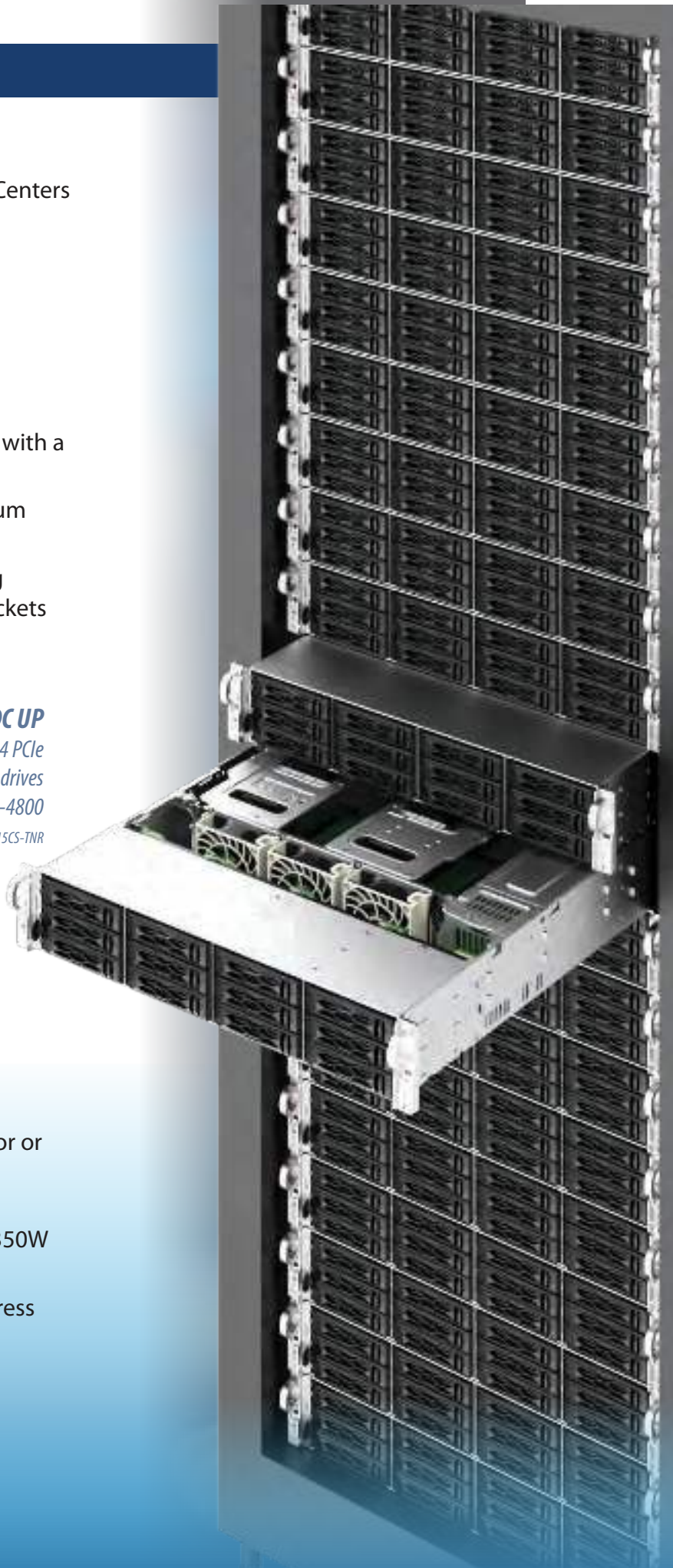
Benefits & Advantages

- UP architecture for maximum performance with a single CPU
- Superior thermal design - Supports maximum power/performance CPUs and GPUs
- Flexible I/O and storage options supporting convenient serviceability with tool-less brackets and hot-swap drive bays

2U CloudDC UP

2 NVIDIA L40 PCIe or 4 NVIDIA L4 PCIe
12 3.5" SATA drives
16 DIMMs DDR5-4800

SYS-521C-NR / AS -2015CS-TNR



Key Features

- Up to 6 GPUs
- Single 4th Gen Intel® Xeon® Scalable processor or AMD EPYC™ 9004 Series processor
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 350W with air cooling
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- 16 DIMM slots DDR5
- Advanced I/O Module (AIOM) for flexible networking options (OCP 3.0 compatible)

2U Hyper-E

Hyper-E- High Performance and Flexibility at the Edge

Benefits & Advantages

- Short-depth chassis ideal for edge deployments
- Front I/O with rear storage access
- AC and DC power options

2U Hyper-E
3 NVIDIA L40 PCIe
6 NVMe drives
32 DIMMs DDR5-4800
SYS-221HE-FTNR / SYS-221HE-FTNRD

Key Features

- 3 NVIDIA L40S/L40 PCIe GPUs
- Dual 4th Gen Intel® Xeon® Scalable processors
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1
- 32 DIMM slots DDR5.
- Networking via AIOM (OCP 3.0 compatible)

Highly Efficient Sustainable Flash

For read-intensive content delivery

Benefits & Advantages

- Maximum density design to support up to 1PB in 2U with next-generation drives
- Direct-attached EDSFF E1.S and E3.S media for the best thermal and I/O performance
- CPUs with built-in Intel Accelerator Engines to offload storage functions and improve performance
- Flexible topology allows distribution of PCIe lanes based on performance and density requirements

1U 24-Bay E1.S
SSG-121E-NES24R

1U 24-Bay E1.S
SSG-121E-NE316R /
ASG-1115S-NE316R

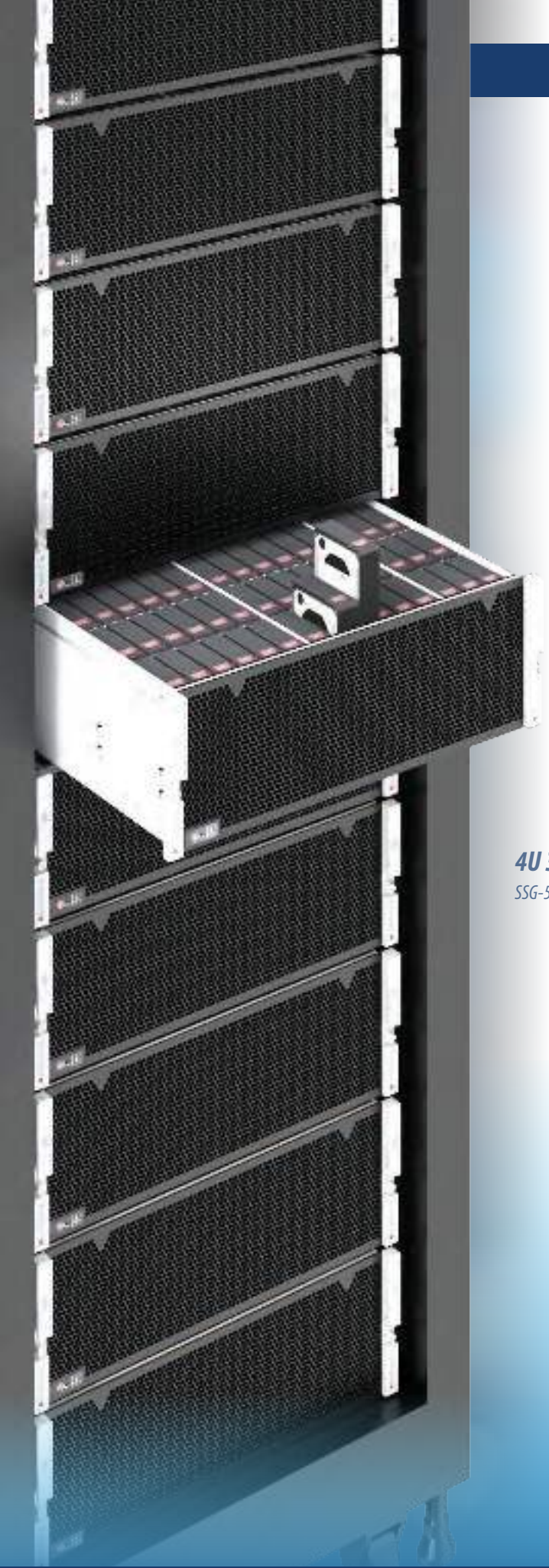
2U 32-Bay E3.S
SSG-221E-NE332R /
ASG-2115S-NE332R

Key Features

- Dual 4th Gen Intel® Xeon® Scalable processors or single AMD EPYC 9004 Series processors
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1
- Up to 24 drives in 1U or 32 drives in 2U
- 2 PCIe 5.0 x16 slots + 2 PCIe 5.0 x16 AIOM slots

Scale-Out Origin Storage

For active archive, user-licensed content, copyright compliance



Benefits & Advantages

- Storage Bays divided between 2x nodes to create scale-out architectures with maximum density
- Optimal Configurations using 30 or 45 HDD per node
- Top-loading drawer with tool-less drive brackets for easy servicing and maintenance
- Designed to be maintained with minimal datacenter staff

4U 30/45-Bay Top-Loading

SSG-540P-E1CTR45L

Key Features

- Dual node twin design
- Dual 3rd Gen Intel® Xeon® Scalable processors per node
- 3 PCIe 4.0 x16 slots per node for I/O
- Designed to be maintained with minimal datacenter staff

6 AI Edge Workloads

Edge Video Transcoding, Edge Inference, Edge Training

Workload Sizes

Extra Large

Hyper-E
Multi-GPU Inferencing and Training

Large

Compact
Multi-GPU Inferencing

Medium

Short-Depth Multi-GPU Edge Server

Small

Embedded
CPU (or ASIC) based Inference

Use Cases

- Video processing: decode, encode, and transcode
- Edge inference: vision, speech, anomaly detection, etc.
- Markets: security and surveillance, retail, manufacturing, healthcare, and medical devices

Opportunities and Challenges

- Size, weight, and power constraints
- Data throughput for video and audio
- Cost of storage, bandwidth constraints
- Latency impacting decision response times
- Data security, privacy, and sovereignty laws
- Resiliency in face of network outages
- Long product lifecycle requirements

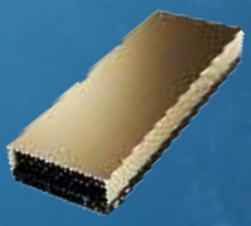
Key Technologies

- CPU or GPU-based AI edge Inferencing, GPU-based AI edge training, and video transcoding/encoding/decoding
- NVIDIA L4, L40S, L40, A30, A40, T4, A2 GPUs
- Short-depth chassis design for edge locations with AC or DC power supply options
- Front I/O with broad range of expansion and I/O port for flexibility and serviceability
- Ruggedized systems designed to be placed outside of the data center

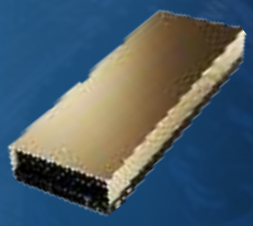
Solution Stack

- NVIDIA® TensorRT™ and Triton Inference Server
- NVIDIA DeepStream, Clara, Merlin, Metropolis, Morpheus, Omniverse, and Riva
- NVIDIA Fleet Command
- Intel® OpenVINO

L40S
FHFL DW
PCIe 4.0 x16
350W
48GB GDDR6



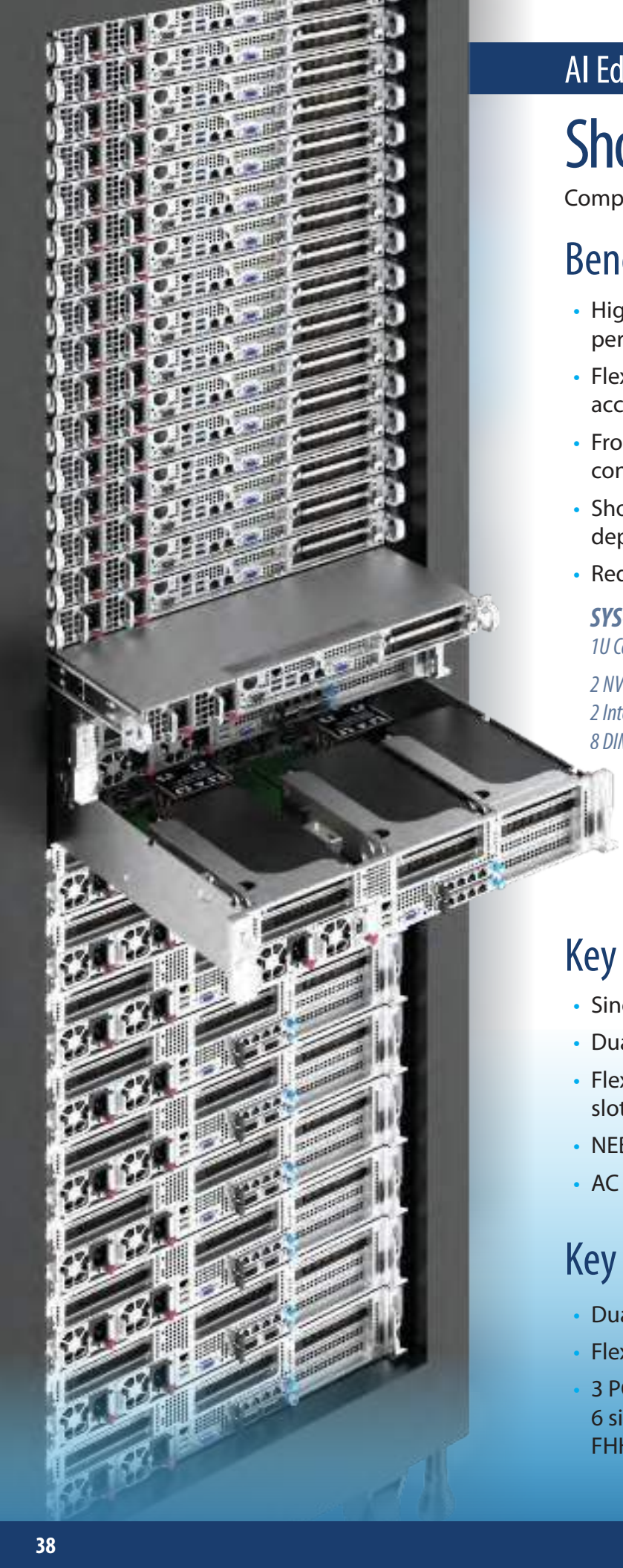
L40
FHFL DW
PCIe 4.0 x16
300W
48GB GDDR6



L4
HHHL SW
PCIe 4.0 x16
72W
24GB GDDR6

Short-Depth 5G/Edge & Hyper E

Compute and AI Performance at the Edge



Benefits & Advantages

- High-density systems for data center level performance at the Edge
- Flexible configurations with broad AI accelerator and AOC options
- Front I/O for easier serviceability in space-constrained environments
- Short-depth chassis design for easy deployment at edge locations
- Redundant AC or DC power supply options

SYS-111E-FWTR

1U Compact Edge/5G Server

2 NVIDIA L4
2 Internal Drive Bays
8 DIMMs DDR5-4800

2U Hyper-E

3 NVIDIA H100 PCIe
6 NVMe drives
32 DIMMs DDR5-4800

Key Features (SYS-111E-FWTR)

- Single 4th Gen Intel® Xeon® Scalable processor
- Dual 10 GbE connectivity
- Flexible configuration with 3 PCIe 5.0 x16 slots (2x FHFL and 1x LP)
- NEBS Level 3 design
- AC and DC power options available

Key Features (Hyper-E)

- Dual 4th Gen Intel® Xeon® Scalable processors
- Flexible network options with 2 AIOM slots
- 3 PCIe 5.0 x16 FHFL double-width slots or 6 single-width slots 2 PCIe 5.0 single width FHHL slots

Fanless and Wallmount Edge

Compact Systems for the Intelligent Edge

Benefits & Advantages

- Compact form factors for deployments at the edge and remote edge
- Designed for ruggedized environments outside the data center
- Deliver low-latency AI inferencing for intelligent edge applications
- Broad range of expansion and I/O port options

SYS-E100-13AD

Ultra-compact Fanless Edge Server
CPU (or ASIC) based Inference

Key Features (SYS-E100-13AD)

- 12th Gen Intel® Core™ processors
- Fanless design for best durability and silent operations
- 3 M.2 expansion slots (NVME, Wi-Fi, LTE/5G)
- USB, HDMI, DP, COM and GPIO ports

SYS-E403-13E

Powerful expandable Server for the Edge

1 NVIDIA L40S OR 2 NVIDIA L4

8 DIMM slots DDR5-4800
4 NVMe Drives

Key Features (SYS-E403-13E)

- 4th Gen Intel® Xeon® Scalable processor
- 3 PCIe 5.0 x16 FHFL slots
- Dual 10 GbE Ethernet
- Optional wall-mounted installation

AI GPU WORKLOADS

LARGE SCALE AI TRAINING

GPU OPTIMIZED

HPC

MULTI-NODE BUILDING BLOCKS

ENTERPRISE AI INFERENCING & TRAINING

VISUALIZATION AND OMNIVERSE

RACKMOUNT BUILDING BLOCKS

VIDEO DELIVERY

EDGE

EDGE OPTIMIZED

NVIDIA GPUs

LARGE SCALE AI TRAINING HPC

TRAINING & INFERENCE

OMNIVERSE

VIDEO DELIVERY

EDGE

H100 SXM5
4 or 8
H100 GPU Board
NVLink &
NVSwitch Fabric
PCIe 5.0
700W per GPU
80GB HBM3 per GPU

H100 NVL
2 FHFL
H100 GPU
with NVLink Bridge
(4x faster than PCIe)
PCIe 5.0
400W per GPU
94GB HBM3 per GPU

H100 PCIE
FHFL DW
PCIe 5.0 x16
350W
80GB HBM2e

L40S
FHFL DW
PCIe 4.0 x16
350W
48GB GDDR6

L40
FHFL DW
PCIe 4.0 x16
300W
48GB GDDR6

RTX 6000 ADA
FHFL DW
PCIe 4.0 x16
300W
48GB GDDR6

L4
HHHL SW
PCIe 4.0 x16
72W
24GB GDDR6

Supermicro System GPU Compatibility

	H100 (SXM)	H100 (NVL)	H100 (PCIe)	L40S	L40	L4	RTX 6000 Ada
4U/5U/8U GPU	4 (4U/5U) 8 (8U)						
4U/5U 10-GPU		10 (4U/5U)	10 (4U/5U)	10 (4U/5U)	10 (4U/5U)	10 (4U/5U)	8 (4U/5U)
SuperBlade		20 (8U) 10 (6U)	20 (8U) 10 (6U)	20 (8U) 10 (6U)	20 (8U) 10 (6U)	40 (8U) 20 (6U)	
BigTwin			4 (2U2N)	4 (2U2N)	4 (2U2N)	4 (2U2N) 4 (2U4N)	2 (2U)
CloudDC			2 (2U)	2 (2U)	2 (2U)	4 (2U) 2 (1U)	
Hyper			4 (2U) 1 (1U)	4 (2U) 1 (1U)	4 (2U) 1 (1U)	4 (2U) 2 (1U)	
WIO						2 (2U) 2 (1U)	
Hyper-E			3	3	3	4	
Short-Depth Edge						2	
Compact Edge/IoT				1	1	2	
Workstation		4	4	4	6	4	

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

Worldwide Headquarters
Super Micro Computer, Inc.
980 Rock Ave.
San Jose, CA 95131, USA
Tel: +1-408-503-8000
Fax: +1-408-503-8008
E-mail: Marketing@Supermicro.com

EMEA Headquarters
Super Micro Computer, B.V.
Het Sterrenbeeld 28, 5215 ML,
's-Hertogenbosch, The Netherlands
Tel: +31-73-640-0390
Fax: +31-73-641-6525
E-mail: Sales_Europe@supermicro.com

APAC Headquarters
Super Micro Computer, Taiwan Inc.
3F, No. 150, Jian 1st Rd., Zhonghe Dist.,
New Taipei City 235, Taiwan
Tel: +886-2-8226-3990
Fax: +886-2-8226-3991
E-mail: Marketing@Supermicro.com.tw

MKT-0002-09/2023-R39

www.supermicro.com