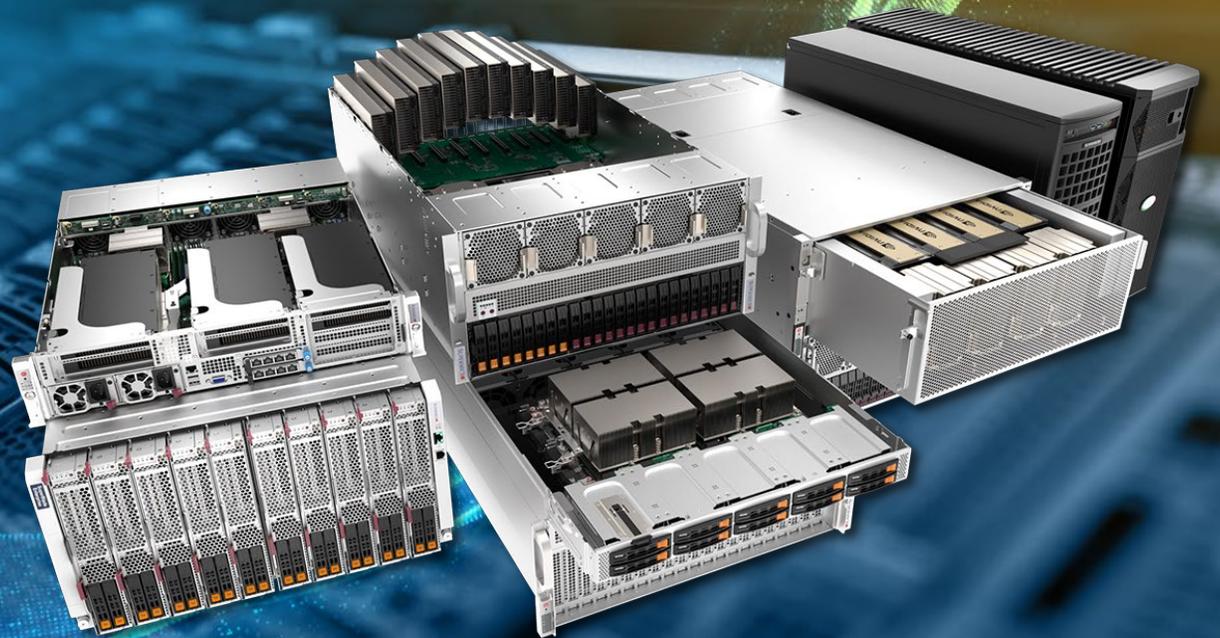




From Large Language Model to the AI Edge

GPU Acceleration for
Broad Range of Workloads



Accelerate Everything AI Campaign

SUPERMICRO Accelerate Everything

Large Language Models (LLM) to the AI Edge

September 2023

Large Scale AI Training Workloads

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 L1/L12 testing and validation
- Factory tuned power and cooling design
- Single source liquid cooling solution available with reduced (weeks) lead time

Large Scale AI Training Workloads

HGX H100 Systems

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

Medium 4U 4-GPU

(Codename: RedStone-Next)
 NVIDIA HGX H100 SXM 4-GPU
 8 U.2 NVMe Drives
 8 PCIe 5.0 x16 networking slots
 95-4160-7000

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

(Codename: Delta-Next)
 NVIDIA HGX H100 SXM 8-GPU
 16 U.2 NVMe Drives
 8 PCIe 5.0 x16 networking slots
 95-4216G-7000 / 95-4216S-7000

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

Accelerate Everything AI
 Maximize GPU Acceleration, Maximize AI Model Inference

Accelerate Every Workload

Large Scale AI Training, HPC/AI, Enterprise AI Inference & Training, Visual Analytics & Bridge, Cloud AI Software and Virtualization, AI Edge

Large Scale AI Training
 4U 4-GPU, 8U 8-GPU

HPC/AI
 4U 4-GPU, 8U 8-GPU

Enterprise AI Inference & Training
 4U 4-GPU, 8U 8-GPU

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

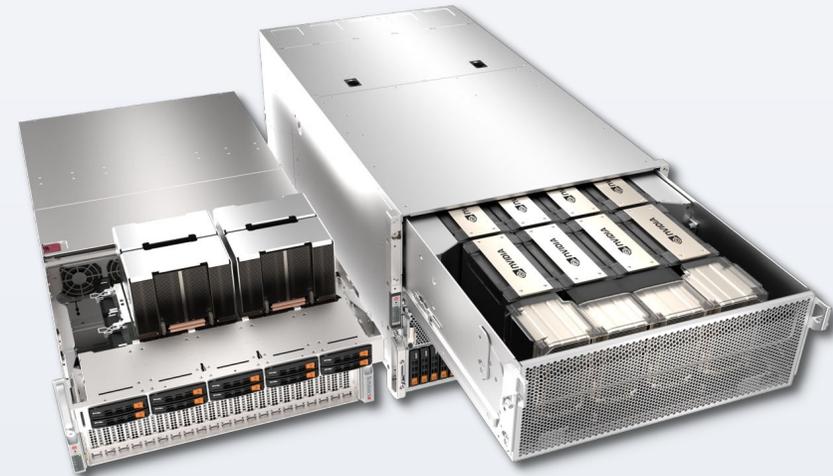
Genomic sequencing
and analysis speed-up



Graphics, Rendering, Video

~100X

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





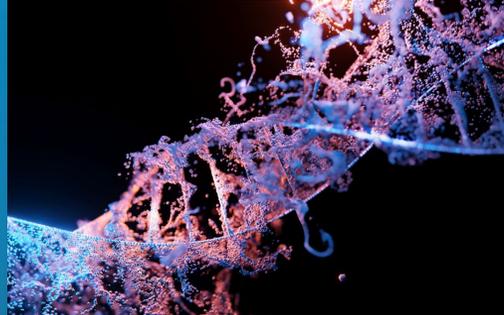
Accelerate Everything

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

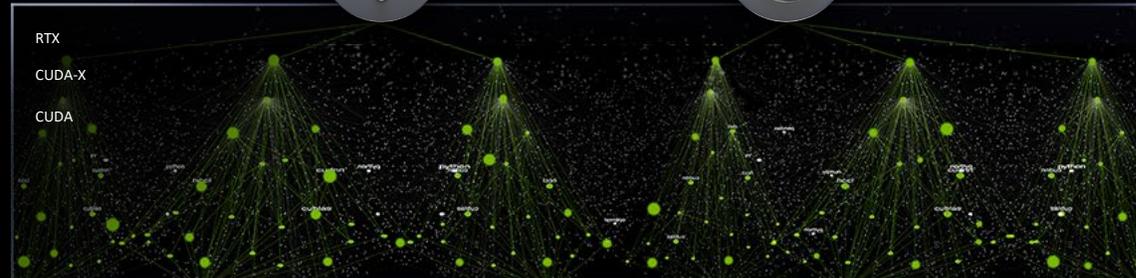
APPLICATION
FRAMEWORKS



PLATFORM



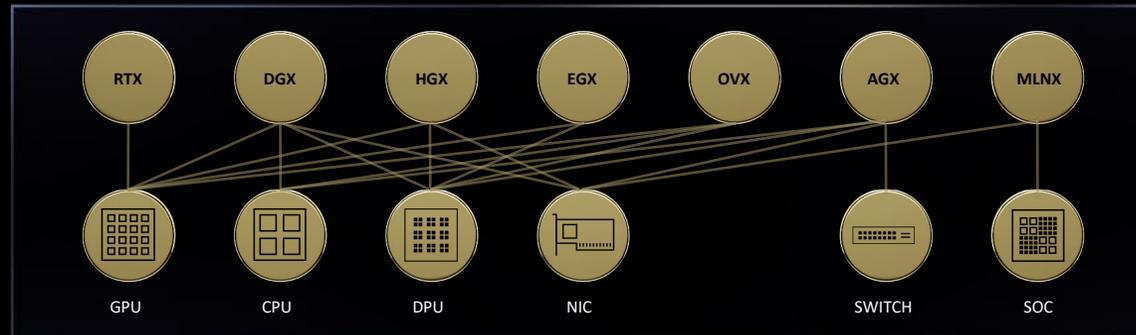
ACCELERATION
LIBRARIES



SYSTEM
SOFTWARE



HARDWARE



Accelerated Computing is Sustainable Computing

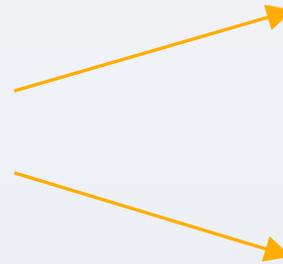


\$12M | 960 CPU Servers | 11 GWh | 1X LLM



\$12M | 48 GPU Servers | 4.0 GWh | 44X LLM

ISO-BUDGET



\$600K | 2 GPU Servers | 0.13 GWh | 1X LLM

ISO-PERFORMANCE

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 Acceleration
Compute	HGX H100 	80GB HBM3 per GPU	SXM ★★★★	SXM ★★★★	SXM ★★★★				
	H100 NVL 	94GB HBM3 per GPU	NVL ★★★★	NVL ★★★★	NVL ★★★★				
	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		★★		★★	★★★★	★★★★	★★★★
	A2 	16GB GDDR6 60W		★			★	★	★

*More than 40 weeks wait time....

*NVL means 2x H100 PCIe paired with NVLink Bridge by default. Regular H100 PCIe and A100 PCIe cards can be paired with NVLink Bridge as option.

NVIDIA L40S

The highest performance universal GPU
for AI, graphics, and video

Fine Tuning LLM

4hrs

GPT-175B 860M Tokens¹

AI Training

1.7X

Performance vs. HGX A100²

AI Inference

1.5X

Performance vs. HGX A100³

GPT3 Training

<4 days

GPT-175 300B Tokens⁴

Image Gen AI

>82

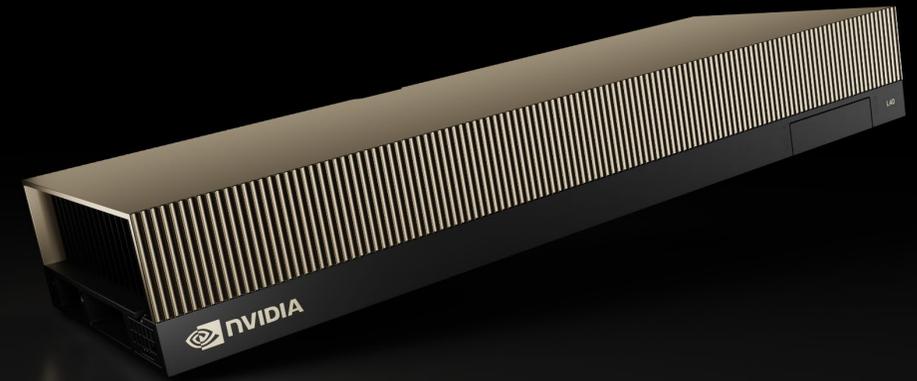
Images per minute⁵

Full Video Pipeline

184

AV1 Encode Streams⁶

Dual-Slot | FHFL | 350W



Preliminary performance projections, subject to change

1. Fine-Tuning LoRA (GPT-175B), bs: 128, sl: 256; 64 GPUs: 16 systems with 4xL40S
2. Fine-Tuning LoRA (GPT-40B), bs: 128, sl: 256; Two systems with 4x L40S, vs HGX A100 8 GPU
3. Hugging Face SWIN Base Inference (BS=1, Seq 224); L40S vs. A100 80GB SXM
4. GPT 175B, 300B tokens, Foundational Training; 4K GPUs; 1000 systems with 4xL40S
5. Image Generation, Stable Diffusion v2.1, 512 x 512 resolution; 1xL40S
6. Concurrent Encoding Streams; 720p30; 1xL40S



GPU Optimized Systems by Workloads

Large Scale AI Training



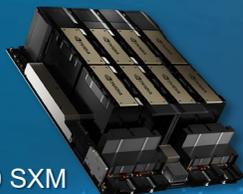
8U 8-GPU System (HGX H100 SXM)
(codenamed: Delta-Next)
SYS-821GE-TNHR, AS -8125GS-TNHR



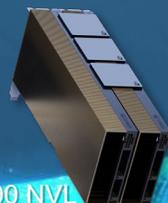
4U 4-GPU System (HGX H100 SXM)
(codenamed: Redstone-Next)
SYS-421GU-TNXR, SYS-521GU-TNXR



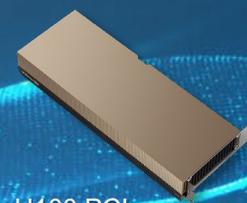
Petabyte Scale All-Flash Storage
SSG-121E-NE316R, ASG-1115S-NE316R



HGX H100 SXM
8-GPU or 4-GPU



H100 NVL



H100 PCIe



Grace Hopper Superchip
(Grace CPU + H100 GPU)

HPC/AI Workloads



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



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



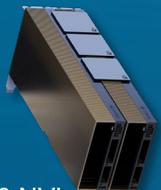
4U/5U 8-10 GPU System
(NVIDIA OVX™ reference design available)
SYS-521GE-TNRT, SYS-421GE-TNRT, AS -4125GS-TNRT



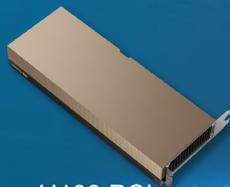
2U Hyper (Up to 4 GPUs)
SYS-221H-TNR, AS -2015HS-TNR



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



H100 NVL



H100 PCIe



L40S



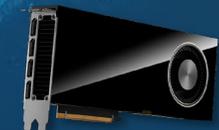
L40



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



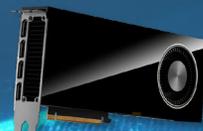
2U CloudDC
(Up to 2 DW or 4 SW GPUs)
SYS-521C-NR, AS -2015CS-TNR



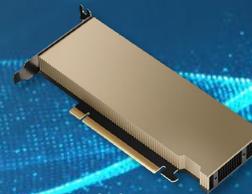
2U Hyper-E Short-Depth
(Up to 3 DW GPUs or 4 SW GPUs)
SYS-221HE-FTNR, SYS-221HE-FTNRD



L40



RTX 6000 Ada



L4

AI Edge



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



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



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



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



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

1:1 Networking Slots
for GPUs up to 400Gbps

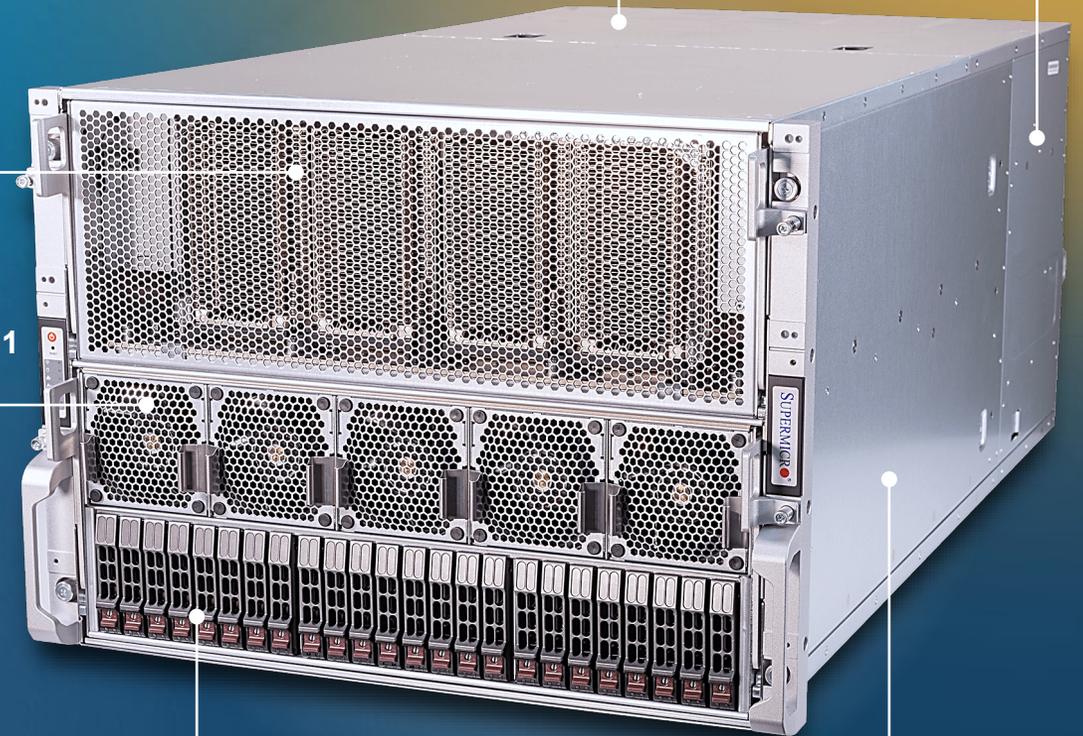
Optimized Thermal
and liquid cooling option

NVIDIA HGX H100
SMX5 8-GPU

PCIe 5.0, DDR5, CXL1.1
latest tech stack

2.5" Drive Bays
Up to 16 NVMe drives

Dual 4th Gen[®] Intel[®]
Xeon Scalable or AMD EPYC
9004 Series Processors

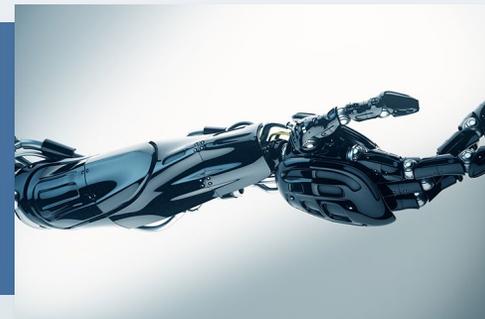
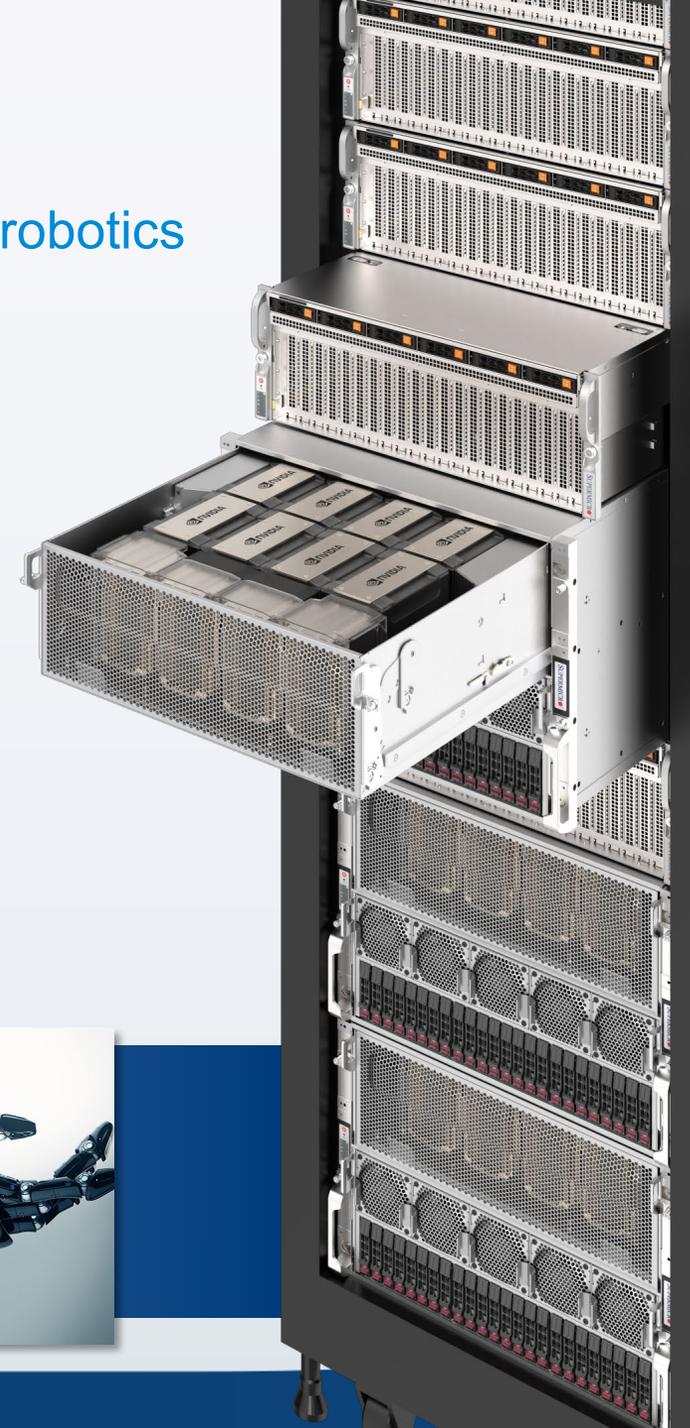


Large Scale AI Training

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



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



HGX H100 SXM
8-GPU or 4-GPU



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

Well-designed Networking is Crucial for AI Data Centers

AI Factories



- Single or few workloads
- Extremely large AI models

NVLINK and InfiniBand AI Fabric

AI Cloud



- Multi-tenant
- Variety of workloads

Ethernet network

NVIDIA Quantum-2 InfiniBand Platform

Unprecedented Performance, Scalability, and Security for Scientific Computing

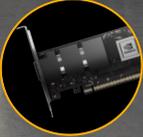
Bare-Metal Secured Multi-Tenant Infrastructure
 Performance Isolation with Congestion Control
 Advanced Adaptive Routing
 In-Network Computing
 400Gb/s InfiniBand



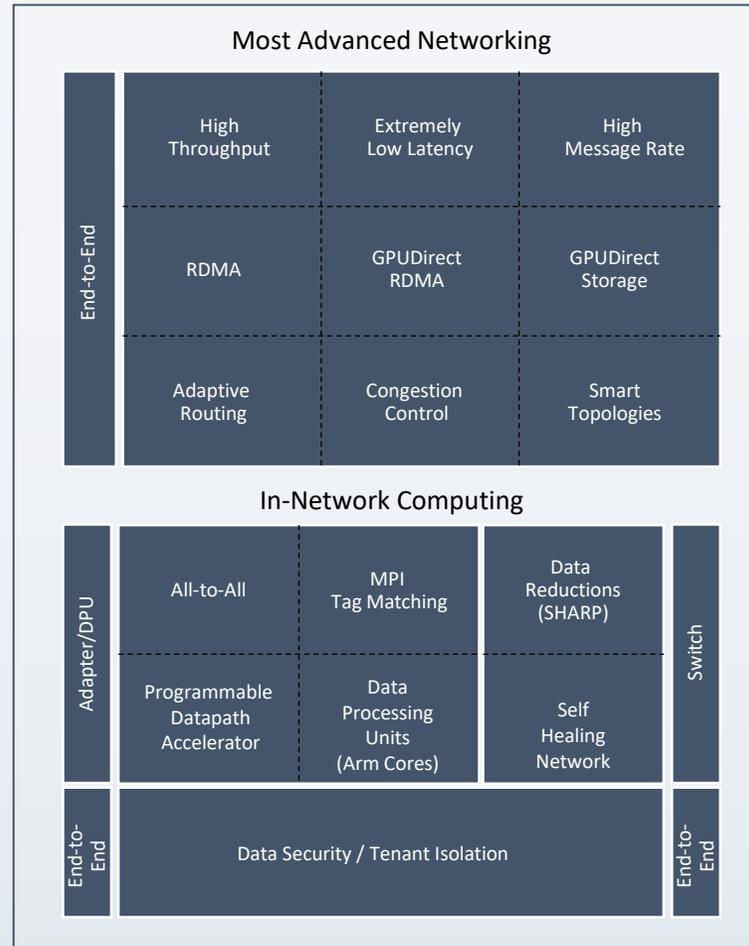
NVIDIA Quantum-2
Switch



BlueField-3
DPU



ConnectX-7



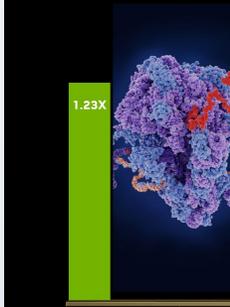




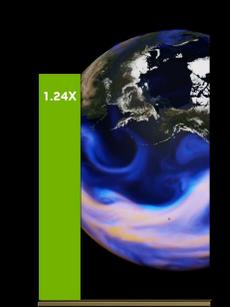




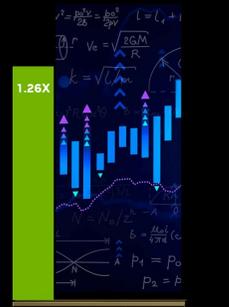
1.2x Higher Application Performance with BlueField DPU and Quantum InfiniBand In-Network Computing



1.23X
Physics/Chemistry



1.24X
Weather



1.26X
FFT

NVIDIA Spectrum Ethernet Platform

Unprecedented Performance, Scalability, and Security for Every Data Center Network

Accelerated Ethernet Fabric for AI and Cloud Data Centers



Spectrum Ethernet Switches

Ethernet Network Software Suite

NetQ

Cumulus

Air

Monitoring

Linux NOS

Simulation



BlueField DPU



ConnectX NIC



LinkX Cables

End-to-End Ethernet Platform Tuned for Accelerated Workloads



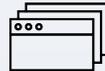
Highest Bandwidth



Lowest Latency



Fastest AI Run Times



Data Center Simulation



Network Visibility



Energy Efficient



NVIDIA Integration



End-to-End Security



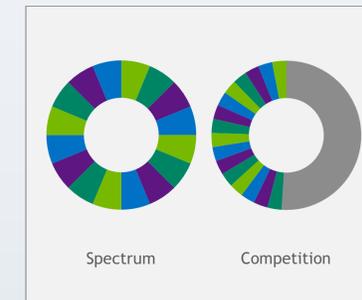
Software Driven Automation

Highest Performing Ethernet Fabric for AI and Cloud



Lowest Latency
Packet Latency (ns)

Spectrum provides **10X** lower latency



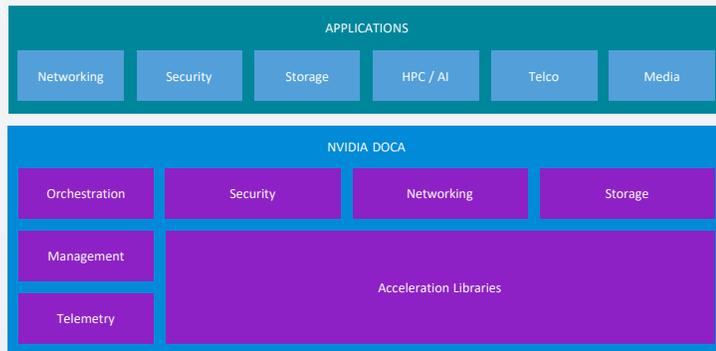
Fair Bandwidth Sharing
Bandwidth Sharing Across Ports

Shared buffer delivers fairness & **100%** reduction in buffer usage

NVIDIA BlueField DPU Platform

Software-defined, hardware-accelerated infrastructure compute platform

Offload, Accelerate, Isolate



BlueField-2
200Gb/s, 9 SPECint
8x Arm A72 cores
Single DDR4 channel



BlueField-3
400Gb/s, 42 SPECint
16x Arm A78 cores
Dual DDR5 channels
Datapath Accelerator



Accelerated Performance
Meet the most stringent performance requirements, run the most demanding workloads



Cloud-Scale Efficiency
Free up x86 cores, save power, and achieve unprecedented scale and efficiency levels



Robust Zero-Trust Security
Ensure comprehensive data center security without compromising on performance



Fully Programmable Infrastructure
Develop and run applications consistently with maximum performance

Broad and Expansive Ecosystem

Applications



Cloud Platforms

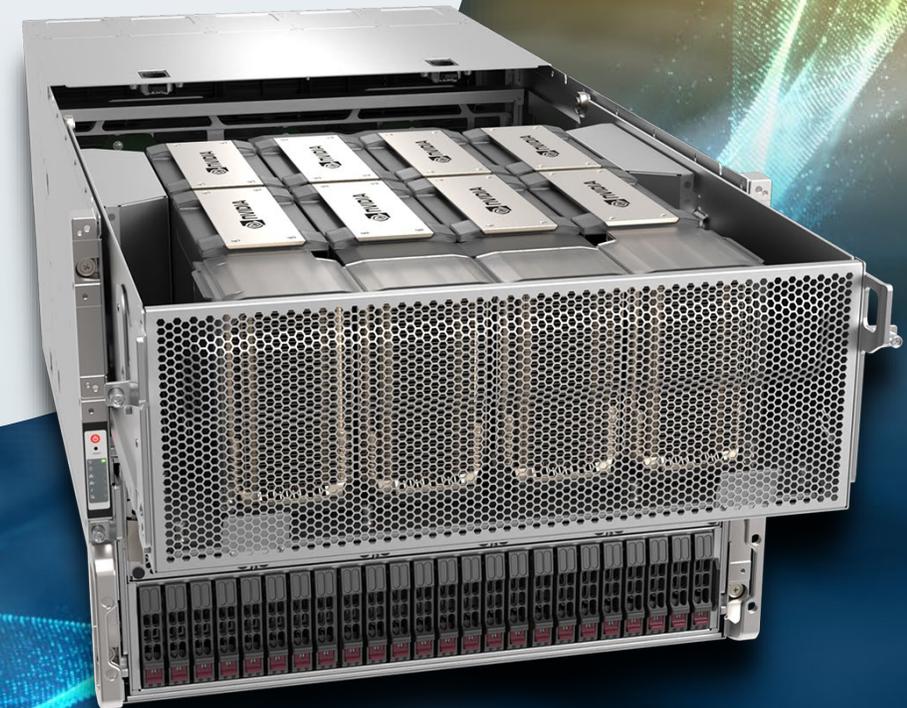


OEM Systems



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, billions of user profiling
- Telco – chatbot for customer support
- Financial Services - retraining GPT-3 level model with 50B parameters for inquiry services



HPC/AI Workloads

Engineering simulation, scientific research, genomic sequencing, drug discovery

Opportunities and Challenges

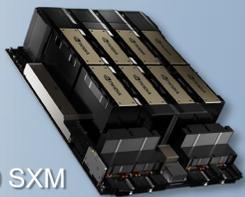
- Augmenting machine learning algorithms and GPU accelerated parallel computing to HPC workloads to achieve faster results and discoveries
- Parallel processing with massive datasets for data-intensive simulations and analytics
- Simulations requiring double precision (FP64)
- High-resolution and real-time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



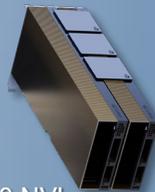
HPC/AI Workloads

Key Technologies

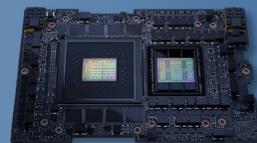
- Double-precision Tensor Cores delivering 535/268 teraFLOPs with HGX H100 SXM 8-GPU/4-GPU, or 134 teraFLOPs with H100 NVL (2 GPUs with NVLink Bridge) at FP64
- High CPU compute and high GPU compute – e.g, up to 20 CPUs and 20 GPUs in 8U
- High bandwidth GPU memory and CPU cache/integrated memory – HBM3,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



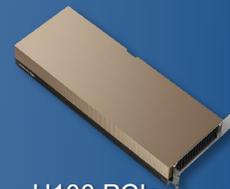
HGX H100 SXM
8-GPU or 4-GPU



H100 NVL



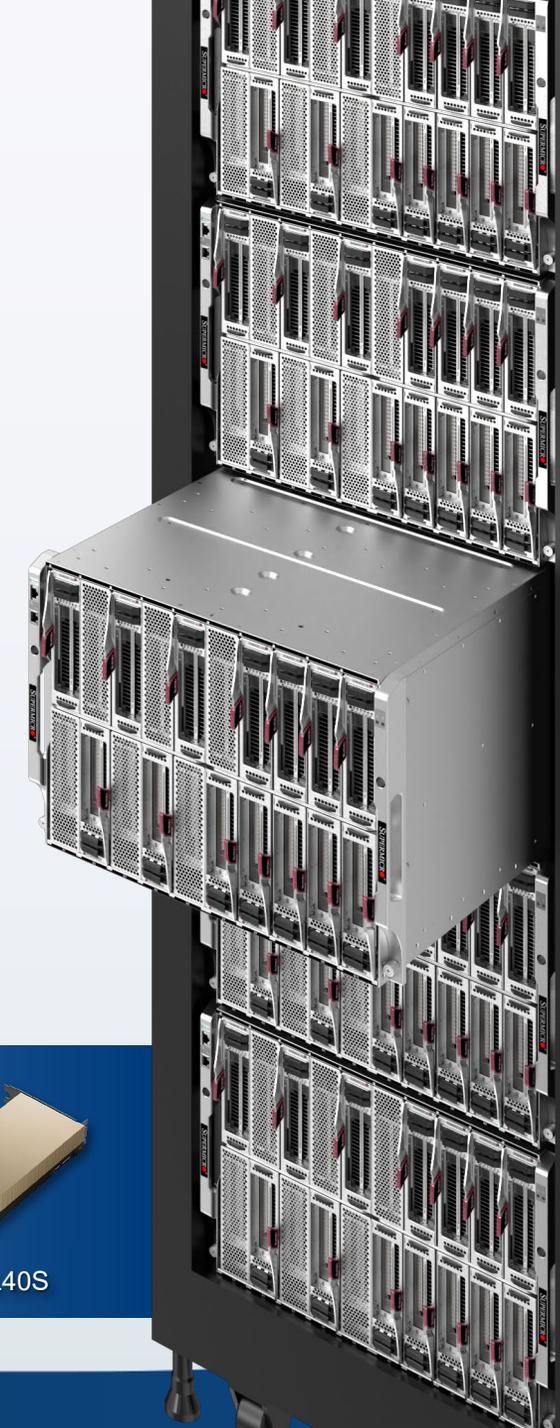
Grace Hopper Superchip
(Grace CPU + H100 GPU)



H100 PCIe



L40S





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



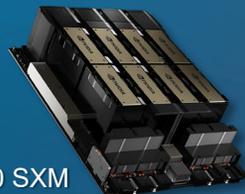
4U/5U 8-10 GPU System
SYS-521GE-TNRT, SYS-421GE-TNRT/TNRT3
AS -4125GS-TNRT/TNRT1/TNRT2



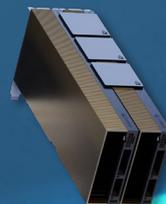
8U SuperBlade (Up to 20 nodes)
SBI-411E-1G / SBI-411E-5G



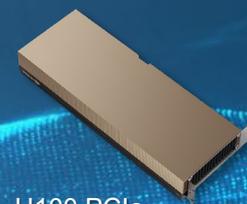
1U Grace Hopper MCX System
SYS-421GU-TNXR / SYS-521GU-TNXR



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

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

Dual Root Complex



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

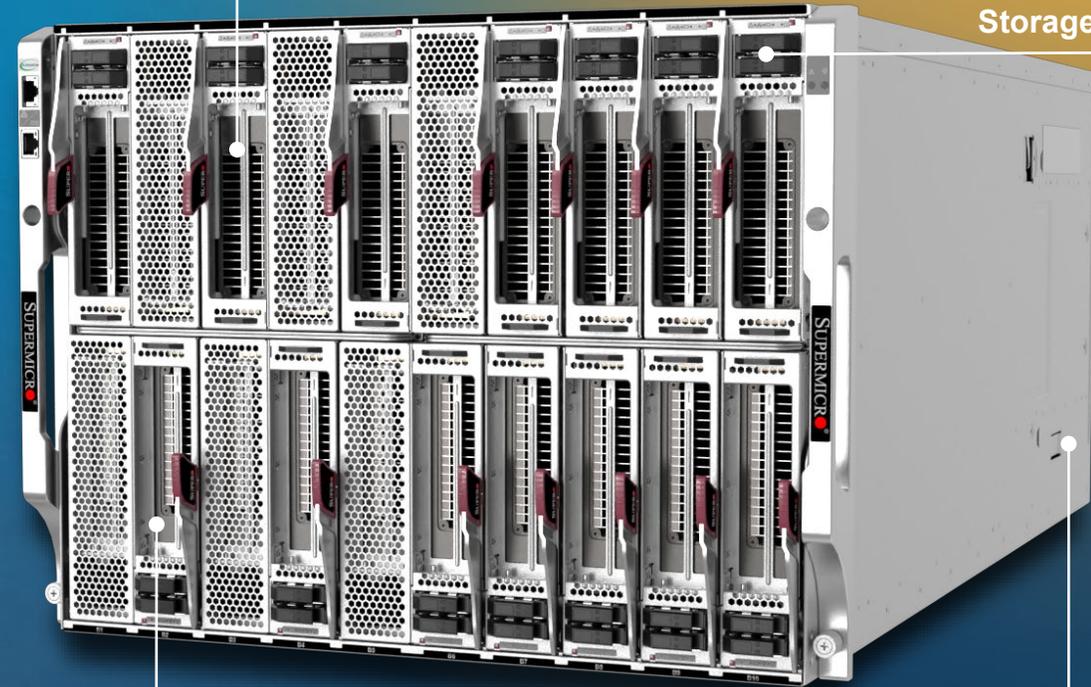
Optional 5U Chassis
Enhanced Thermal
Capacity and I/O



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



Up to 20 NVIDIA H100 PCIe GPUs in 8U

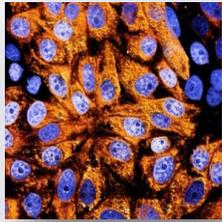
2x E1.S PCIe 5.0 Storage per blade

1: 1 GPU to CPU per blade
or 2 : 1 GPU to CPU

Shared power and cooling,
integrated switch and
management console

Workloads of the Modern Supercomputer

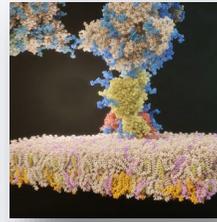
EDGE



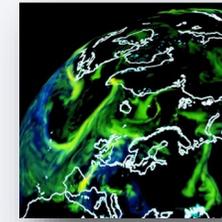
SIM + AI



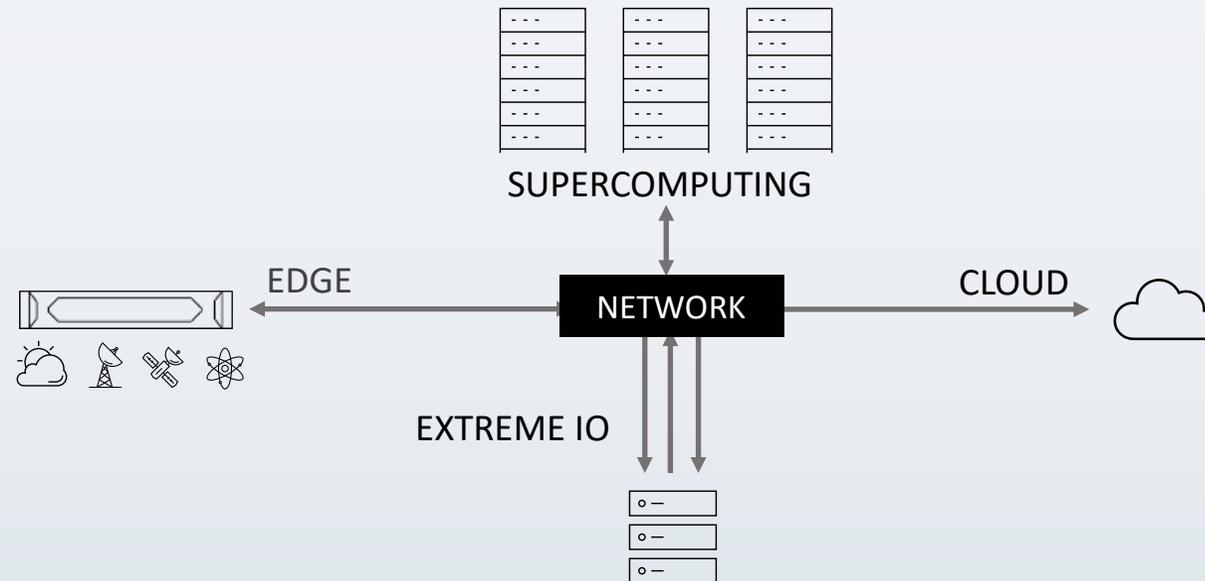
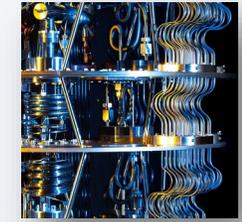
SIMULATION



DIGITAL TWIN



QUANTUM COMPUTING

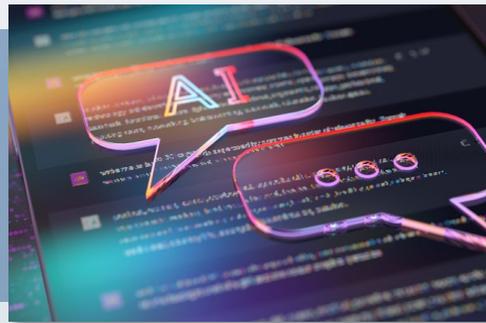


Enterprise AI Inference and Training

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, or hybrid
- Utilization of frameworks, pre-trained models, open-source AI models with fine-tuning and embeddings (with their own dataset)

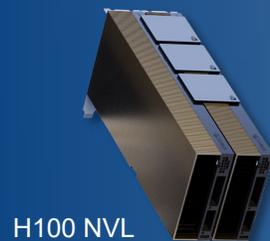


Enterprise AI Inference and Training

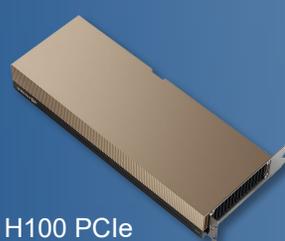
AI-enabled services/applications, chatbots, business automation

Key Technologies

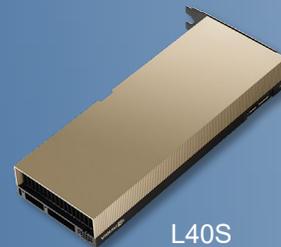
- Flexible, modular, highly configurable rackmount servers with different form factors to balance compute, storage, networking, and cost for various enterprise AI workload needs for today and the future
- PCIe 5.0 supported platforms for future proofing – GPUs, storage, networking
- FP8 and FP16 support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



H100 NVL



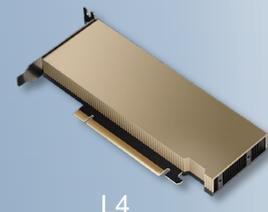
H100 PCIe



L40S



L40



L4



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

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

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



E1.S Drive Bays
Up to 8 Drives

Final system configuration subject to change

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



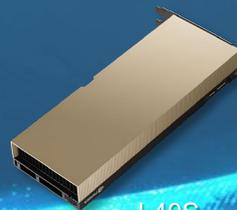
4U/5U 8-10 GPU System
(NVIDIA OVX™ reference design available)
SYS-521GE-TNRT, SYS-421GE-TNRT, AS -4125GS-TNRT



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

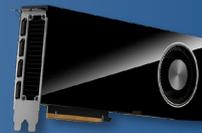
- GPU media engines with transcoding acceleration including AV1 encoding and decoding
- NVIDIA RTX GPUs handling both real-time 3D graphic rendering and media streaming for cloud gaming and VDI.
- NVIDIA BlueField-2, -3 (DPU) for low latency, secure and fast data management
- Dense, resource-saving multi-node,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



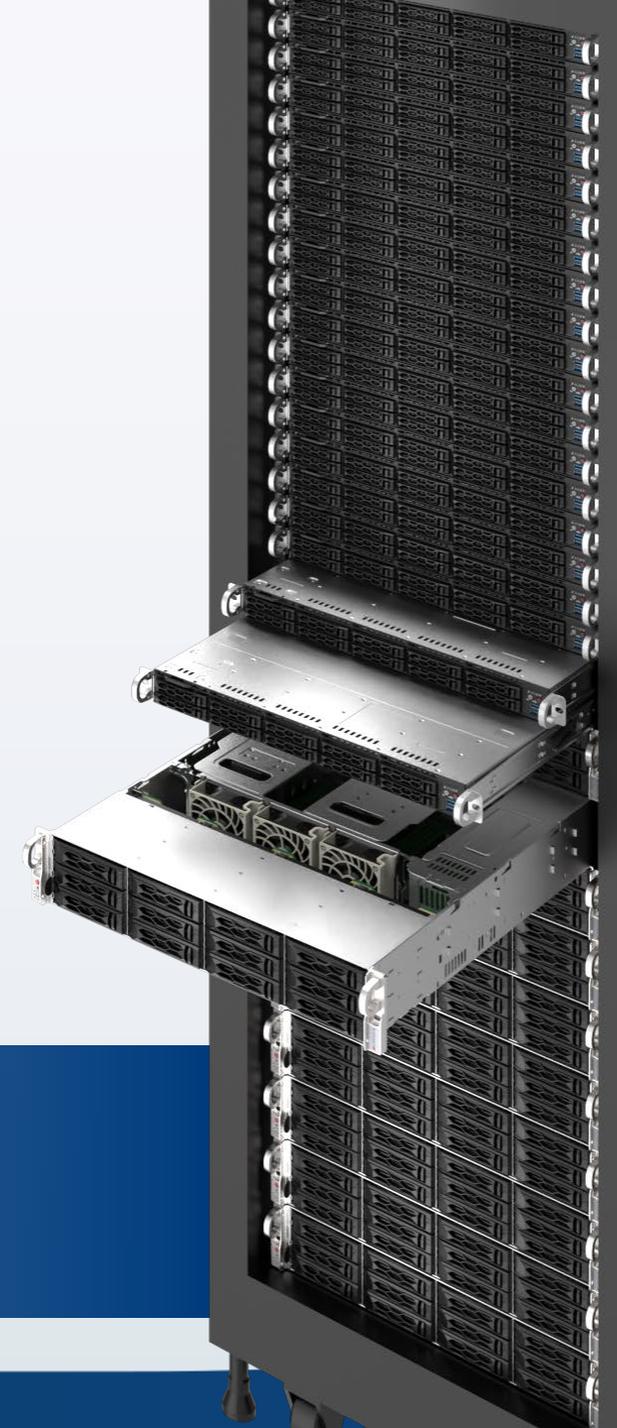
L40S



L40



RTX 6000 Ada



AI Edge

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

Opportunities and Challenges

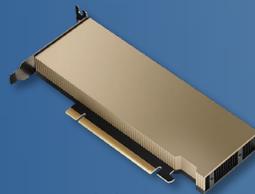
- Space and weight limitation, power constraints
- Balancing data throughput for video and audio requirements with cost of storage and bandwidth constraints
- Latency impacting response time and service quality
- Data privacy and security, regulatory compliance
- Resiliency in face of network outages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

Short-depth chassis
for edge locations

Flexible Configurations
GPU, NICs, storage

PCIe 5.0
Up to 4 x16 or 8 x8 slots

AC or DC Power Supplies
Ideal for Edge Deployments

Front I/O
Cold-aisle Serviceability





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



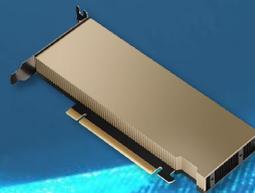
Compact Edge Server
(Up to 3 SW GPUs)
SYS-E403-13E



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



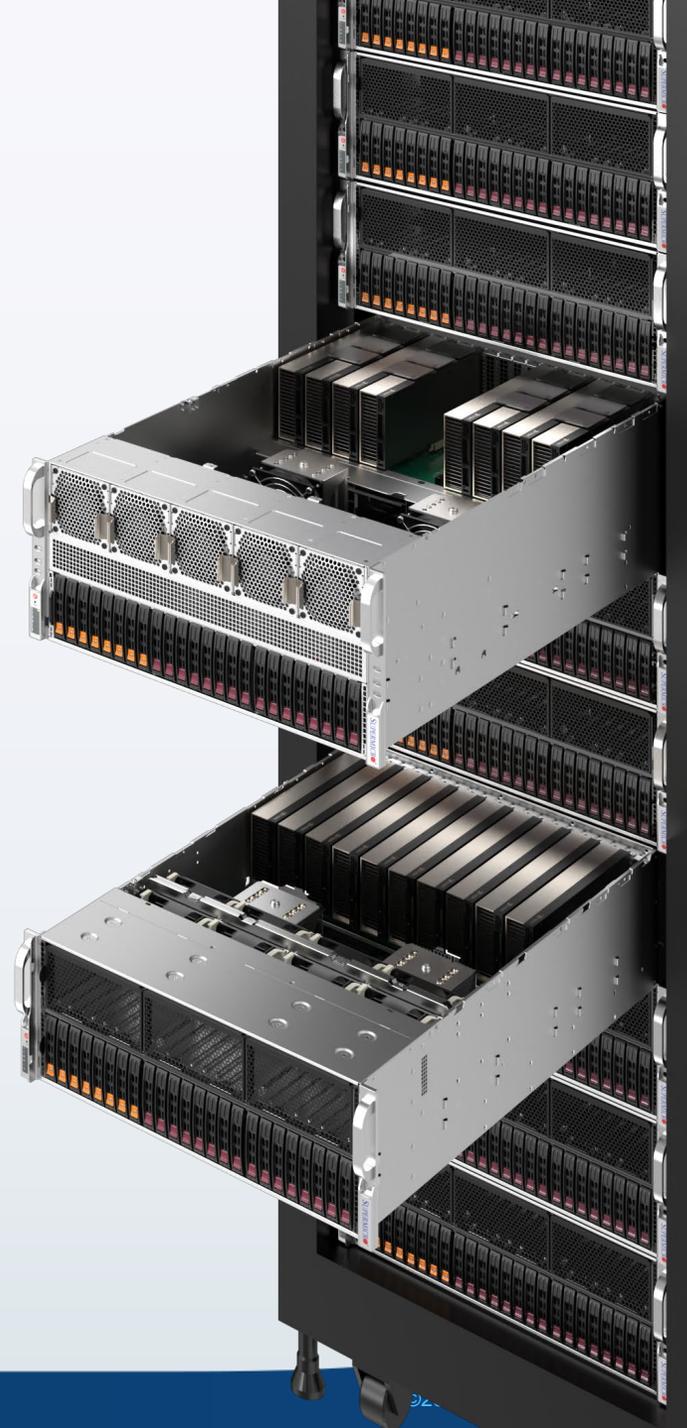
L4



L40S/L40

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
- Give us feedback
- Happy selling!



Leverage Sales Assets

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

The screenshot displays the AI Solution Page with a blue and white color scheme. The main heading is "Accelerate Everything AI" with a sub-heading "Workload GPU Acceleration, Large Scale AI Training, AI Inference". Below this is a navigation menu with tabs for "Home", "Solutions", "Products", "Support", and "Partners". The main content area is divided into four sections: "Accelerate Every Workload", "Large Scale AI Training", "HPC/AI", and "Enterprise AI Inference & Training". Each section features a brief description and a "Learn More" button.

AI Solution Page
www.supermicro.com/ai

The brochure cover features the Supermicro logo and the slogan "Accelerate Everything". The main image shows a server rack with a GPU card being inserted. The text at the bottom reads "Large Language Models(LLM) to the AI Edge".

AI GPU Brochure

The product brief cover has a blue and white design. The title is "OPTIONS FOR ACCESSING PCIe GPUS IN A HIGH PERFORMANCE SERVER ARCHITECTURE". Below the title is a sub-heading "Understanding Configuration Options for Supermicro GPU Servers Delivers Maximum Performance for Workloads". The cover also features the Supermicro logo and a "PRODUCT BRIEF" label.

Product Brief

The datasheet for Large Scale AI Training lists various system configurations under "AI High Systems" and "FlexScale Scale Storage". It includes details on components like CPUs, GPUs, and storage. A section titled "Recommended NVIDIA GPUs" lists several GPU models with their specifications.

The datasheet for HPC/AI lists configurations for "AI High Systems" and "AI High Systems". It details system components and provides a list of "Recommended NVIDIA GPUs" with their respective specifications.

The datasheet for Enterprise AI Inference & Training lists configurations for "AI High Systems" and "AI High Systems". It includes system specifications and a list of "Recommended NVIDIA GPUs" with their details.

The datasheet for Visualization & Design lists configurations for "AI High Systems" and "AI High Systems". It details system components and provides a list of "Recommended NVIDIA GPUs" with their specifications.

The datasheet for Content Delivery & Virtualization lists configurations for "AI High Systems" and "AI High Systems". It includes system specifications and a list of "Recommended NVIDIA GPUs" with their details.

The datasheet for AI Edge lists configurations for "AI High Systems" and "AI High Systems". It details system components and provides a list of "Recommended NVIDIA GPUs" with their specifications.

AI Workload Datasheets