SUPERMICRO®

Accelerate Everything

Large Language Models (LLM)
to the AI Edge

November 2023

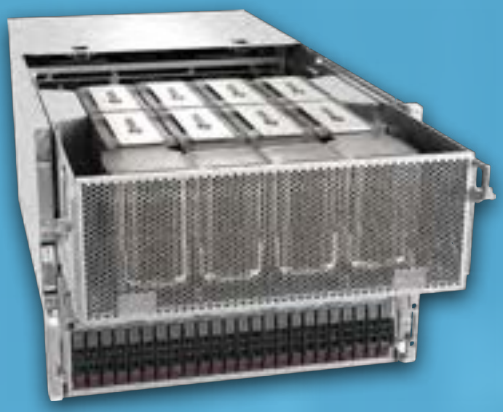
AI GPU Solution Portfolio

Unlock Unprecedented Performance Leveraging GPU Optimized Systems

GPU technology can bring unprecedented performance to a broad spectrum of workloads – up to 5X, 10X, … 100X improvements in performance and efficiency. These workloads span from the rapidly growing generative AI market to enterprise inferencing, product design, visualization, and to the intelligent edge. Supermicro has built a portfolio of workload-optimized systems for optimal GPU performance and efficiency across this broad spectrum of workloads.

TABLE OF CONTENTS

#1 GPU SOLUTIONS IN THE MARKET



8U HGX H100 8-GPU System
- Large Language Models (LLM)
- 900GB/s NVLink 7x better performance than PCIe
- 1:1 networking slots for GPUs up to 400Gbps each

4U HGX H100 4-GPU System
- HPC/AI Workloads
- Double-precision Tensor Cores delivering up to 268 teraFLOPS
- Superior thermal design and liquid cooling option

SuperBlade®
- Up to 20 GPUs in 8U
- Highest Density
- Multi-Node Architecture

Petabyte Scale Storage
- Maximum density design to support up to 1PB in 2U
- Up to 32 E3.S NVMe drives in 2U

2U MGX System
- Modular Building Block Platform Supporting Today's and Future GPUs, CPUs, and DPUs

1U Grace Hopper System
- CPU+GPU Coherent Memory System

1 Large Scale AI Training Workloads

Generative AI, Natural Language Processing (NLP), Computer Vision

Workload Sizes

Extra Large

Liquid Cooled AI Rack Solutions
NVIDIA HGX™ H100 SXM 8-GPU
Up to 80 kW/Rack

Large

8U 8-GPU System
NVIDIA HGX H100 SXM 8-GPU

Medium

4U 4-GPU System
NVIDIA HGX H100 SXM 4-GPU

Storage

Petabyte Scale Storage
High throughput and High Capacity
for AI Data Pipeline

Use Cases

- Large Language Models (LLMs)
- Autonomous Driving Training
- Recommender Systems

Opportunities and Challenges

- Continuous growth of data set size
- High performance everything: GPUs, memory, storage and network fabric
- Pool of GPU memory to fit large AI models and interconnect bandwidth for fast training

Key Technologies

- NVIDIA HGX H100 SXM 8-GPU/4-GPU
- GPU/GPU interconnect (NVLink and NVSwitch), up to 900GB/s – 7x greater than PCIe 5.0
- Dedicated high performance, high capacity GPU memory
- High throughput networking and storage per GPU enabling NVIDIA GPUDirect RDMA and Storage.

Solution Stack

- DL Frameworks: TensorFlow, PyTorch
- Transformers: BERT, GPT, Vision Transformer
- NVIDIA AI Enterprise Frameworks (NVIDIA Nemo, Metropolis, Riva, Morpheus, Merlin
- NVIDIA Base Command (infrastructure software libraries, workload orchestration, cluster management)
- High performance storage (NVMe) for training cache
- Scale-out storage for raw data (data lake)

HGX H100 Systems

- H100 SXM5 board with 4-GPU or 8-GPU
- NVLink & NVSwitch Fabric
- Up to 700W per GPU

AI Rack Solutions

Multi-Architecture Flexibility with Future-Proof Open-Standards-Based Design for POD, and SuperPOD with Liquid Cooling

Benefits & Advantages

- Proven AI rack cluster deployment in some of the world's largest AI clusters
- AI POD, SuperPOD customizable architecture
- Turn-key proven solutions accelerates time to market
- Traditional, free-air and liquid cooled configurations for optimal TCE/TCO

Key Features

- Factory integrated and fully tested multi-rack cluster
- Server, storage, networking, software, management total solutions designed, built and deployed to your specification
- Rack Scale L11/L12 testing and validation
- Factory tuned power and cooling design
- Single source liquid cooling solution available with reduced (weeks) lead time

HGX H100 Systems

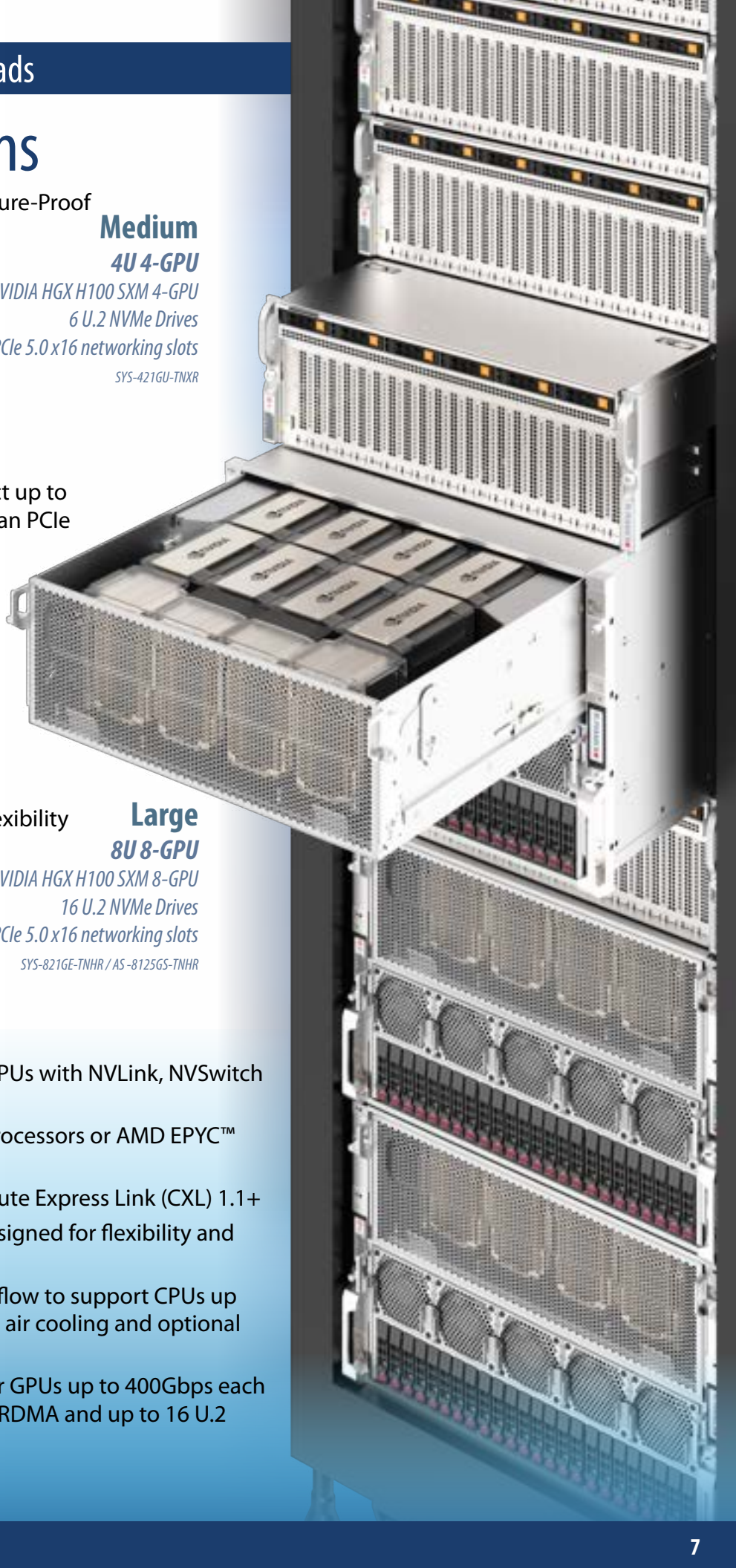
Multi-Architecture Flexibility with Future-Proof Open-Standards-Based Design

Medium

4U 4-GPU

NVIDIA HGX H100 SXM 4-GPU
6 U.2 NVMe Drives
8 PCIe 5.0 x16 networking slots

SYS-421GU-TNXR



Benefits & Advantages

- High performance GPU interconnect up to 900GB/s - 7x better performance than PCIe
- Superior thermal design supports maximum power/performance CPUs and GPUs
- Dedicated networking and storage per GPU with up to double the NVIDIA GPUDirect throughput of the previous generation
- Modular architecture for storage and I/O configuration flexibility with front and rear I/O options

Large

8U 8-GPU

NVIDIA HGX H100 SXM 8-GPU
16 U.2 NVMe Drives
8 PCIe 5.0 x16 networking slots

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

Key Features

- 4 or 8 next-generation H100 SXM GPUs with NVLink, NVSwitch interconnect
- 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+
- Innovative modular architecture designed for flexibility and futureproofing in 8U or 4U.
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling and optional liquid cooling
- PCIe 5.0 x16 1:1 networking slots for GPUs up to 400Gbps each supporting GPUDirect Storage and RDMA and up to 16 U.2 NVMe drive bays

Petabyte Scale NVMe Flash

High Throughput and High Capacity Storage
for AI Data Pipeline

1U 24-Bay E1.S

SSG-121E-NES24R

Benefits & Advantages

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

1U 16-Bay E3.S

SSG-121E-NE316R / ASG-1115S-NE316R

2U 24/32-Bay E3.S

SSG-221E-NE324R / ASG-2115S-NE332R

Key Features

- Dual 4th Gen Intel Xeon Scalable processors or single AMD EPYC™ 9004 Series processor
- Up to 32 E3.S NVMe drives in 2U
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+

Petabyte Scale HDD

Top-Loading Data Lake Storage

Benefits & Advantages

- Fully redundant dual-ported high availability/failover clustering for use with Parallel File Systems
- Dual ported SAS architecture with 60 and 90 Bay configurations
- Top-loading drawer with tool-less drive brackets for easy servicing and maintenance
- Industry standard SAS controllers and expander infrastructure to support the most popular SDS platforms like ZFS and Lustre

4U 60/90-Bay Top-Loading

SSG-640SP-E1CR60 / SSG-640SP-E1CR90

Key Features

- Two hot-pluggable system nodes
- Dual 3rd Gen Intel® Xeon® Scalable processors per node
- 3 PCIe 4.0 x16 slots per node for I/O

2 HPC/AI Workloads

Simulation: Stress Analysis, Aerodynamics, Device Performance Prediction, Fluid Dynamics, Research, Exploration, Weather Prediction

Workload Sizes

Large

8U 8-GPU or 4U 4-GPU System
NVIDIA HGX H100 SXM
8-GPU or 4-GPU

SuperBlade®
Highest Density
Multi-Node Architecture

Medium

4U/5U 8-10 GPU PCIe
Maximum Performance
and Flexibility

1U NVIDIA MGX™ System
NVIDIA GH200 Grace Hopper
with CPU+GPU Coherent
Memory

Use Cases

- Manufacturing and engineering simulations (CAE, CFD, FEA, EDA)
- Bio/life sciences (genomic sequencing, molecular simulation, drug discovery)
- Scientific simulations (astrophysics, energy exploration, climate modeling, weather forecasting)

Opportunities and Challenges

- Infusing machine learning algorithms to HPC workloads to achieve faster results and discoveries with more iterations.
- Parallel processing with massive datasets for data-intensive simulations and analytics
- High-resolution and real-time visualization of scientific simulations and modeling

Key Technologies

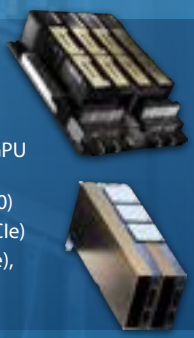
- NVIDIA H100 (SXM, NVL, PCIe), L40S, A100
- NVIDIA Grace Hopper™ Superchip (Grace CPU and H100) with NVLink® Chip-2-Chip (C2C) interconnect
- Dual socket Intel and AMD-based solutions with high CPU core counts
- CPUs integrated with High Bandwidth Memory/bigger L3 cache
- PCIe 5.0 storage and networking
- Liquid cooling

Solution Stack

- NVIDIA HPC Software Development Kit (SDK)
- NVIDIA CUDA
- Commercial and in-house CAE software

HGX H100, H100 NVL, and H100 PCIe

- H100 SXM5 board with 4-GPU or 8-GPU (HGX H100)
- NVLink & NVSwitch Fabric (HGX H100)
- NVLink Bridge (H100 NVL or H100 PCIe)
- 80GB HBM3 (HGX H100 or H100 PCIe), 96GB HBM3 (H100 NVL) per GPU

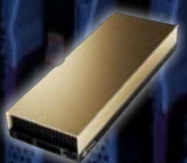


GRACE HOPPER SUPERCHIP

- Grace Arm Neoverse V2 CPU
- NVIDIA H100 with NVLink-C2C
- Up to 480GB LPDDR5X and 96GB HBM3

L40S

FHFL DW
PCIe 4.0 x16
300W
48GB GDDR6



HGX H100 Systems

Designed for Largest AI-fused HPC Clusters

Benefits & Advantages

- Double-precision Tensor Cores delivering up to 535/268 teraFLOPS at FP64 in the 8-GPU/4-GPU respectively.
- TF32 precision to reach nearly 8000 teraFLOPs for single-precision matrix-multiplication
- Superior thermal design and liquid cooling option supports maximum power/perfomance CPUs and GPUs.
- Dedicated networking and storage per GPU with up to double the NVIDIA GPUDirect throughput of the previous generation

4U 4-GPU

NVIDIA HGX H100 SXM 4-GPU
6 U.2 NVMe Drives
8 PCIe 5.0 x16 networking slots

SYS-421GU-TNXR

Key Features

- 4 or 8 H100 SXM GPUs with NVLink, interconnect with up to 900GB/s
- 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+
- Innovative modular architecture designed for flexibility and futureproofing in 8U, 5U, or 4U
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling and optional liquid cooling
- PCIe 5.0 x16 1:1 networking slots for GPUs up to 400 Gbps each supporting GPUDirect Storage and RDMA, and up to 16 U.2 NVMe drive bays, high throughput data pipeline and clustering

8U SuperBlade®

SuperBlade® - Highest Density Multi-Node Architecture for HPC, AI and Cloud Applications

Benefits & Advantages

- Up to 20 nodes in 8U – 100 blades per rack
- Single NVIDIA H100 PCIe GPU per blade
- High CPU to GPU ratio
- Integrated power, cooling, switch and management console
- Up to 95% cable reduction compared to traditional rackmount servers

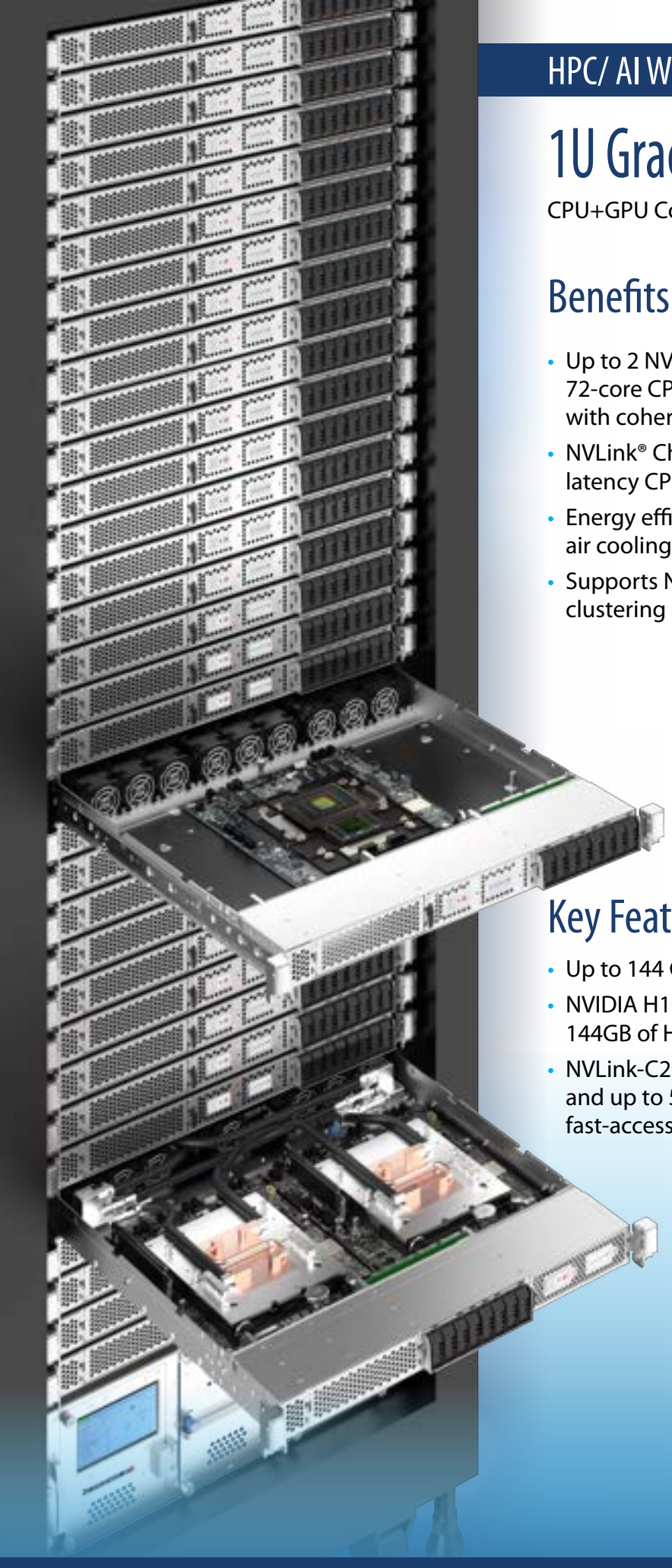
8U SuperBlade®
1 NVIDIA H100 PCIe
2 M.2 NVMe Drives
2 E1.S Drives
200G HDR InfiniBand
SBI-411E-1G/5G

Key Features

- 1 H100 or L40S PCIe GPU per blade
- Single 4th Gen Intel® Xeon® Scalable processor per blade
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1
- Flexible storage options including U.2 NVMe, SAS including M.2 NVMe and EDSFF E1.S
- Shared power, cooling and switch for maximum efficiency with optional liquid cooling
- 2-port 25GbE (3rd and 4th LAN), 1x 200G HDR InfiniBand or 1x 100G EDR InfiniBand via mezzanine card

1U Grace Hopper MGX Systems

CPU+GPU Coherent Memory System for AI and HPC Applications



Benefits and Advantages

- Up to 2 NVIDIA GH200 Grace Hopper Superchips featuring 72-core CPU and H100 Tensor Core GPU tightly coupled with coherent memory
- NVLink® Chip-2-Chip (C2C) high-bandwidth and low-latency CPU-GPU interconnect
- Energy efficient 1000W per Grace Hopper Superchip with air cooling and liquid cooling options.
- Supports NVIDIA BlueField®-3 or ConnectX®-7 for fast clustering and advanced data processing with E1.S drives

1U Grace Hopper MGX System
(air-cooled)
1 NVIDIA GH200 Grace Hopper Superchip
8 E1.S + 2 M.2 drives
96GB HBM3+480GB LPDDR5X
Up to 400G NDR InfiniBand
ARS-111GL-NHR

Key Features

- Up to 144 Grace Arm Neoverse V2 CPU cores in 1U
- NVIDIA H100 Tensor Core GPU with 96GB of HMB3 or 144GB of HBM3e (coming soon) per node
- NVLink-C2C with 900GB/s of CPU-GPU interconnect and up to 576GB (480GB LPDDR5X + 96GB HMB3) of fast-access memory available to the GPU
- Up to 3 PCIe 5.0 x16 slots (1U 1-node) or 2 PCIe 5.0 x16 slots per node (1U 2-node)
- Up to 8 hot-swap E1.S drives and 2 M.2 NVMe drives

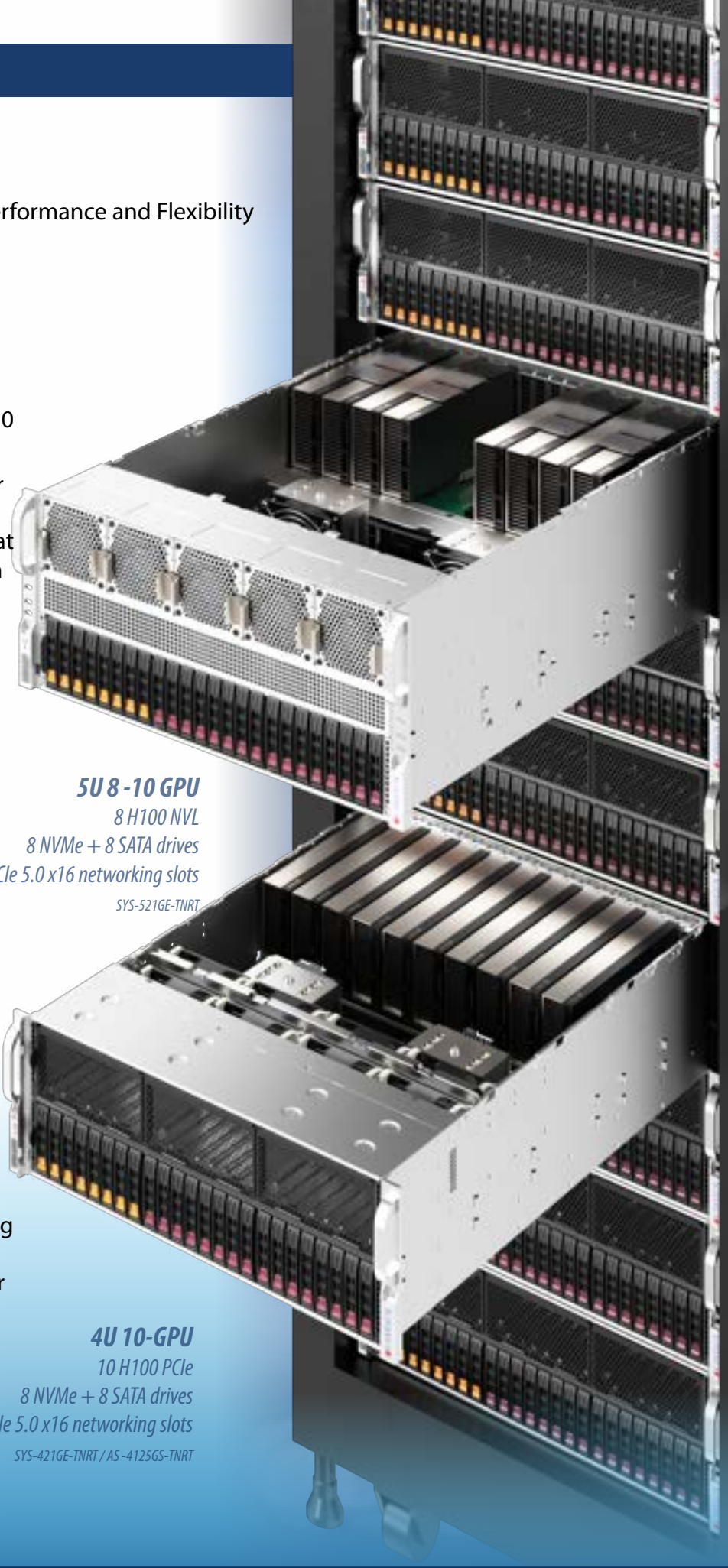
1U 2-Node Grace Hopper MGX System *(liquid-cooled)*
2 NVIDIA GH200 Grace Hopper Superchips (1 per node)
4 E1.S + 2 M.2 drives per node
96GB HBM3+480GB LPDDR5X per node
Up to 400G NDR InfiniBand
ARS-111GL-DNHR-LCC

10 GPU Systems

4U/5U 8 or 10 GPU PCIe - Maximum Performance and Flexibility

Benefits & Advantages

- 13 PCIe 5.0 x16 slots with up to 10 PCIe FHFL GPUs supporting 8 NVIDIA H100 NVL (4 NVLink Bridge pairs) or 10 H100 PCIe GPUs.
- 4U or 5U configurations with superior thermal design supporting max power/performance CPUs and GPUs at up to 32°C ambient temperature with optional air cooling
  - *Single Root, Dual Root or Direct Connect GPU configurations*



5U 8 -10 GPU
8 H100 NVL
8 NVMe + 8 SATA drives
4-5 PCIe 5.0 x16 networking slots
SYS-521GE-TNRT

Key Features

- Up to 8 or 10 H100 PCIe GPUs with optional NVLink Bridge (H100 NVL), or up to 10 L40S
- Dual 4th Gen Intel Xeon Scalable processors or AMD EPYC 9004 Series processors
- Supports PCIe 5.0 DDR5 and Compute Express Link 1.1+
- Configurable with 2 400G networking per root (4 for Dual Root) and Advanced I/O Module (AIOM) slot for high throughput data pipeline and clustering

4U 10-GPU
10 H100 PCIe
8 NVMe + 8 SATA drives
4-5 PCIe 5.0 x16 networking slots
SYS-421GE-TNRT / AS -4125GS-TNRT

3

Enterprise AI Inferencing & Training

Generative AI Inference, Large Language Model Inference, Speech Recognition, Recommendation, Computer Vision

Workload Sizes

Extra Large

4U/5U 8-10 GPU PCIe
GPU-based Inference and Training

Large

6U SuperBlade®
High Density, Disaggregated

Medium

2U MGX System
Modular Building Block Platform Supporting Today's and Future GPUs, CPUs, and DPUs

2U Grace MGX System
Modular Building Block Platform with Energy-efficient Grace CPU Superchip

Enterprise AI Inferencing & Training

Use Cases

- Content creation (image, audio, video, writing)
- AI-enabled office applications and services
- Enterprise business process automation

Opportunities and Challenges

- Total solution complexity
- Open architecture, vendor flexibility, and fast deployment for rapidly evolving technologies
- High computational and resource costs, cloud vs. on-prem
- Utilization of frameworks, pre-trained or open-source AI models with fine-tuning

Key Technologies

- NVIDIA H100 (NVL, PCIe), A100, L40S, L40, and L4 GPUs
- PCIe 5.0 storage and networking
- Intel and AMD CPU options
- NVIDIA Grace™ Superchip (2 Grace CPUs on one Superchip) with NVLink® Chip-2-Chip (C2C) interconnect
- Flexible rackmount servers from 1U to 6U to balance compute, storage, and networking for various enterprise AI workload needs

Solution Stack

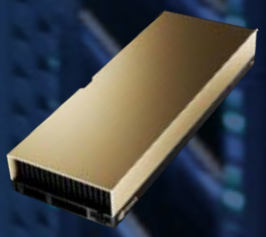
- NVIDIA AI Enterprise software
- NVIDIA NGC™ catalog: containers, pre-trained models
- RedHat OpenShift, VMWare

H100 NVL

- 2 FHFW H100 GPU with NVLink Bridge (4x faster than PCIe)
- PCIe 5.0 x16
- 400W per GPU
- 94GB HBM3 per GPU

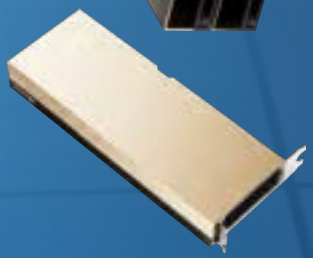
L40S\L40

- FHFL DW
- PCIe 4.0 x16
- 350W (L40S)/300W (L40)
- 48GB GDDR6



H100 PCIE

- FHFL DW
- PCIe 5.0 x16
- 300W per GPU
- 80GB HBM2e



L4

- HHHL SW
- PCIe 4.0 x16
- 72W
- 24GB GDDR6

10 GPU Systems

4U/5U 8 or 10 GPU PCIe — Highly Flexible Architecture

Benefits & Advantages

- Up to 13 PCIe 5.0 slots for flexible GPUs, I/O and networking options
- 4U or 5U configurations with superior thermal design supporting max power/performance CPUs and GPUs at up to 32°C ambient temperature with air cooling
- *Single Root, Dual Root or Direct Connect GPU configurations*

8-10 GPU (PCIe)
8 NVIDIA H100 NVL
or 10 H100 PCIe
8 NVMe and 8 SATA Drives
32 DIMMs DDR5-4800

SYS-421GE-TNRT / AS -4125GS-TNRT / SYS-521GE-TNRT

Key Features

- Up to 8 or 10 H100 PCIe GPUs with optional NVLink Bridge (H100 NVL), or L40S
- 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+
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling.

6U SuperBlade®

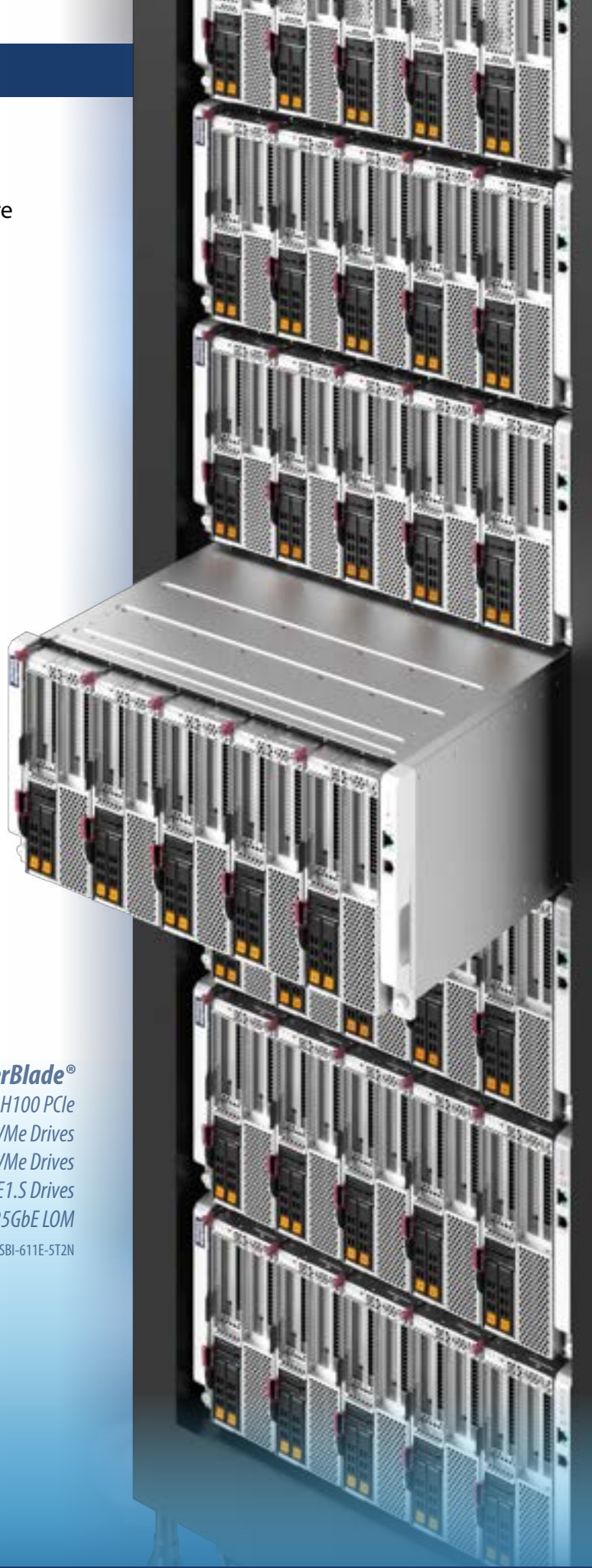
SuperBlade® - Highest Density Multi-Node Architecture for HPC, AI and Cloud Applications

Benefits & Advantages

- Up to 10 single-width nodes in 6U with up to 2 GPUs per blade, or 5 double-width nodes with up to 4 GPUs per blade
- Integrated power, cooling, switch and management console
- Up to 95% cable reduction compared to traditional rackmount servers
- High CPU to GPU Ratio

Key Features

- Up to 2 H100 PCIe or L40S GPUs per blade
- Single 4th Gen Intel® Xeon® Scalable processor per blade
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- Flexible storage options including U.2 (NVMe, SAS, SATA), M.2 (SATA/NVMe), and EDSFF E1.S
- Shared power, cooling and switch for maximum efficiency with optional liquid cooling
- Flexible networking up to 400G NDR InfiniBand



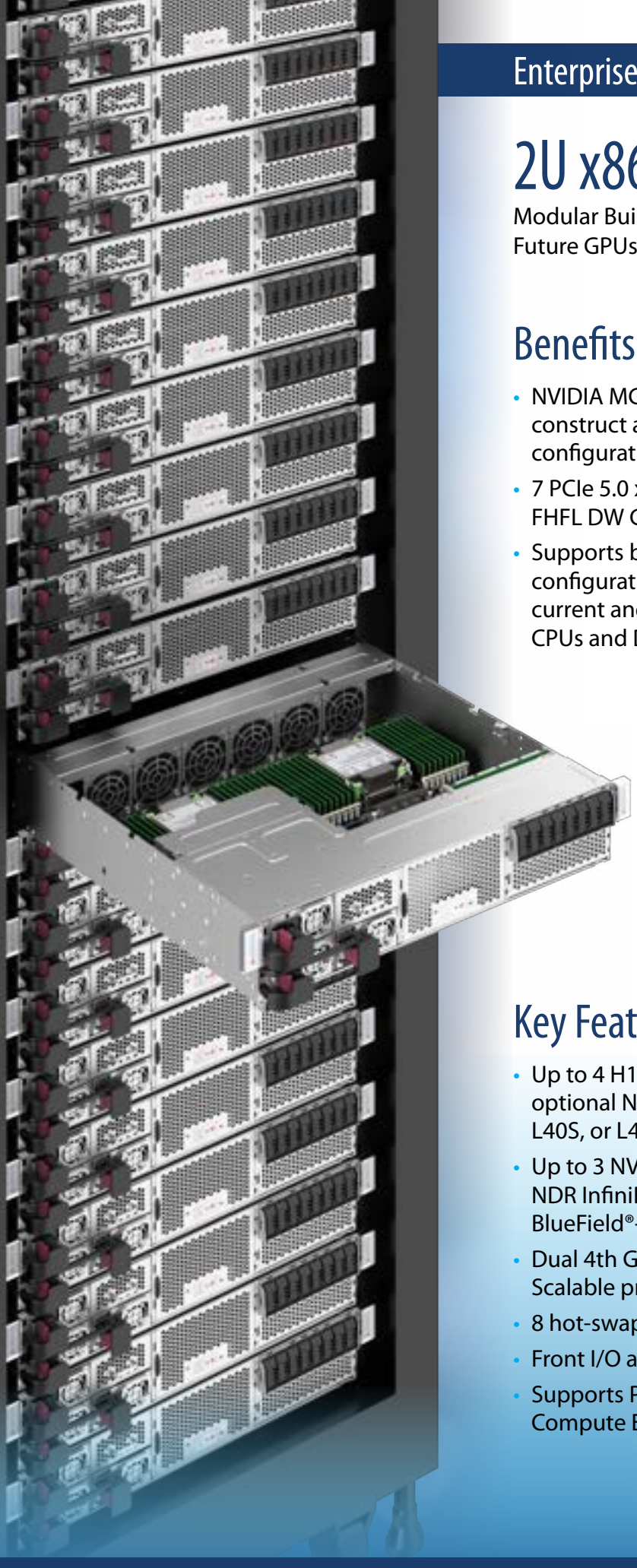
6U SuperBlade®
2 NVIDIA H100 PCIe
2 U.2 NVMe Drives
3 M.2 NVMe Drives
2 E1.S Drives
2x25GbE LOM
SBI-611E-5T2N

2U x86 MGX Systems

Modular Building Block Platform Supporting Today's and Future GPUs, CPUs, and DPUs

Benefits & Advantages

- NVIDIA MGX reference design enabling to construct a wide array of platforms and configurations
- 7 PCIe 5.0 x16 slots in 2U with up to 4 PCIe FHFL DW GPUs and 3 NICs or DPUs.
- Supports both ARM and x86-based configurations and is compatible with current and future generations of GPUs, CPUs and DPUs



2U MGX System
4 NVIDIA H100 PCIe or NVL
8 E1.S + 2 M.2 drives
16 DIMMs DDR5-4800
SYS-221GE-NR

Key Features

- Up to 4 H100 PCIe GPUs with optional NVLink Bridge (H100 NVL), L40S, or L40
- Up to 3 NVIDIA ConnectX-7 400G NDR InfiniBand cards or 3 NVIDIA BlueField®-3 cards
- Dual 4th Gen Intel Xeon Scalable processors
- 8 hot-swap E1.S and 2 M.2 slots
- Front I/O and Rear I/O configuration
- Supports PCIe 5.0 DDR5 and Compute Express Link 1.1+

2U Grace MGX System

Modular Building Block Platform with Energy-efficient Grace CPU Superchip

Benefits & Advantages

- Two NVIDIA Grace CPUs on one Superchip with 144-core and up to 500W CPU TDP
- 900GB/s NVLink® Chip-2-Chip (C2C) high-bandwidth and low-latency interconnect between Grace CPUs
- NVIDIA MGX reference design enabling to construct a wide array of platforms and configurations
- 7 PCIe 5.0 x16 slots in 2U with up to 4 PCIe FHFL DW GPUs and 3 NICs or DPUs.

2U Grace MGX System
4 NVIDIA H100 PCIe, NVL, or L40S
8 E1.S + 2 M.2 drives
960GB LPDDR5X
ARS-221GL-NR

Key Features

- Up to 144 high-performance Arm Neoverse V2 Cores with up to 960GB LPDDR5X onboard memory
- Up to 4 H100 PCIe GPUs with optional NVLink Bridge (H100 NVL), L40S, or L40
- Up to 3 NVIDIA ConnectX-7 400G NDR InfiniBand cards or 3 NVIDIA BlueField®-3 cards
- 8 hot-swap E1.S and 2 M.2 slots
- Front I/O and Rear I/O configuration

4 Visualization and Omniverse Workloads

Real-Time Collaboration, 3D Design, Game Development

Workload Sizes

Large

4U/5U 8 GPU
Tailored Architecture for
NVIDIA Omniverse™

Medium

2U Hyper
4 FHFL DW GPUs
Compute Optimized Architecture

GPU Workstation
4-GPU Rackmount/Full Tower

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

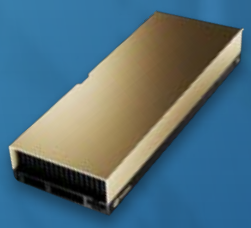
- NVIDIA OVX™ certified architecture
- NVIDIA L40S, L40, and RTX 6000 Ada GPUs
- NVIDIA BlueField®-2, or BlueField®-3 (DPU)
- NVIDIA RTX GPUs with ray tracing
- Rack-scale integration

Solution Stack

- Universal Scene Description Connectors
- NVIDIA Omniverse™ Enterprise

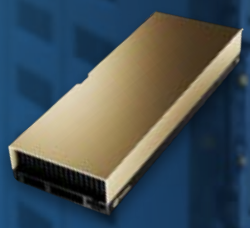
L40S

- FHFL DW
- PCIe 4.0 x16
- 350W
- 48GB GDDR6



L40

- FHFL DW
- PCIe 4.0 x16
- 300W
- 48GB GDDR6

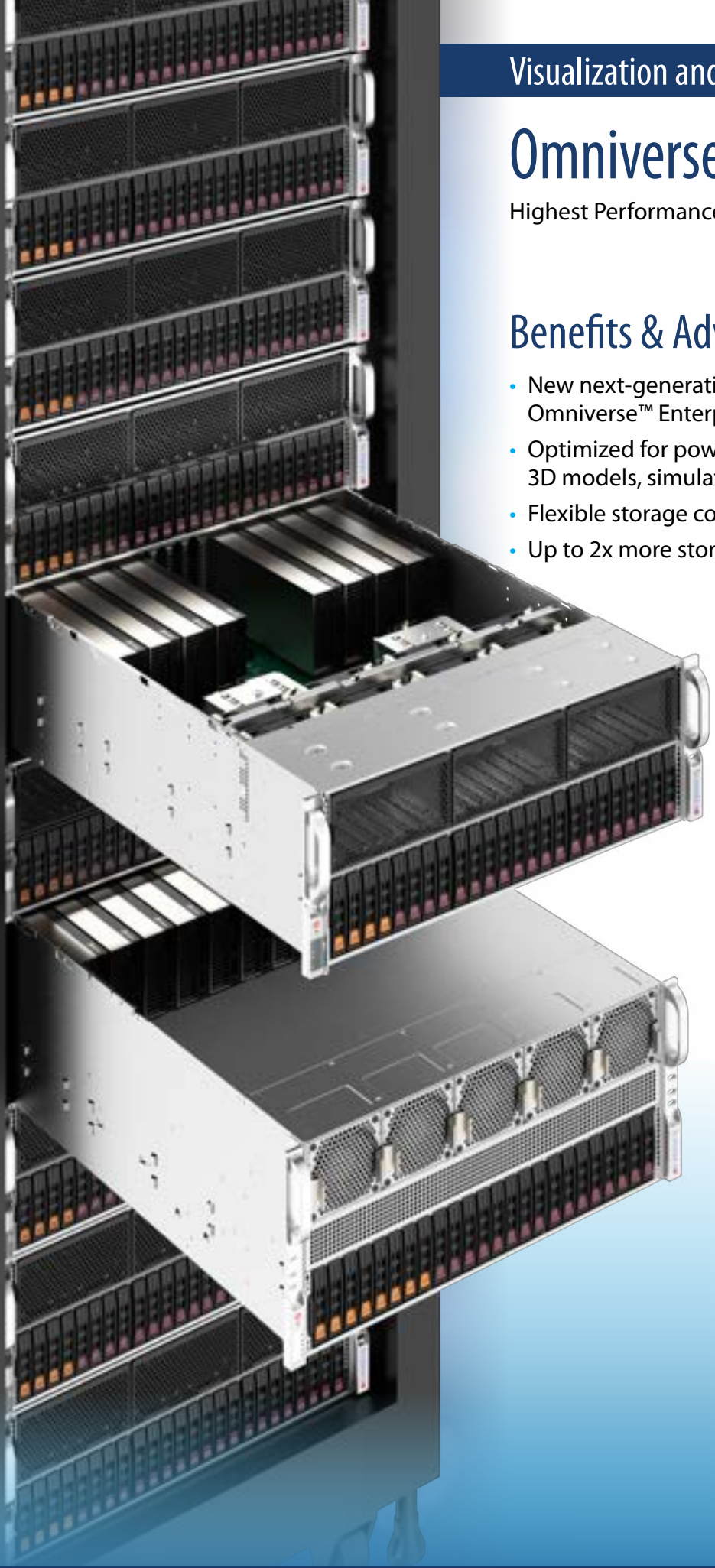


RTX 6000 ADA

- 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

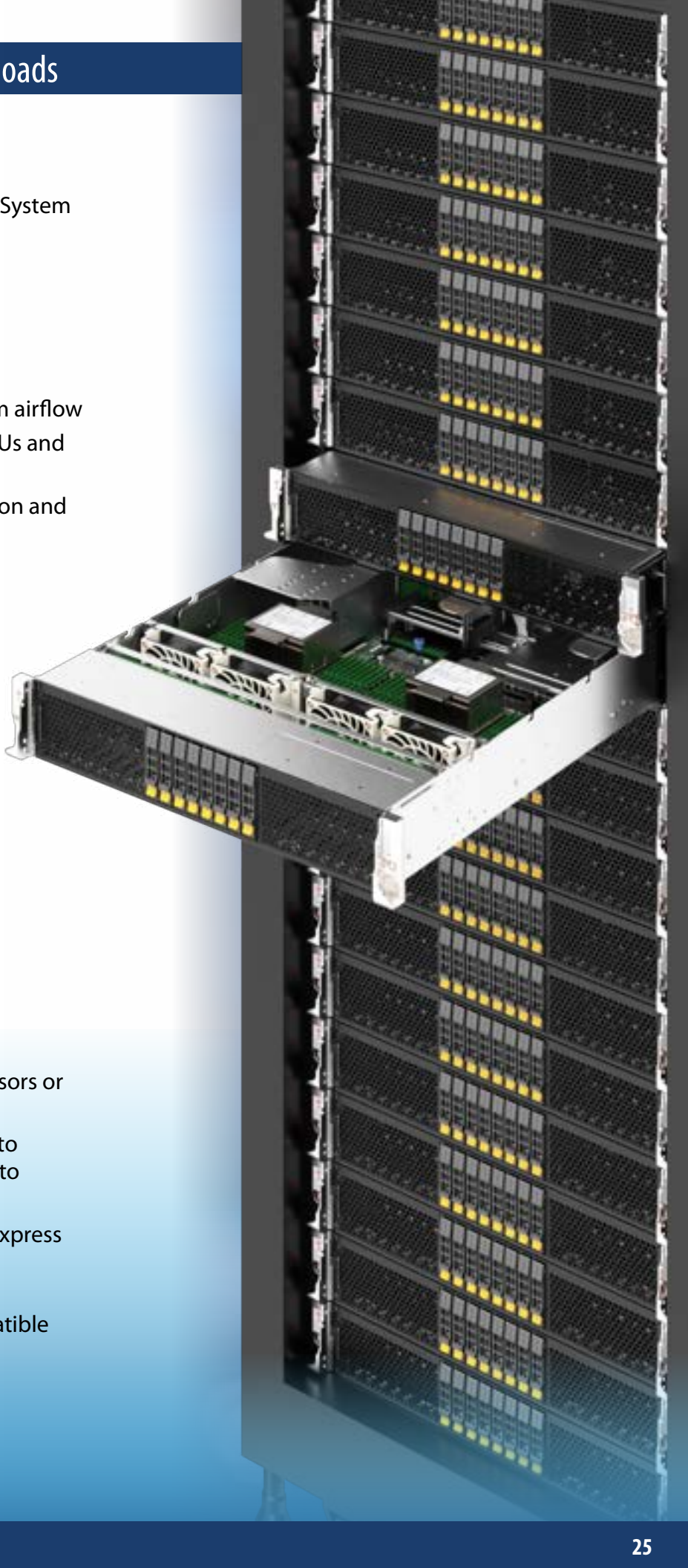
- 8 NVIDIA L40S/L40 PCIe GPUs
- Dual 4th Gen Intel® Xeon® Scalable processors or AMD EPYC™ 9004 Series processors
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- 3 NVIDIA ConnectX-7
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling.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

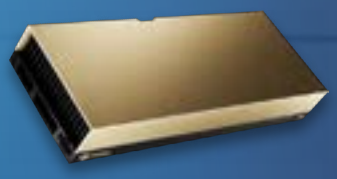
- GPU media engines with transcoding acceleration including AV1 encoding and decoding
- NVIDIA L40, L4, and RTX GPUs
- NVIDIA BlueField®-2 or BlueField-3 (DPU)
- Dense, resource-saving multi-node, multi-GPU systems for space and power efficiency
- High-capacity, high-throughput hot-swap storage

Solution Stack

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

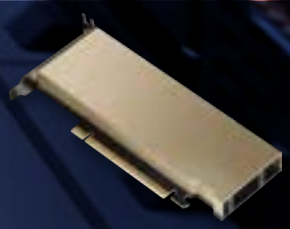
L40

- FHFL DW
- PCIe 4.0 x16
- 300W
- 48GB GDDR6



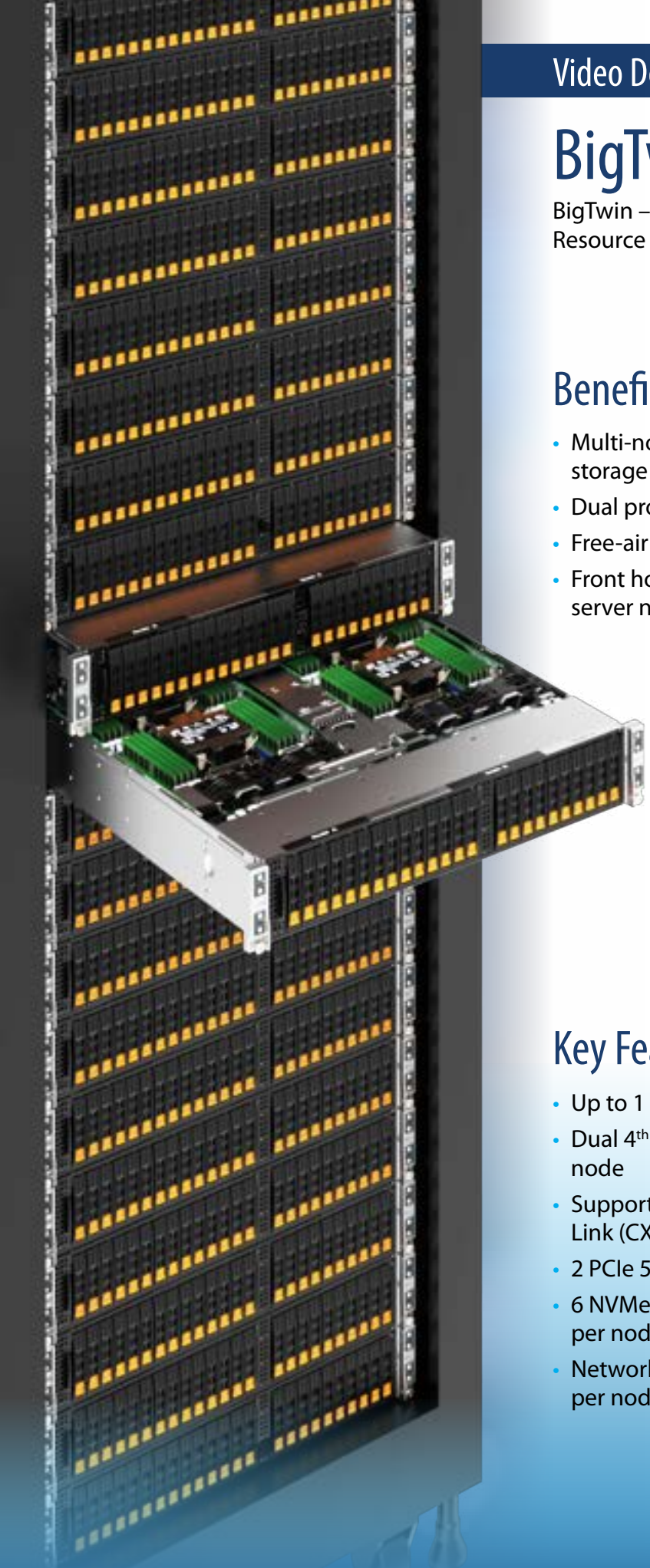
L4

- 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

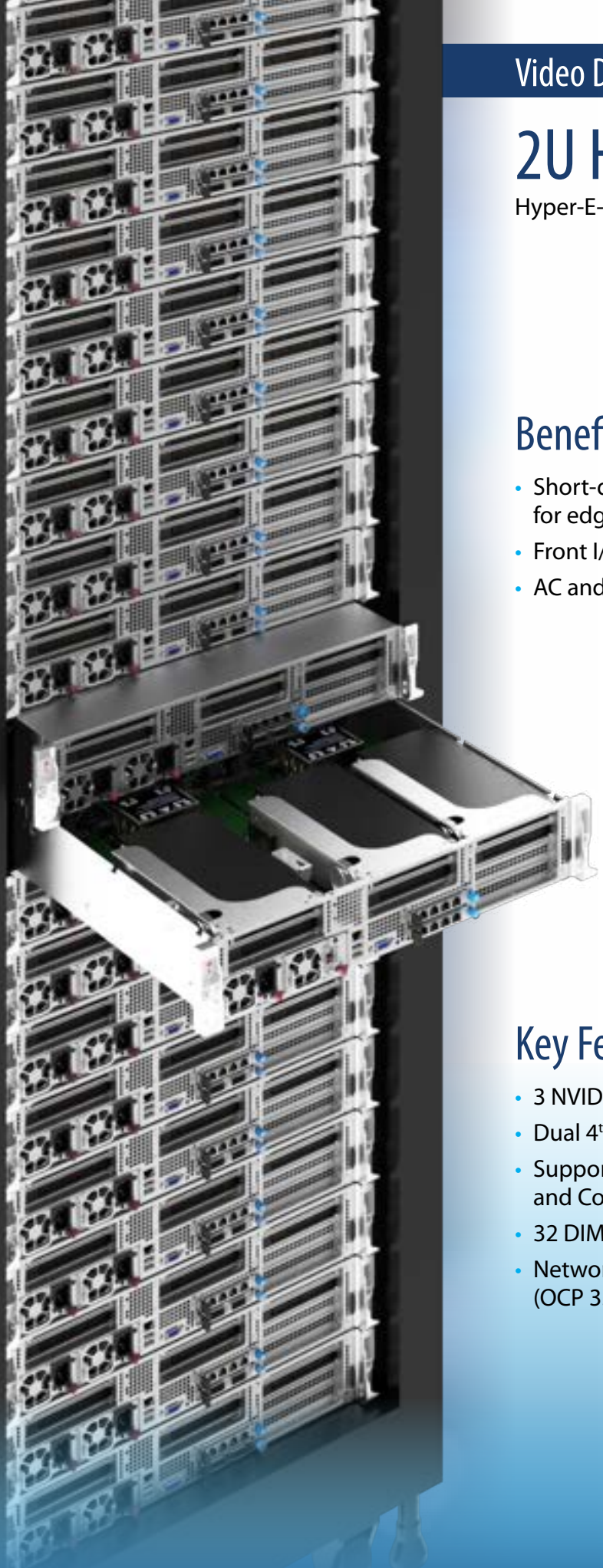
- Up to 6 GPUs
- Single 4th Gen Intel® Xeon® Scalable processor or AMD EPYC™ 9004 Series processor
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 350W with air cooling
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- 16 DIMM slots DDR5
- Advanced I/O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

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

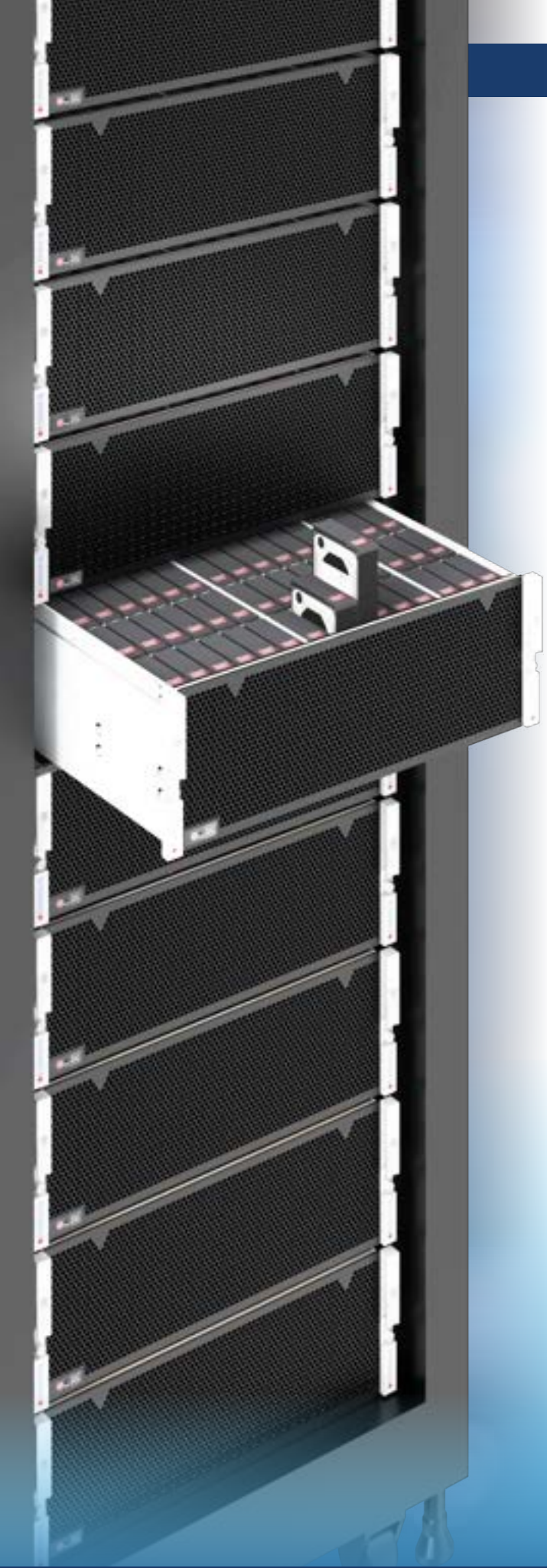
1U 24-Bay E1.S
SSG-121E-NES24R

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

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

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

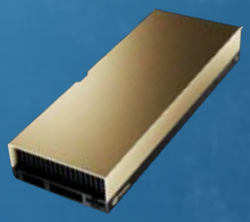
- CPU or GPU-based AI edge Inferencing, GPU-based AI edge training, and video transcoding/encoding/decoding
- NVIDIA L4, L40S, L40, A30, A40, T4, A2 GPUs
- Short-depth chassis design for edge locations with AC or DC power supply options
- Front I/O with broad range of expansion and I/O port for flexibility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

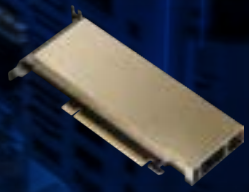


L40

- FHFL DW
- PCIe 4.0 x16
- 300W
- 48GB GDDR6

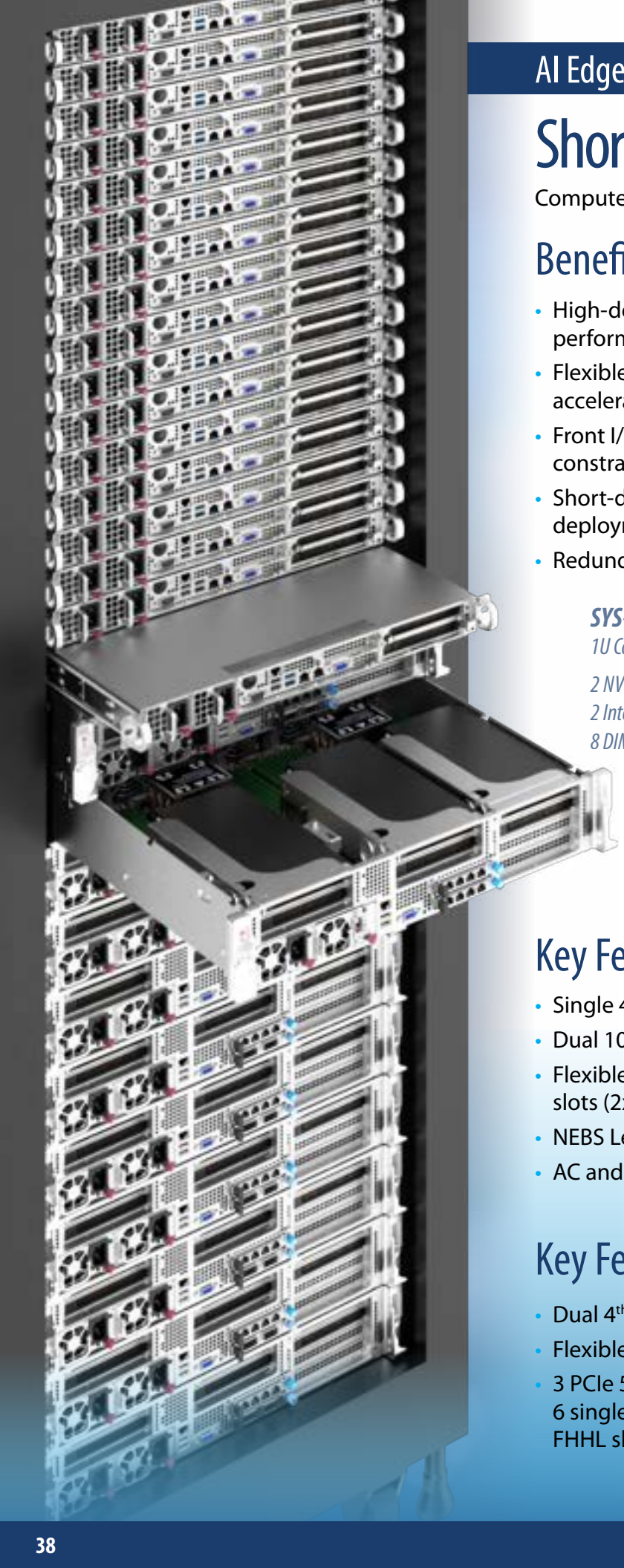
L4

- 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)

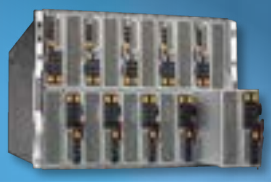
- 4th Gen Intel® Xeon® Scalable processor
- 3 PCIe 5.0 x16 FHFL slots
- Dual 10 GbE Ethernet
- Optional wall-mounted installation

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

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

TRAINING & INFERENCE

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

OMNIVERSE

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

VIDEO DELIVERY

L4
HHHL SW
PCIe 4.0 x16
72W
24GB GDDR6

EDGE

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-0002-11/2023-R47

www.supermicro.com