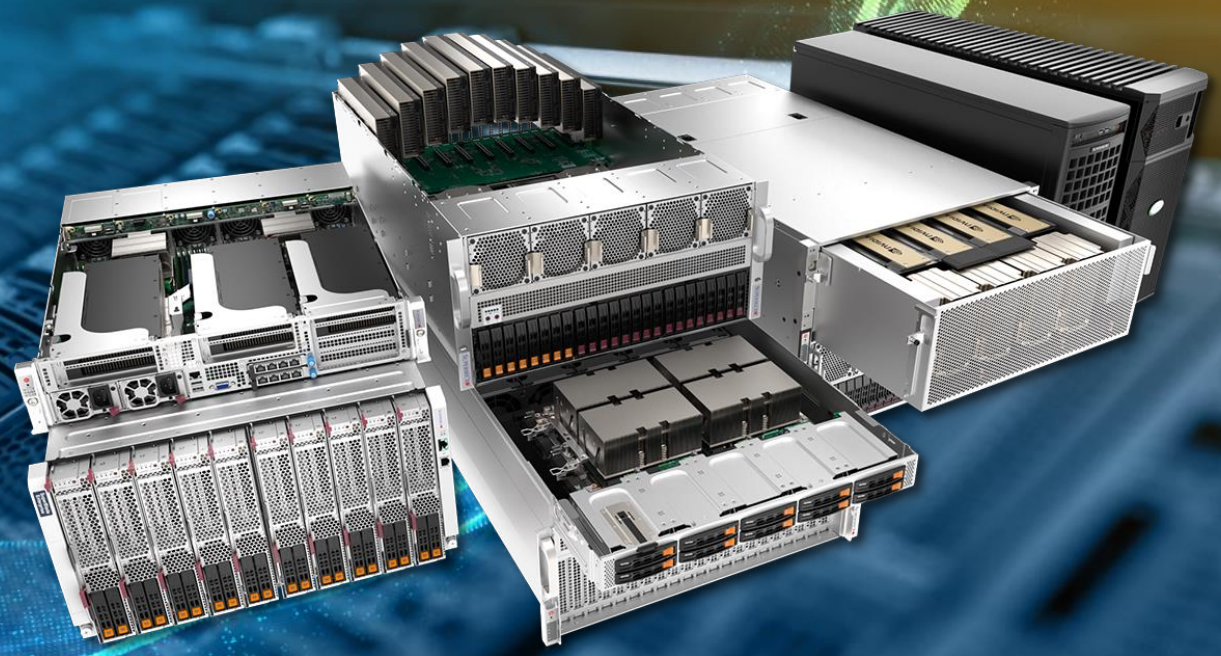
From Large Language Model to the AI Edge

GPU Acceleration for Broad Range of Workloads



Why is GPU Acceleration Booming Everywhere?

Parallel processing for complex problems that can be broken down into similar operations

AI, Machine Learning

~20X

Neural Networks are "embarrassingly parallel"

HPC, Scientific Computing

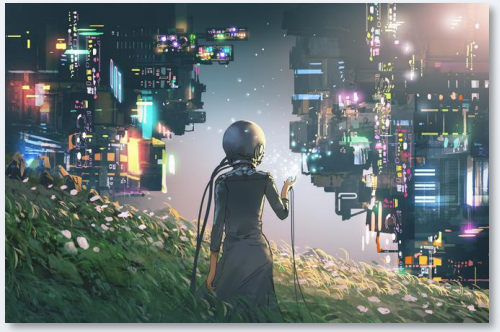
~40X

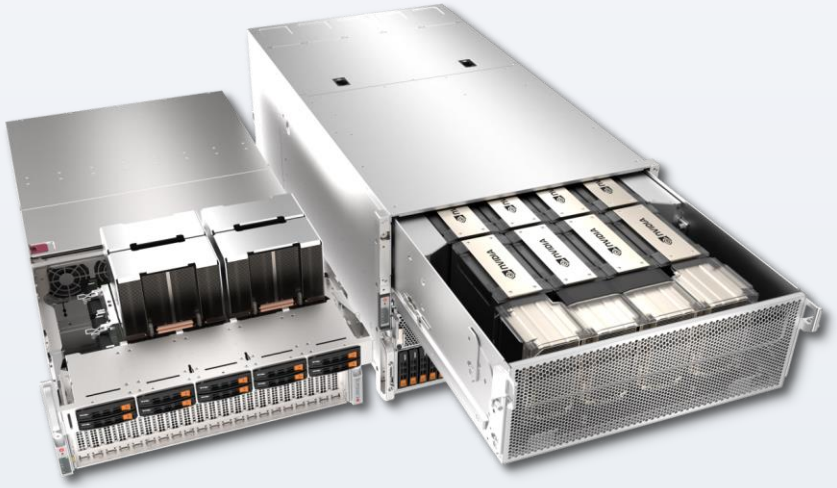
Genomic sequencing and analysis speed-up

Graphics, Rendering, Video

~100X

Real-time 3D graphics rendering, video encoding, and decoding



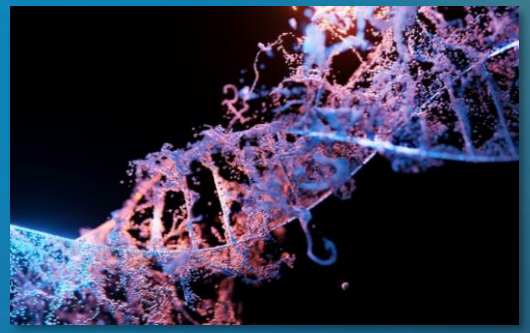


Accelerate Everything

GPU Optimized Systems to Achieve 5X, 10X,… 100X Performance

Large Scale AI Training Workloads

Large language models, Generative AI training, autonomous driving, robotics



HPC/AI Workloads

Engineering simulation, scientific research, genomic sequencing, drug discovery

Enterprise AI Inference & Training

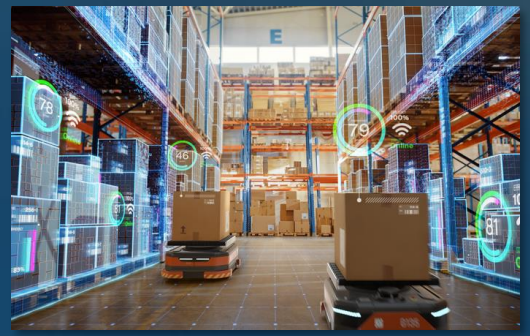
AI-enabled services/applications, chatbots, business automation

Visualization and Design

Graphical content development and automatic generation, digital twins, 3D collaboration

Content Delivery and Virtualization

Content delivery networks (CDNs), video transcoding, live streaming, VDI



AI Edge

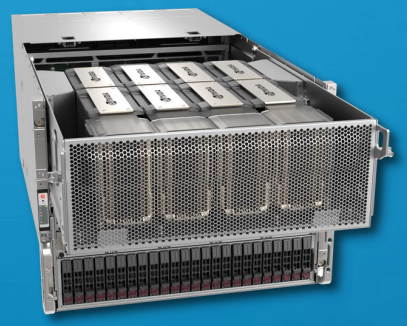
Retail automation, manufacturing/logistics automation, medical diagnosis/predictive care, security, and many more

What GPU Fits The Best for Your Workload?

	GPU	Memory (VRAM)	DL Training & DA	DL Inference	HPC / AI	Omniverse/ Render Farms	Virtual Workstation	Virtual Desktop(VDI)	Edge Accleration
Compute	HGX H100	80GB HBM3 per GPU	SXM ★★★	SXM ★★★	SXM ★★★				
	H100 NVL	94GB HBM3 per GPU	NVL ★★★	NVL ★★★	SXM ★★★				
	H100 PCIe	80GB HBM2e	PCIE ★★★	PCIE ★★★	PCIE ★★★				
	A100	80GB HBM2e	SXM PCIE ★★	SXM PCIE ★★	SXM PCIE ★★				
Graphics/ Compute	L40S	48GB GDDR6	★★	★★★	★	★★★	★★★		★
	L40	48GB GDDR6	★	★★		★★★	★★★		★
	RTX 6000 ADA	48GB GDDR6	★	★★		★★	★★★	★★★	
Small Form Factor Compute /Graphics	L4	24GB GDDR6 72W		★★		★★	★★★	★★★	★★★
	T4	16GB GDDR6 70W		★			★	★	★

GPU Optimized Systems by Workloads

Large Scale AI Training



8U 8-GPU System (HGX H100 SXM)
(codenamed: Delta-Next)
SYS-821GE-TNHR, AS -8125GS-TNHR

4U 4-GPU System (HGX H100 SXM)
(codenamed: Redstone-Next)
SYS-421GU-TNXR, SYS-521GU-TNXR

Petabyte Scale All-Flash Storage
SSG-121E-NE316R, ASG-1115S-NE316R

HPC/AI Workloads

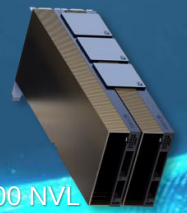
4U 4-GPU System (HGX H100 SXM)
SYS-421GU-TNXR

8U SuperBlade (Up to 20 nodes)
SBI-411E-1G / SBI-411E-5G

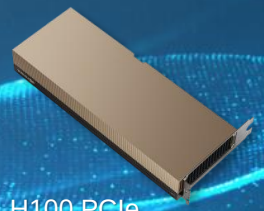
4U/5U 8-10 GPU System
SYS-521GE-TNRT, SYS-421GE-TNRT/TNRT3
AS -4125GS-TNRT/TNRT1/TNRT2

1U Grace Hopper MGX System
SYS-421GU-TNXR / SYS-521GU-TNXR

HGX H100 SXM
8-GPU or 4-GPU



H100 NVL



H100 PCIe

Grace Hopper Superchip
(Grace CPU + H100 GPU)

GPU Optimized Systems by Workloads

Enterprise AI Inference & Training

4U/5U 8-10 GPU System
SYS-521GE-TNRT, SYS-421GE-TNRT/TNRT3
AS -4125GS-TNRT/TNRT1/TNRT2

6U SuperBlade (Up to 10 GPUs)
SBI-611E-5T2N

2U MGX System (Up to 4 GPUs)
SYS-221GE-NR

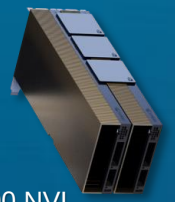
2U Grace MGX System (Up to 4 GPUs)
ARS-221GL-NR

Visualization and Design

4U/5U 8-10 GPU System
(NVIDIA OVX™ reference design available)
SYS-521GE-TNRT, SYS-421GE-TNRT, AS -4125GS-TNRT

GPU Workstation (Up to 4 GPUs)
SYS-741GE-TNRT, AS -5014A-TT

2U Hyper (Up to 4 GPUs)
SYS-221H-TNR, AS -2015HS-TNR



H100 NVL

H100 PCIe

L40S

L40

L40S

L40

RTX 6000 Ada

GPU Optimized Systems by Workloads

Content Delivery and Virtualization

2U 4-Node BigTwin
(Up to 2 SW GPUs per node)
SYS-221BT-HNTR, SYS-621BT-HNTR

2U CloudDC
(Up to 2 DW or 4 SW GPUs)
SYS-521C-NR, AS -2015CS-TNR

2U Hyper-E Short-Depth
(Up to 3 DW GPUs or 4 SW GPUs)
SYS-221HE-FTNR, SYS-221HE-FTNRD

AI Edge

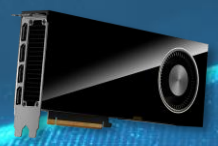
2U Hyper-E Short-Depth
(Up to 3 DW GPUs or 4 SW GPUs)
SYS-221HE-FTNR, SYS-221HE-FTNRD

Compact Fanless Edge Server
(Up to 3 SW GPUs)
SYS-E403-13E

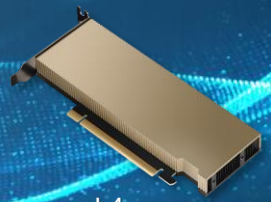
1U Compact Short-Depth Edge/5G Server
(Up to 2 SW GPUs)
SYS-111E-FWTR

Embedded Fanless Edge Server
(CPU or ASIC based Inference)
SYS-E100-13AD

L40



RTX 6000 Ada



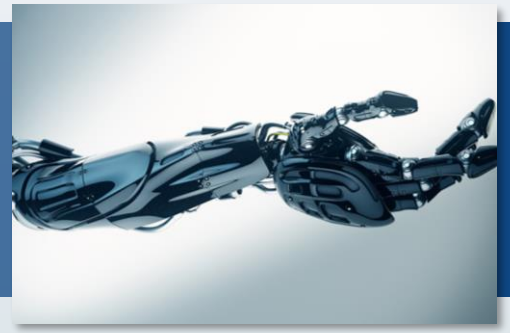
L4

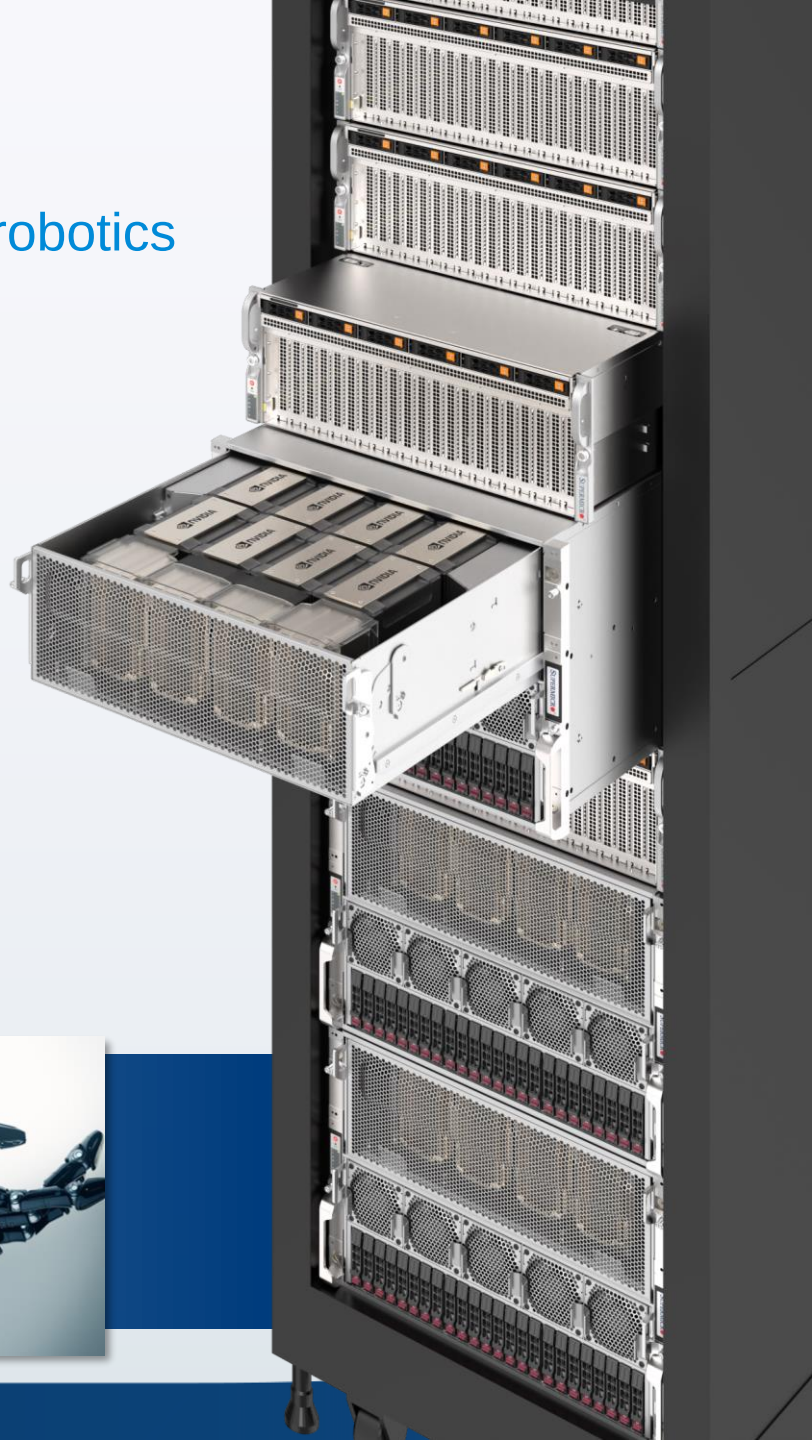
Large Scale AI Training

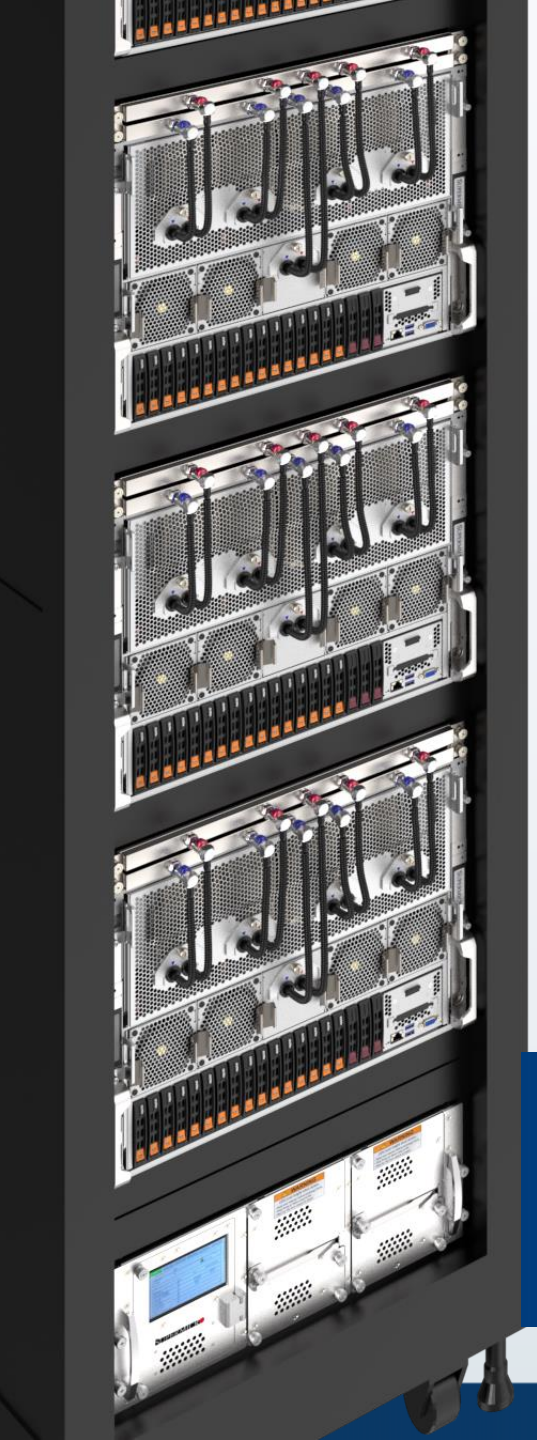
Large language models, generative AI training, autonomous driving, robotics

Opportunities and Challenges

- Pool of 10,000+ GPUs and GPU memory to fit large AI models to maximize parallel computing and minimize training time
- Training with massive amount of data with continuous growth of data size (e.g., over 1 trillion tokens)
- Serve AI models (inference) to millions of concurrent users
- High performance everything: GPUs, memory, storage, and network fabric







Large Scale AI Training

Key Technologies

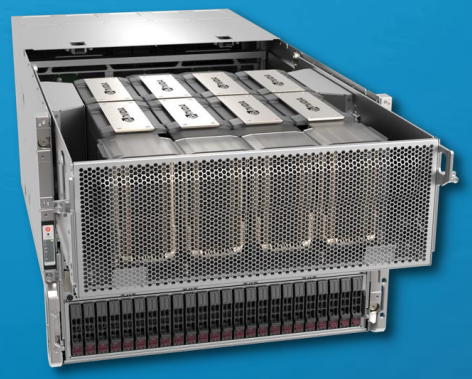
- NVIDIA HGX H100 SXM 8-GPU/4-GPU with 900GB/s NVLink interconnect
- Dedicated, lots of high performance, high bandwidth GPU memory - HBM3, HBM2e
- 400GbE networking (Ethernet or InfiniBand), PCIe 5.0 storage for fast AI data pipe
- NVIDIA GPUDirect RDMA and Storage to keep feeding data to GPUs with minimum latency
- Liquid cooling for GPUs and CPUs
- All-flash storage and file systems to support petabytes of hot-tier data cache

- NVIDIA HGX H100 SXM5 board with 4- GPU or 8-GPU
- NVLink and NVSwitch
- 80GB HBM3 per GPU
- Up to 700W TDP

- NVIDIA ConnectX-7
- Up to 400GbE or 400G NDR InfiniBand
- x16/x32 PCIe 5.0

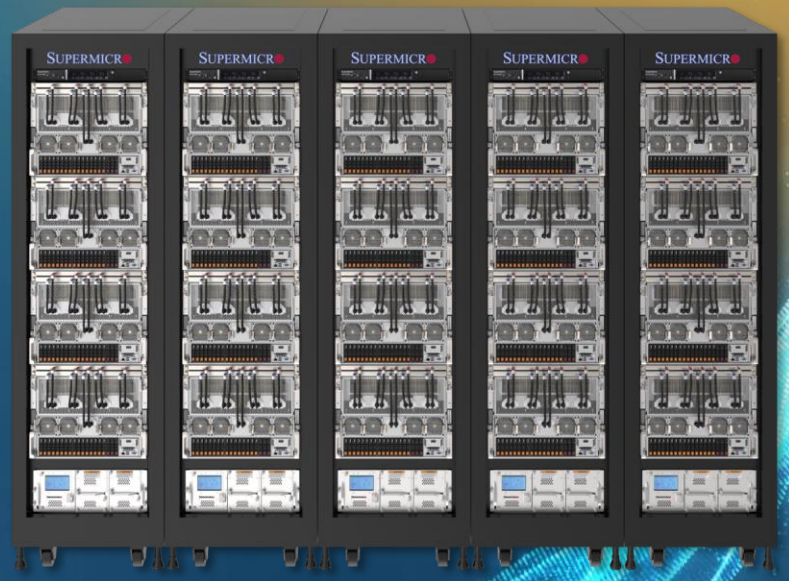
GPU Optimized Systems by Workloads

Large Scale AI Training



8U 8-GPU System (HGX H100 SXM)
(codenamed: Delta-Next)
SYS-821GE-TNHR, AS -8125GS-TNHR

4U 4-GPU System (HGX H100 SXM)
(codenamed: Redstone-Next)
SYS-421GU-TNXR, SYS-521GU-TNXR



Liquid-cooled AI Rack Integrated Solutions
SYS-821GE-TNHR, AS -8125GS-TNHR

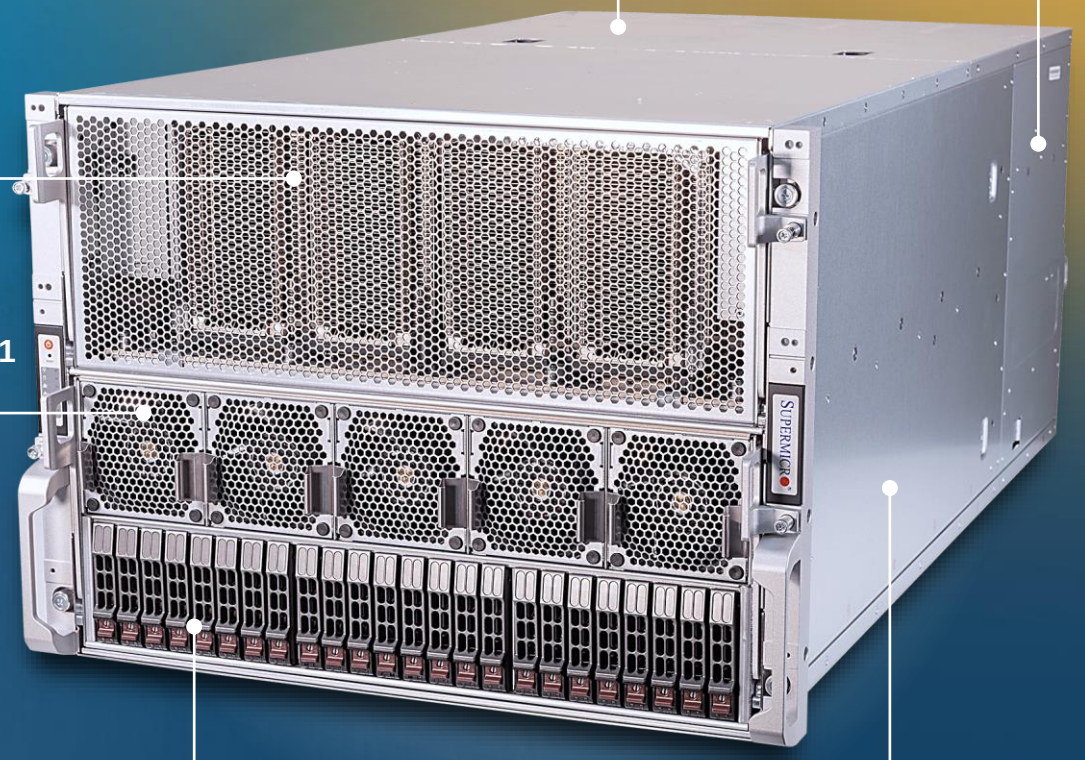
Petabyte Scale All-Flash Storage
SSG-121E-NE316R, ASG-1115S-NE316R

HGX H100 SXM
8-GPU or 4-GPU

8U HGX H100 8-GPU System
(codenamed: Delta-Next)

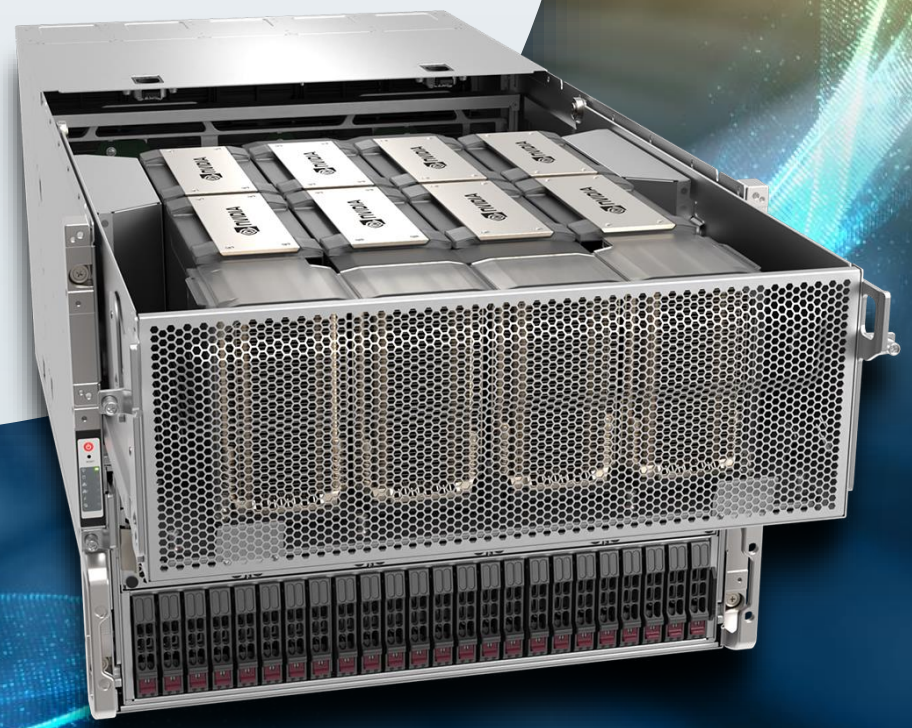
SYS-821GE-TNHR or AS -8125GS-TNHR

- 900GB/s GPU interconnect – 10x better performance than PCIe
- 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
- Liquid cooling options for both GPUs and CPUs to optimize performance and energy cost



Success & Use Cases

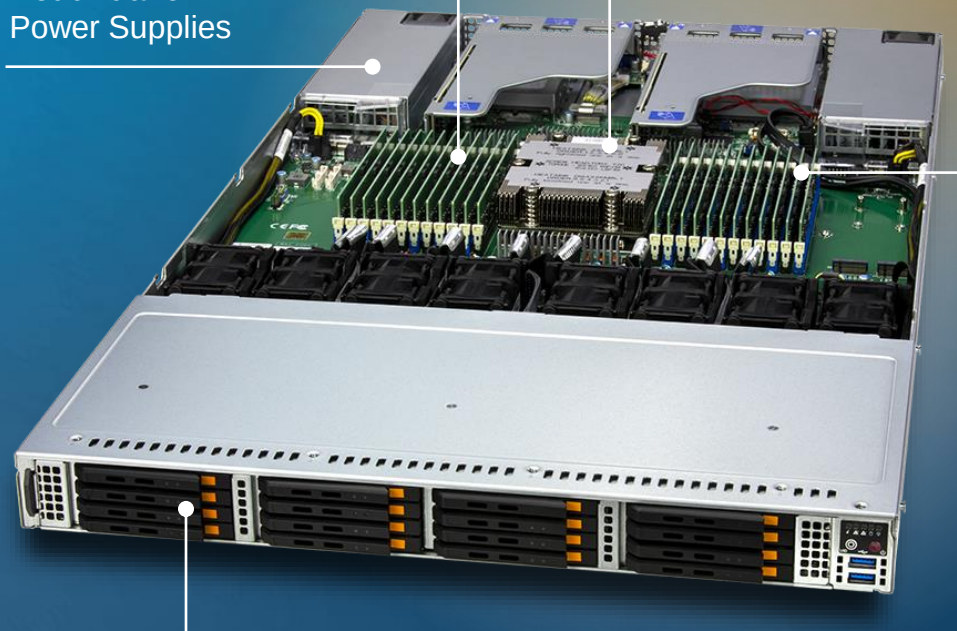
- Cloud Computing – 1000s of 8U systems deployed
- Online Businesses – goods and contents recommendations and personalization
- Automotive Industry – computer vision, autonomous driving, 1000-2000 systems
- Social Media – content recommendation, use profiling
- Telco – chatbot for customer support
- Financial Services - retraining GPT-3 level model with 50B parameters for inquiry services

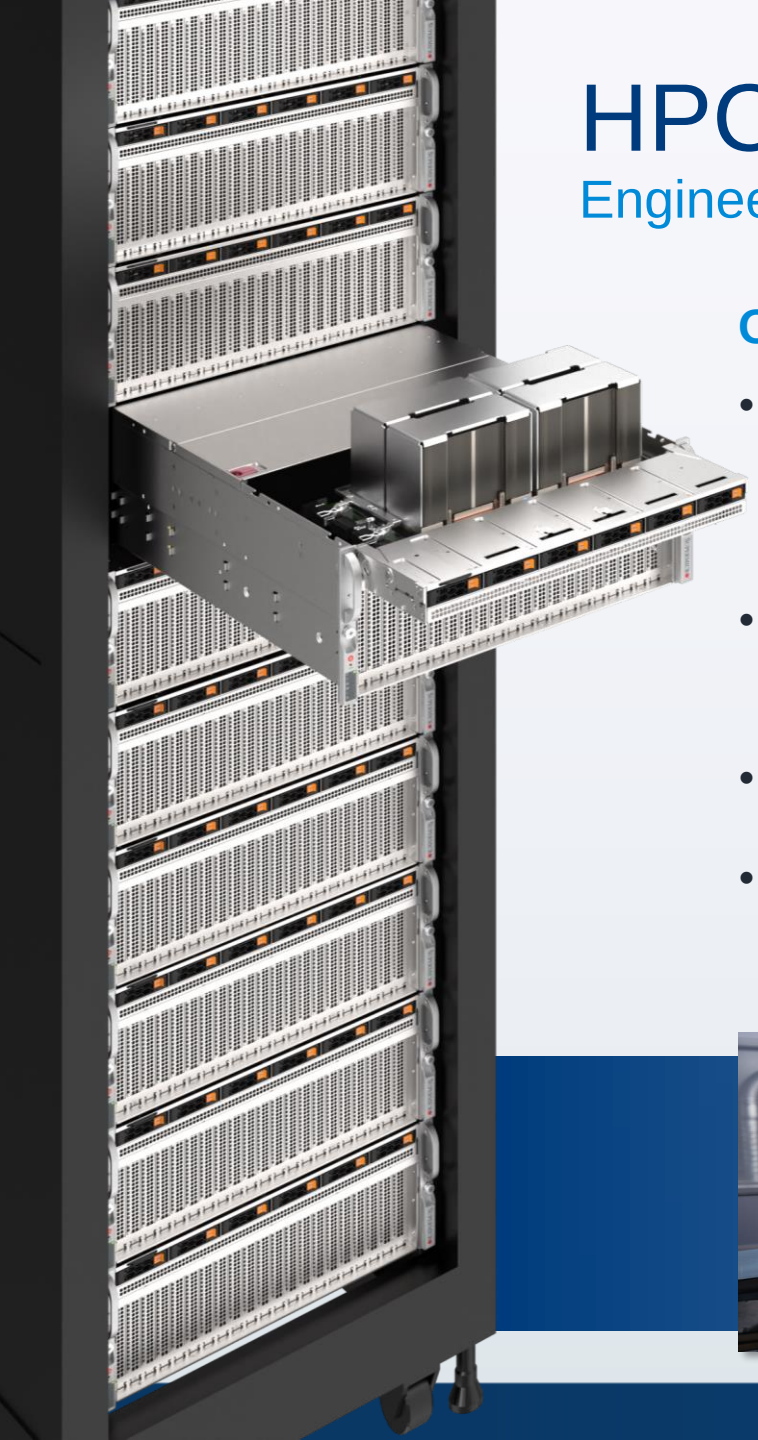


Petascale NVMe Storage

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

- Direct-attached EDSFF E3.S PCIe 5.0 media for the best thermal and I/O performance
- Dual 4th Gen Intel Xeon Scalable or single AMD EPYC 9004 Series processor
- Up to 32 E3.S NVMe drives in 2U
- Up to 2 x16 PCIe 5.0 slots + 2 AIOM slots
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- Optional 4 CXL E3.S 2T form factor memory expansion modules + 8 E3.S NVMe storage configuration





HPC/AI Workloads

Engineering simulation, scientific research, genomic sequencing, drug discovery

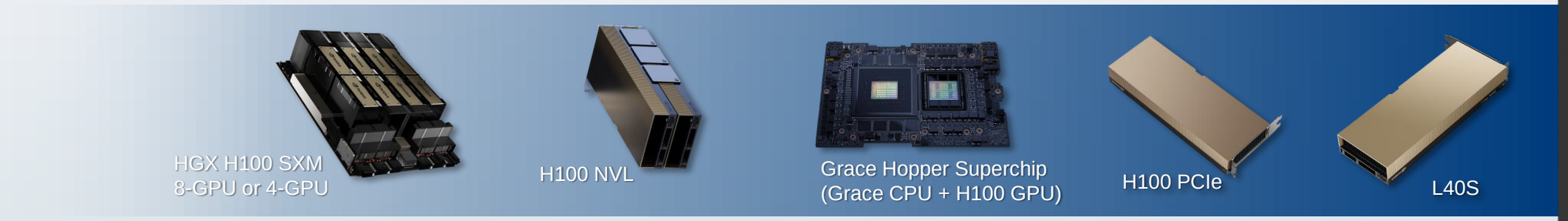
Opportunities and Challenges

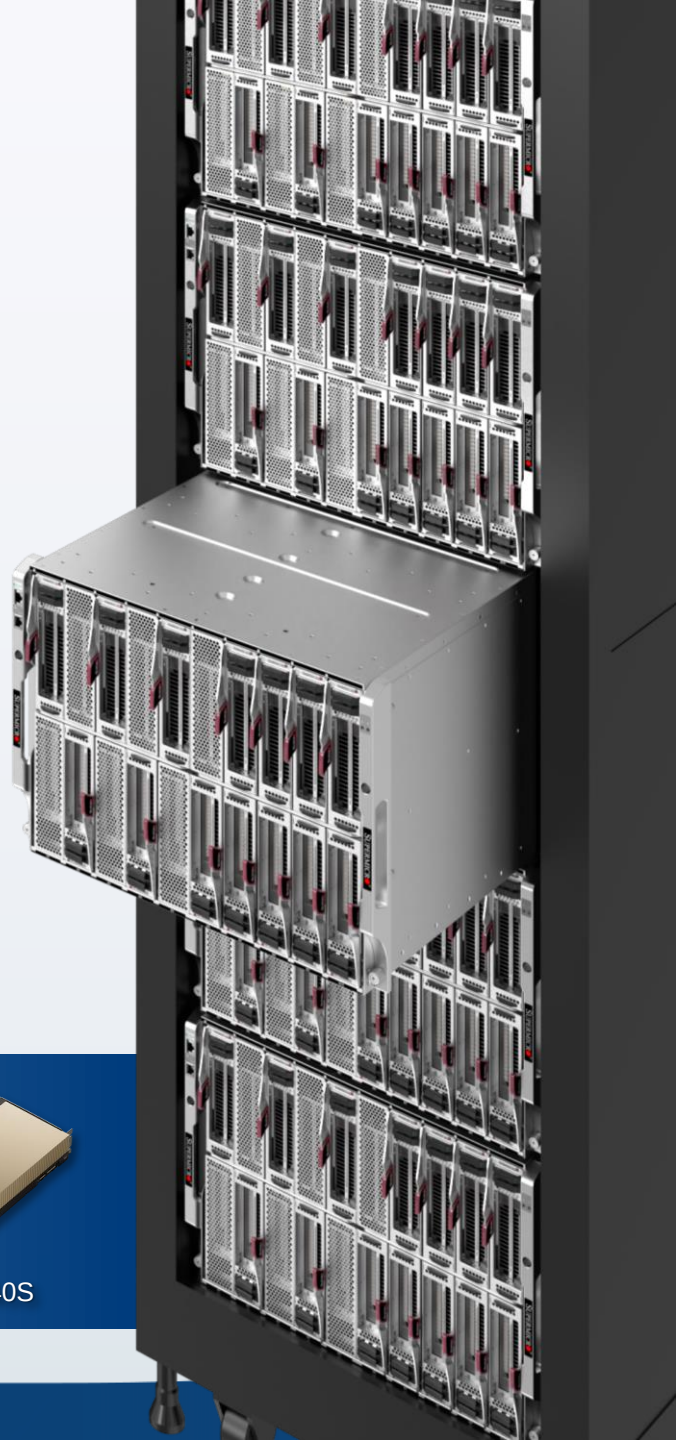
- Augmenting machine learning algorithms and GPU accelerated parallel computing to HPC workloads to achieve faster results and discoveries
- Parallel processing with massive datasets for data-intensive simulations and analytics
- Simulations requiring double precision (FP64)
- High-resolution and real-time visualization of scientific simulations and modeling

HPC/AI Workloads

Key Technologies

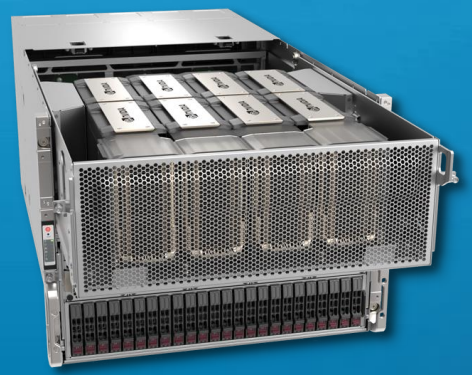
- Double-precision Tensor Cores delivering 535/268 teraFLOPs with HGX H100 SXM 8-GPU/4-GPU, or 134 teraFLOPs with H100 NVL (2 GPUs with NVLink Bridge) at FP64
- High CPU compute and high GPU compute – e.g, up to 20 CPUs and 20 GPUs in 8U
- High bandwidth GPU memory and CPU cache/integrated memory – HBM3, HBM2e
- GPU-GPU Interconnect (NVLink) and 400GbE networking for clustering, PCIe 5.0 storage
- Liquid cooling for GPUs and CPUs





GPU Optimized Systems by Workloads

HPC/AI Workloads



8U 8-GPU System (HGX H100 SXM)
(codenamed: Delta-Next)
SYS-821GE-TNHR, AS -8125GS-TNHR

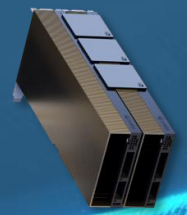
4U 4-GPU System (HGX H100 SXM)
SYS-421GU-TNXR

4U/5U 8-10 GPU System
SYS-521GE-TNRT, SYS-421GE-TNRT/TNRT3
AS -4125GS-TNRT/TNRT1/TNRT2

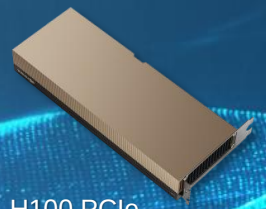
8U SuperBlade (Up to 20 nodes)
SBI-411E-1G / SBI-411E-5G

1U Grace Hopper MGX System
SYS-421GU-TNXR / SYS-521GU-TNXR

HGX H100 SXM
8-GPU or 4-GPU



H100 NVL



H100 PCIe

Grace Hopper Superchip
(Grace CPU + H100 GPU)

8-10 GPU Systems

SYS-521GE-TNRT

- Up to 8 or 10 PCIe GPUs with optional NVLink Bridge (e.g., H100 NVL)
- Dual Root Configuration
- Dual 4th Gen Intel® Xeon® Scalable
- 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

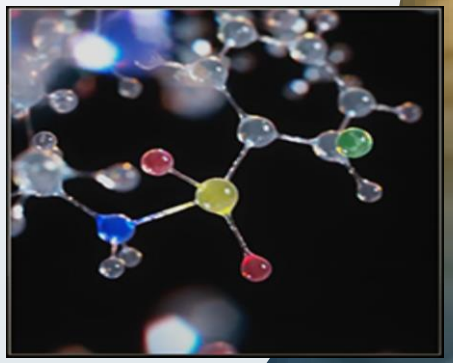
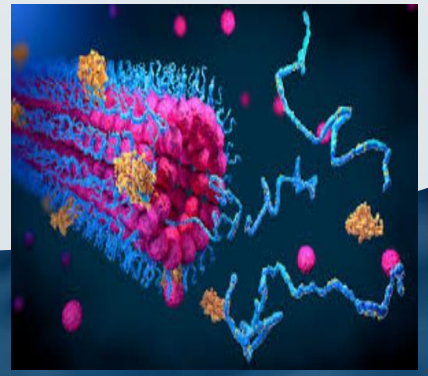
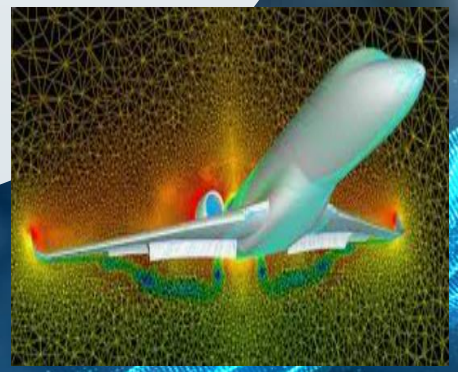
8U SuperBlade®

SBI-411E-1G/5G

- 1 (SW blade) or 2 (DW blade) PCIe GPUs including H100, H100 NVL, L40S
- 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 and cooling, and integrated switch for maximum efficiency with optional liquid cooling
- 2-port 25GbE (3rd and 4th LAN), 1x 200G HDR InfiniBand or 1 x 100G EDR InfiniBand via mezzanine card

Success & Use Cases

- Research Labs – e.g., accelerating “particle accelerators”
- Climate Modeling
- Drug Discovery
- Computational Fluid Dynamics
- Seismic imaging and analysis
- Materials science and engineering
- Astrophysical simulation

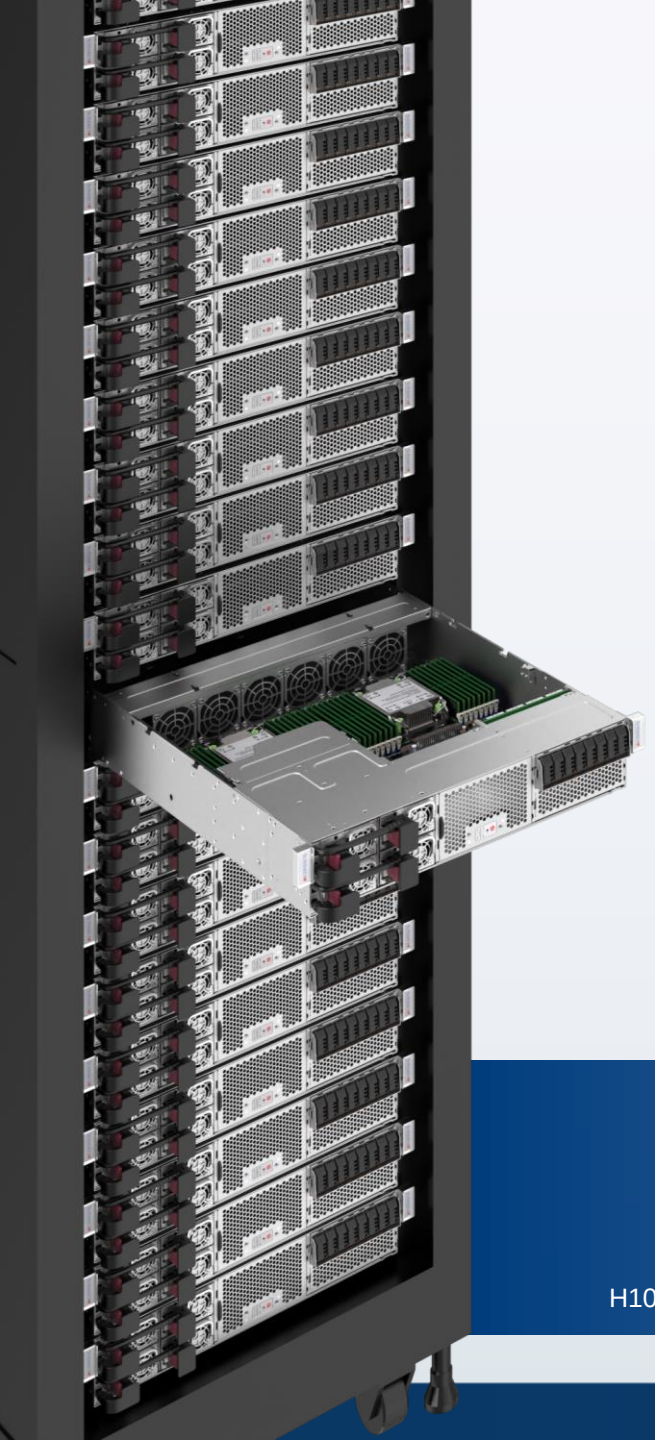





Enterprise AI Inference and Training

AI-enabled services/applications, chatbots, business automation

Opportunities and Challenges

- AI adoption across industries to boost productivity, streamline operations, make data-driven decisions, and improve customer experience
- Open architecture, vendor flexibility, fast/easy deployment for rapidly evolving technologies
- High computational and resource costs, cloud vs. on-prem
- Utilization of frameworks, pre-trained models, open-source AI models with fine-tuning and embeddings (with their own dataset)



Enterprise AI Inference and Training

AI-enabled services/applications, chatbots, business automation

Key Technologies

- Flexible, modular, highly configurable rackmount servers with different form factors to balance compute, storage, networking, and cost for various enterprise AI workload needs for today and the future
- PCIe 5.0 supported platforms for future proofing – GPUs, storage, networking
- FP8 and FP16 support to boost performance with less resources and cost
- Intel, AMD, ARM CPU options
- NVIDIA Certified with NVIDA AI Enterprise and NGC catalog to fully leverage pre-trained models and optimized libraries and toolset

GPU Optimized Systems by Workloads

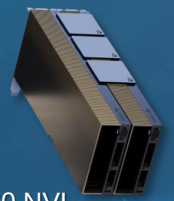
Enterprise AI Inference & Training

4U/5U 8-10 GPU System
SYS-521GE-TNRT, SYS-421GE-TNRT/TNRT3
AS -4125GS-TNRT/TNRT1/TNRT2

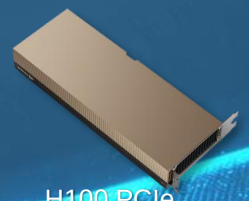
6U SuperBlade (Up to 10 GPUs)
SBI-611E-5T2N

2U MGX System (Up to 4 GPUs)
SYS-221GE-NR

2U Grace MGX System (Up to 4 GPUs)
ARS-221GL-NR



H100 NVL



H100 PCIe

L40S

L40

8-10 GPU Systems

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

- Up to 8 or 10 PCIe GPUs with optional NVLink Bridge (e.g., H100 NVL)
- Single Root , Dual Root, Direct Connect configuration available depending on workload requirements
- Dual 4th Gen Intel® Xeon® Scalable 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

13 PCIe 5.0 Slots with Up to 10 GPUs
+ I/O and networking

Single Root, Dual Root and Direct-Connect Options

2.5″ Drive Bays
Up to 24 drives with Direct-to-CPU option

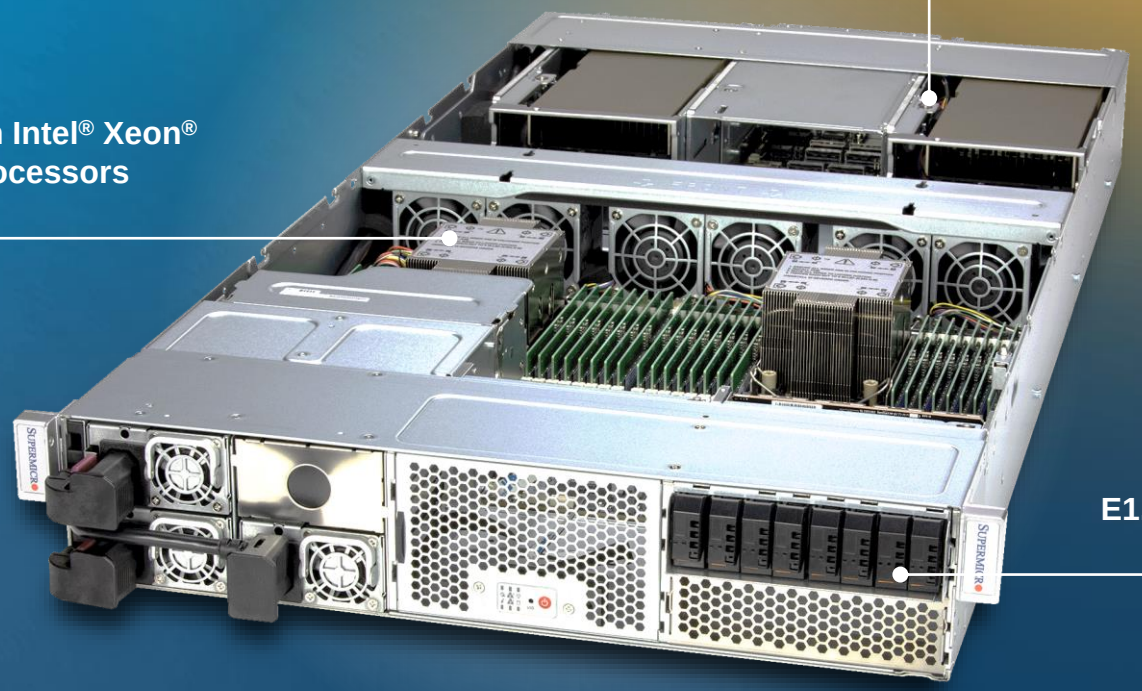
2U MGX Systems

SYS-221GE-NR / ARS-221GL-NR

- 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 (SYS-221GE-NR) or 2 NVIDIA Grace CPUs integrated board with up to 960GB LPDDR5X onboard memory (ARS-221GL-NR)
- 8 hot-swap E1.S and 2 M.2 slots
- Front I/O and Rear I/O configuration
- Compatible with current and future generations of GPUs, CPUs, and DPUs

PCIe GPUs
Up to 4 NVIDIA H100, H100 NVL, L40S

Dual 4th Gen Intel® Xeon® Scalable Processors
up to 350W

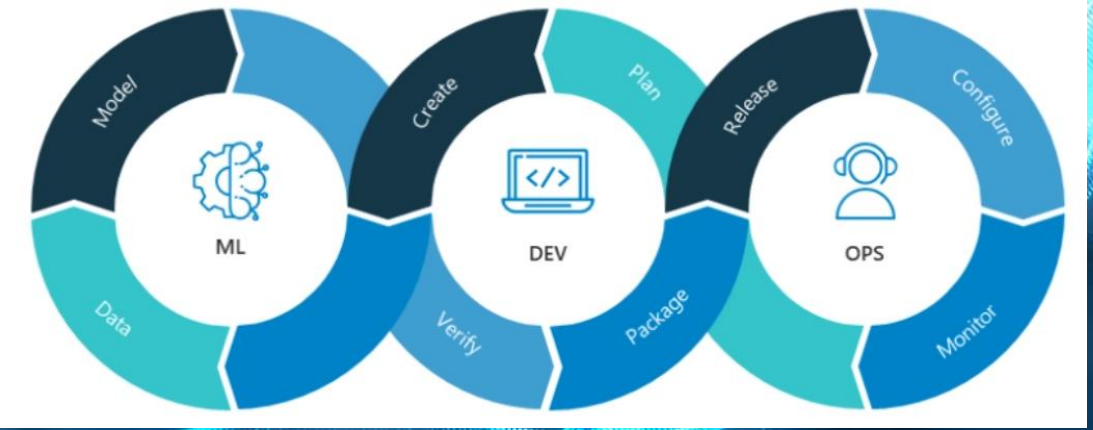


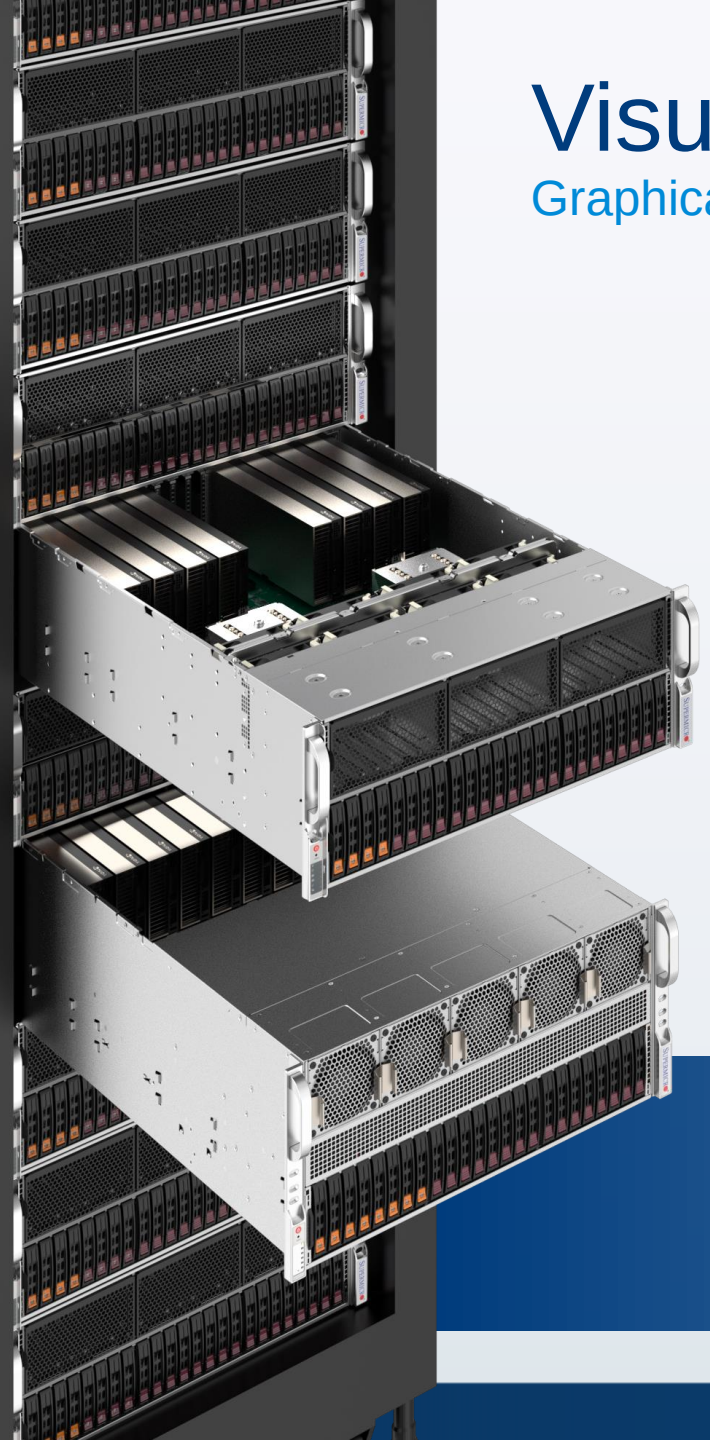
E1.S Drive Bays
Up to 8 Drives

Final system configuration subject to change

Use Cases

- MLOps Data Science in Production at Scale
- Best Practices For Businesses to Run AI Successfully
- A First-Principles Approach to Machine Learning Production
- Machine Learning Platform for AI lets enterprises quickly create and deploy machine learning experiments to achieve business objectives.





Visualization and Design

Graphical content development and automatic generation, digital twins, 3D collaboration

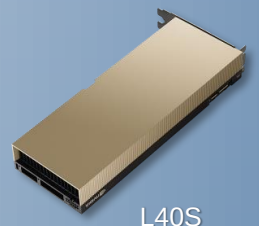
Opportunities and Challenges

- AI-aided 3D graphics, game development, creative asset generation
- Digitizing industrial design and productization process with virtualized real-world scenarios
- Integrated engineering and enterprise-scale simulations
- Cloud and virtual collaboration with low latency

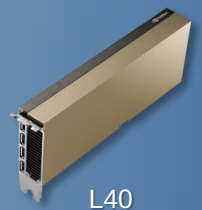
Visualization and Design

Key Technologies

- NVIDIA OVX reference architecture supporting NVIDIA Omniverse Enterprise, Universal Scene Description (USD) connectors
- NVIDIA RTX GPUs with ray tracing for photo realistic visuals
- NVIDIA BlueField 2, 3 (DPU) for low latency, secure and fast data management
- Multi-GPU workstation or virtualized workstations
- Rack-scale integration for virtual production and collaboration infrastructure, speedy rendering, fast and secure data storing and transfer



L40S



L40

RTX 6000 Ada

GPU Optimized Systems by Workloads

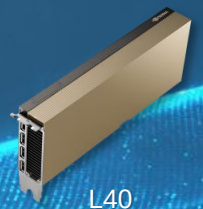
Visualization and Design

4U/5U 8-10 GPU System
(NVIDIA OVX™ reference design available)
SYS-521GE-TNRT, SYS-421GE-TNRT, AS -4125GS-TNRT

2U Hyper (Up to 4 GPUs)
SYS-221H-TNR, AS -2015HS-TNR

GPU Workstation (Up to 4 GPUs)
SYS-741GE-TNRT, AS -5014A-TT

L40S



L40

RTX 6000 Ada

Use Cases

BMW Group

- 31 factories around the world
- 99 percent vehicles produced are custom configurations
- 40 BMW New Models
- 100 options for each car
- 2,100 possible configurations

BIG Challenge Keeping Materials Stocked on the Assembly Line

NVIDIA Omniverse Enterprise is enabling digital twins at BMW

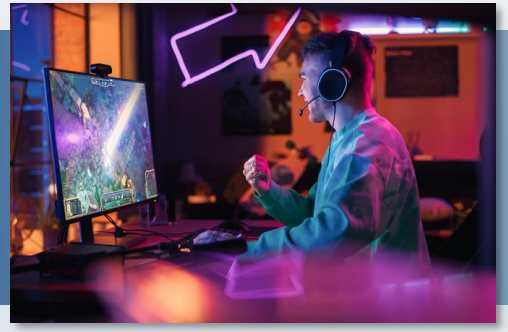
- Run factory simulations to optimize its operations
- Deploy fleet of robots for logistics
- Improved the distribution of materials, production environment

Content Delivery and Virtualization

Content delivery networks (CDNs), video transcoding, live streaming, VDI

Opportunities and Challenges

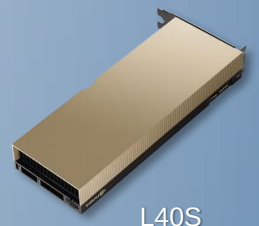
- Contents in 4K and 8K, 120Hz+ refresh rate for cloud gaming
- Save data bandwidth and reduce delivery delays
- Faster, more efficient transcoding and compression
- Reduce power consumption and infrastructure cost
- Balancing hot, warm, cold data storage for data throughput and capacity



Content Delivery and Virtualization

Key Technologies

- GPU media engines with transcoding acceleration including AV1 encoding and decoding
- NVIDIA RTX GPUs handling both real-time 3D graphic rendering and media streaming for cloud gaming and VDI.
- NVIDIA BlueField-2, -3 (DPU) for low latency, secure and fast data management
- Dense, resource-saving multi-node, multi-GPU systems for space and power efficiency
- High-capacity, high-throughput hot-swap storage



L40S

L40

RTX 6000 Ada

GPU Optimized Systems by Workloads

Content Delivery and Virtualization

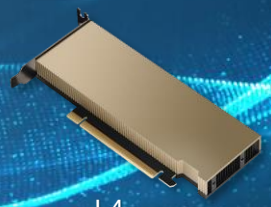
2U 4-Node BigTwin
(Up to 2 SW GPUs per node)
SYS-221BT-HNTR, SYS-621BT-HNTR

2U CloudDC
(Up to 2 DW or 4 SW GPUs)
SYS-521C-NR, AS -2015CS-TNR

2U Hyper-E Short-Depth
(Up to 3 DW GPUs or 4 SW GPUs)
SYS-221HE-FTNR, SYS-221HE-FTNRD

L40

RTX 6000 Ada

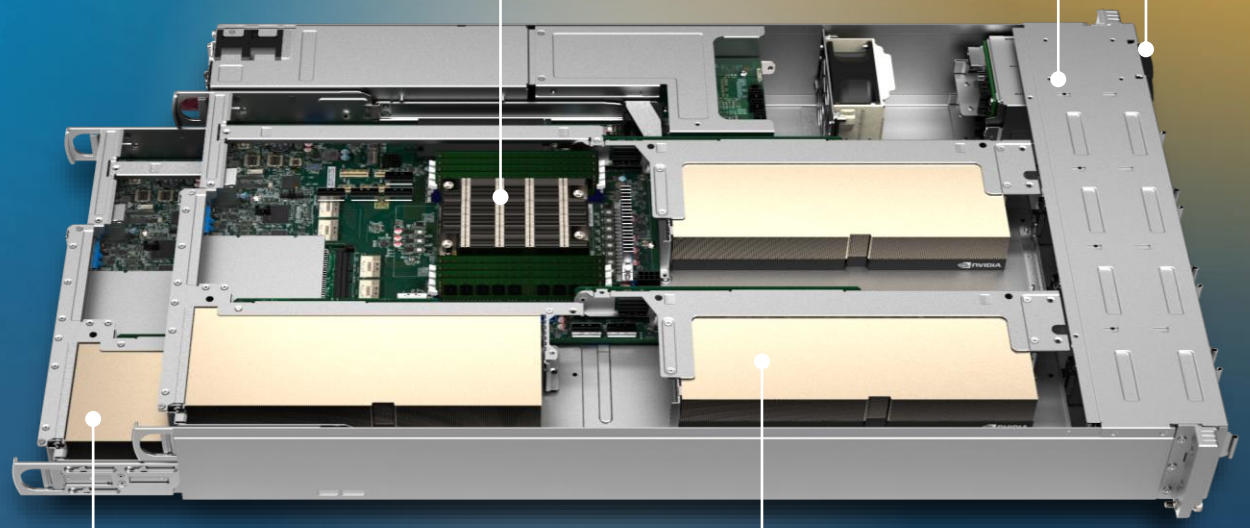


L4

2U 2-Node GPU System

SYS-210GP-DNR / AS -2114GT-DNR

- Up to 3 DW GPUs per node, both passive and active cooling GPUs such as A100, A40, RTX A6000, A4000
- Single 3rd Gen Intel® Xeon® Scalable or AMD EPYC™ 7003 Series processor per node
- Networking via PCIe 4.0 x8 AIOM slot per node
- 2 hot-swap U.2 NVMe drives per node
- NAB Show 2022 Product of the Year award winner



Front Hot-Swap NVMe Drives
Up to 2 U.2 drives per Node

Single Socket per Node
3rd Gen Intel Xeon Scalable or AMD 7003 Series Processors

Shared Power and Cooling
Increased Efficiency

Hot-pluggable Nodes
in a 2U form factor

Up to 3 DW or 6 SW GPUs
and 1 AIOM per node

2U 4-Node BigTwin®

SYS-221BT-HNTR/SYS-621BT-HNTR

- Up to 4 SW GPUs such as NVIDIA L4 and 8 CPUs in 2U
- Dual 4th Gen Intel® Xeon® Scalable processors
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- 2 PCIe 5.0 x16 (LP) slots
- 6 NVMe drives per node
- Networking via AIOM (OCP 3.0 compatible) per node

Use Cases

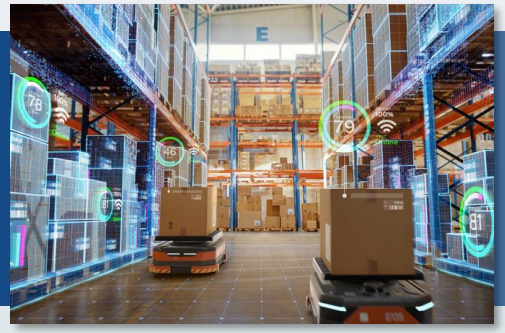
- Media & Entertainment Streaming
- Oscar-Worthy Visual Effects
- AI-Accelerated Production
- Virtual Human Meets Virtual Studio

AI Edge

Intelligent retail, Industry 4.0, smart cities, predictive healthcare, smart security and more

Opportunities and Challenges

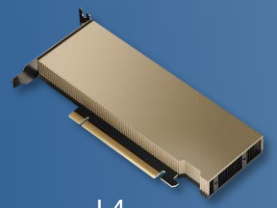
- Space and weight limitation, power constraints
- Balancing data throughput for video and audio requirements with cost of storage and bandwidth constraints
- Latency impacting response time and service quality
- Data privacy and security, regulatory compliance
- Resiliency in face of network outages
- Long product lifecycle requirements



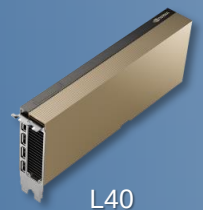
AI Edge

Key Technologies

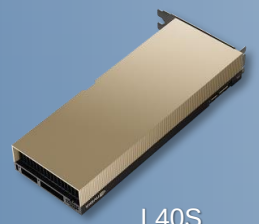
- CPU or GPU-based AI Inferencing, GPU-based video transcoding/encoding/decoding
- Short-depth chassis design for edge locations with AC or DC power supply options
- Front I/O with broad range of expansion and I/O port for flexibility and easy serviceability
- Ruggedized systems designed to be placed outside of the data center
- Edge fleet management software



L4



L40



L40S

2U Hyper-E

SYS-221HE-FTNR, SYS-221HE-FTNRD

- Up to 3 DW GPUs or 4 SW GPUs
- Dual 4th Gen Intel® Xeon® Scalable processors
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- Flexible network options with 2 AIOM slots up to 200GbE each
- AC or DC power option

GPU Optimized Systems by Workloads

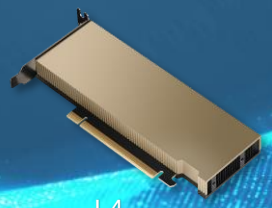
AI Edge

2U Hyper-E Short-Depth
(Up to 3 DW GPUs or 4 SW GPUs)
SYS-221HE-FTNR, SYS-221HE-FTNRD

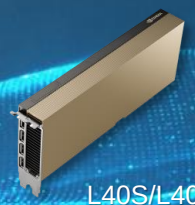
Compact Fanless Edge Server
(Up to 3 SW GPUs)
SYS-E403-13E

1U Compact Short-Depth Edge/5G Server
(Up to 2 SW GPUs)
SYS-111E-FWTR

Embedded Fanless Edge Server
(CPU or ASIC based Inference)
SYS-E100-13AD



L4



L40S/L40

Use Cases

Verticals: Retail (groceries)
Workloads: Computer vision
Number of Nodes/Systems: 1000+

Where Should You Go From Here

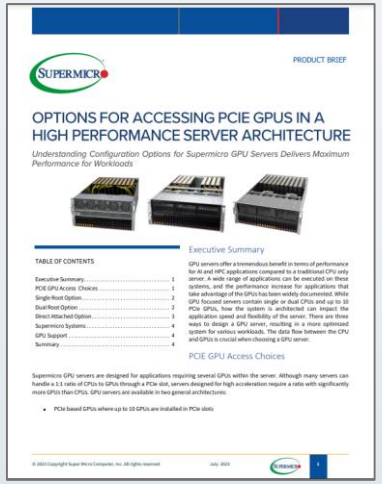
- Download sales assets, get familiar with GPU accelerated workloads
- Feel free to use these slides to engage with your customers
- Get PM's help if more in-depth technical information, benchmarks/proof points needed
- Give us feedback
- Happy selling!

Leverage Sales Assets

New Landing Page, AI/GPU Workload Brochure, Product Brief, Datasheets, and etc.

AI Solution Page
www.supermicro.com/ai

AI GPU Brochure



Product Brief

AI Workload Datasheets