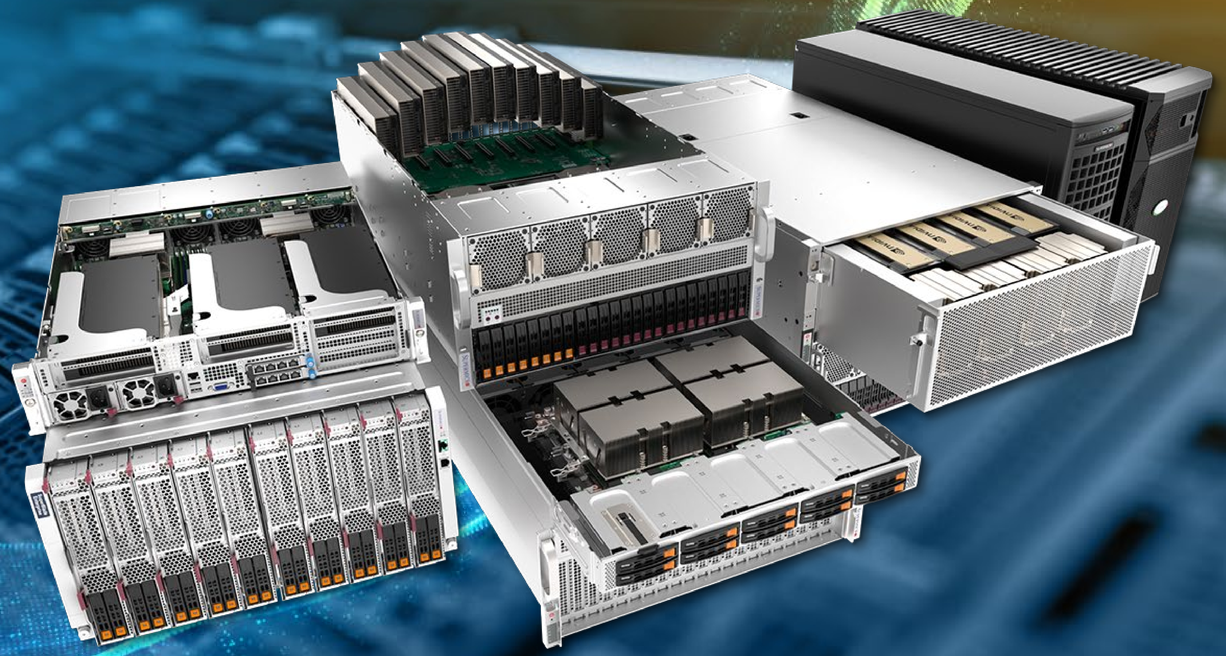
L40S GPU Platform & Systems

GPU Acceleration for Broad Range of Workloads

Bernhard Schimpl

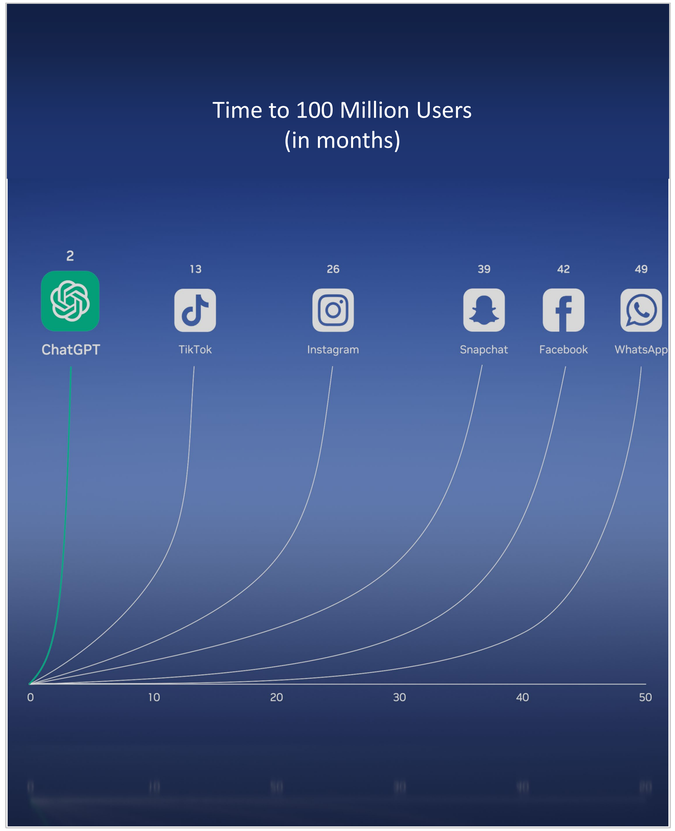
Jeff Kang



The iPhone Moment of AI is Here

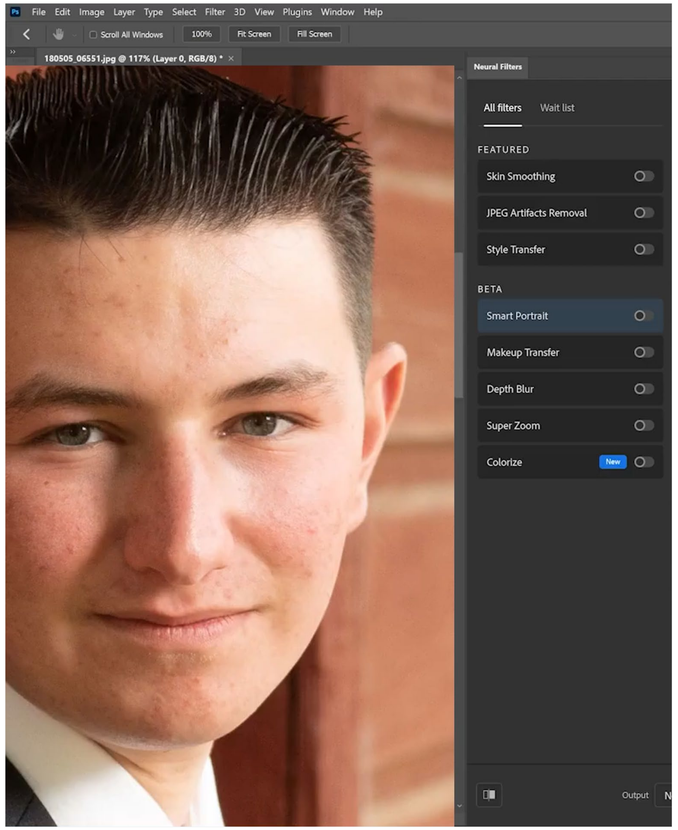
Every major application and workflow is going to include AI

CHATBOTS
Fastest Growing Application Ever



GENERATIVE ART
Over 200M+ Users

AI-AUGMENTED APPLICATIONS
ISVs Accelerating AI Integration



3 Major Changes

The modern AI Data Center

1 Rapid growth in AI adoption

- HUGE growth in generative AI in particular
- LLM (Large Language Models)
- AI integration into HPC, Omniverse, etc.

MM write me a poem about NVIDIA GPUs

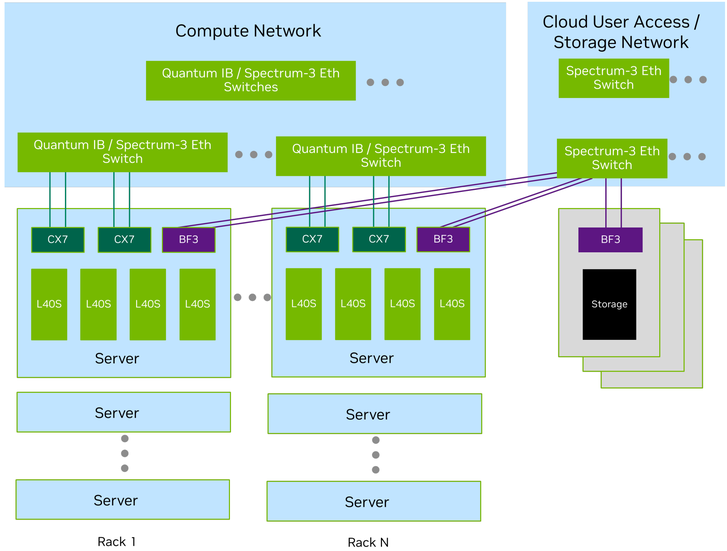
NVIDIA GPUs, a digital delight,
In silicon realms, they shine so bright.
With CUDA cores and ray tracing's grace,
They conquer gaming and AI's embrace.

From pixels to patterns, they unveil the lore,
In data centers, they explore and soar.
With Pascal, Turing, and Ampere's might,
NVIDIA GPUs lead us to the light.

In this world of tech, they take the lead,
A digital masterpiece, indeed.
NVIDIA, you're the star we see,
In the vast cosmos of technology.

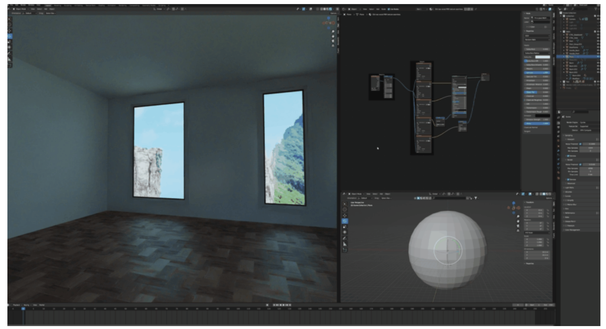
2 These are data center level problems

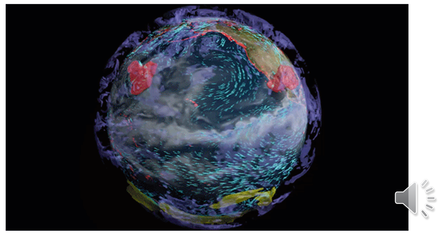
- Optimized node-to-rack-to-room
- Optimized network for compute & storage
- Optimized software stack



3 Datacenters must handle multiple workloads

- AI, both training and inferencing
- Graphics, collaboration, rendering
- Virtualization & multi-tenancy

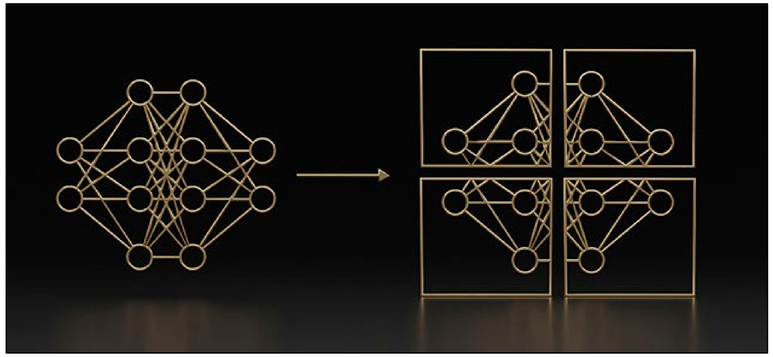




NVIDIA

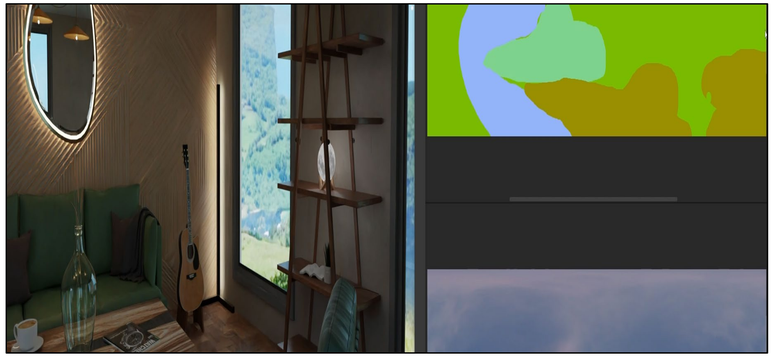
Powerful Multi-Workload Acceleration

Universal Performance to Accelerate a Broad Range of AI and Graphics Use Cases



LLM Inference & Training

Accelerate AI training and inference workloads with 4th Gen Tensor Cores, Transformer Engine and support for FP8.



Generative AI

Breakthrough inference performance for AI-enabled graphics, video, and image generation

3D Graphics and Rendering

Tackle high-fidelity creative workflows with 3rd-Gen RTX , DLSS 3 and 48GB of GPU memory

Mainstream Compute

Powerful FP32 for scientific data analysis and simulation. Life science, geo science, physics, higher-ed, and financial services.

Omniverse Enterprise

Connect, develop and operate Universal Scene Description (OpenUSD)-based 3D industrial digitalization workflows

Streaming and Video Content

Increase end to end video services hosted per GPU with higher encode/decode density and support for AV1

NVIDIA GPU's

NVIDIA L40S - Universal accelerator for a broad range of use cases

AI & COMPUTE WORKLOADS		GRAPHICS & GENERAL-PURPOSE WORKLOADS	
H100 Highest AI, LLM, HPC, & DA Performance	**A100** Powerful DL Training, Inference, AI & HPC	**L40** Powerful Visual Computing and AI	**L4** Universal AI, Video, and Graphics SFF, High-density, Low Power
Limited Availability, Longer Lead Times			
DL Training & DA	Language Processing	Graphics & Rendering	Mainstream Acceleration
DL Inference	Conversational AI	Omniverse	DL Inference
HPC	Recommenders	Virtual Desktops	Media Processing

NVIDIA L40S

The Most Powerful Universal Data Center GPU for AI and Graphics

L40S Value Proposition

Powerful AI & Graphics, Data Center Ready, Available Now!

Performance

Powerful AI + Graphics

Data Center Scale

Value

Better Price-Performance

Accelerate many workloads

Availability

Short Lead Time

Fast deployment

Price-performance L40S vs A100 **

NVIDIA L40S

The Highest Performance Universal GPU for AI, Graphics, and Video

Fine Tuning LLM

4hrs

GPT-175B 860M Tokens[1]

AI Training

1.7X

Performance vs. HGX A100[2]

AI Inference

1.5X

Performance vs. HGX A100[3]

GPT3 Training

<4 days

GPT-175 300B Tokens[4]

Image Gen AI

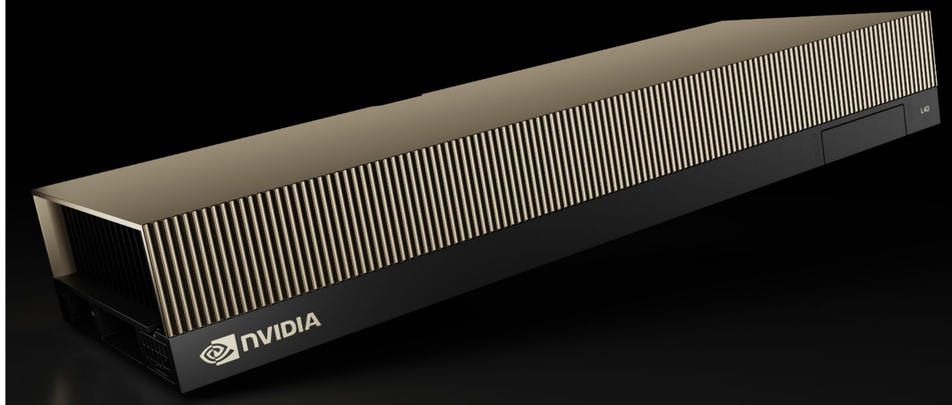
>82

Images per minute[5]

Full Video Pipeline

184

AV1 Encode Streams[6]



Dual-Slot I FHFL I 350W

Preliminary performance projections, subject to change

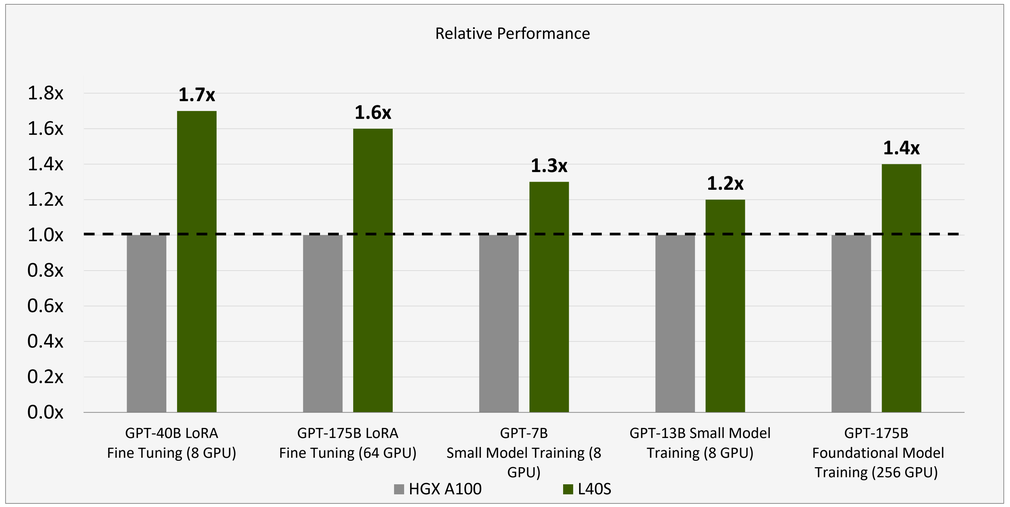
1. Fine-Tuning LoRA (GPT-175B), bs: 128, sl: 256; 64 GPUs: 16 systems with 4xL40S
2. Fine-Tuning LoRA (GPT-40B), bs: 128, sl: 256; Two systems with 4x L40S, vs HGX A100 8 GPU
3. Hugging Face SWIN Base Inference (BS=1,Seq 224); L40S vs. A100 80GB SXM
4. GPT 175B, 300B tokens, Foundational Training; 4K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

L40S Delivers Up to 1.7X A100 Performance[1]

Compared to HGX A100[2]

LLM Training

Up to **1.7X** vs HGX A100



LLM Inference

Up to **1.1X** vs HGX A100

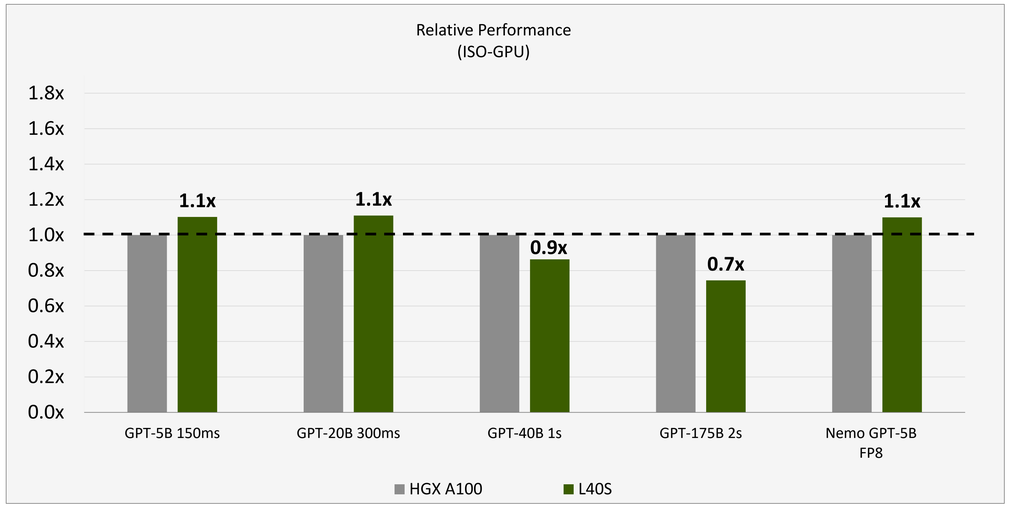
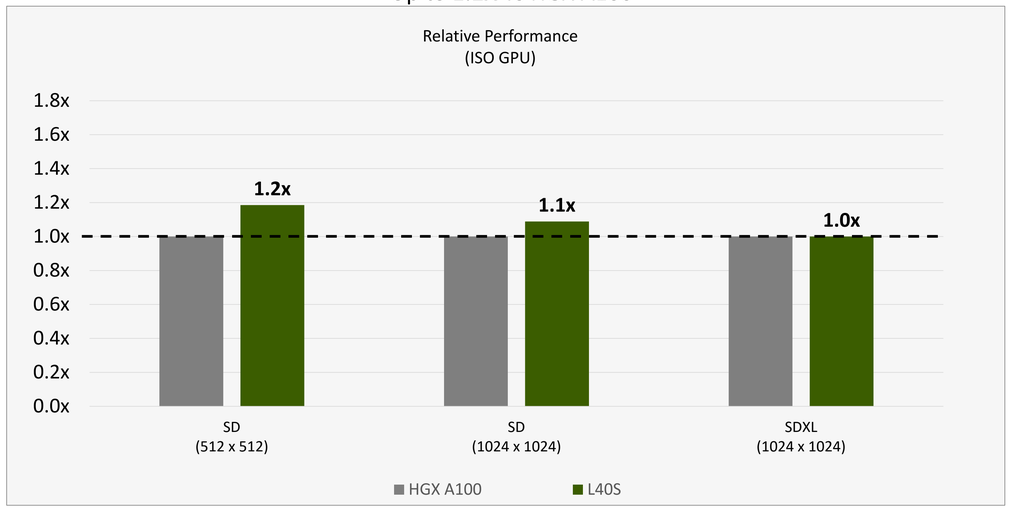


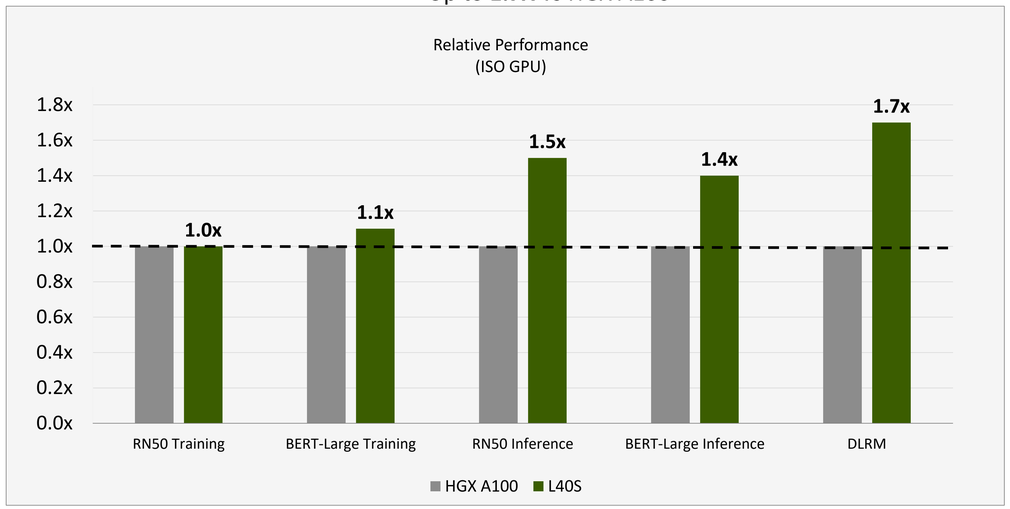
Image Generative AI

Up to **1.2X** vs HGX A100



Traditional DL Inference & Training

Up to **1.7X** vs HGX A100



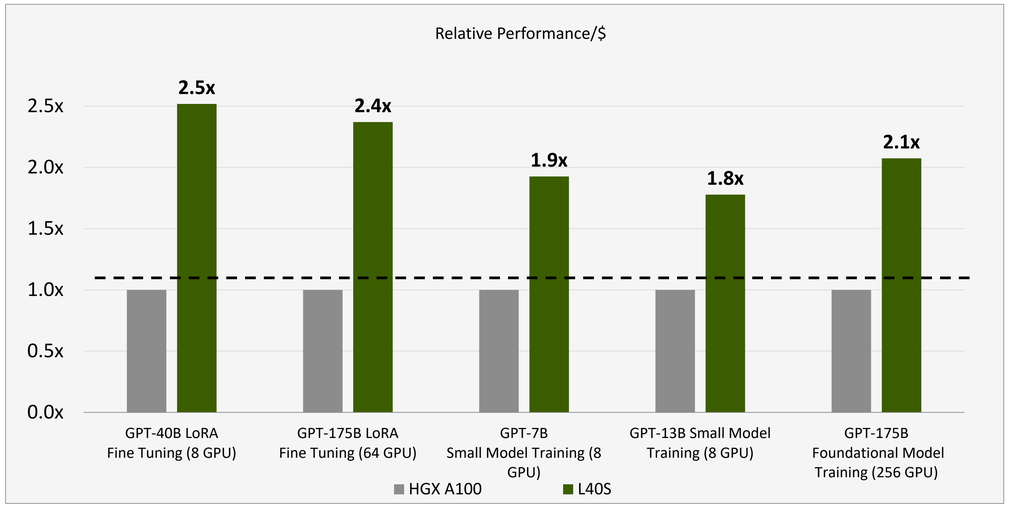
Iso-GPU Performance
1. Two OVX L40S Servers : 8x L40S GPU
2. HGX A100 : HGX A100 8 GPU : 8x A100 80GB SXM

L40S Delivers Up to 2.5X Improved Performance/$

3-year TCO compared to HGX A100[2]

LLM Training

Up to **2.5X** vs HGX A100



LLM Inference

Up to **1.6X** vs HGX A100

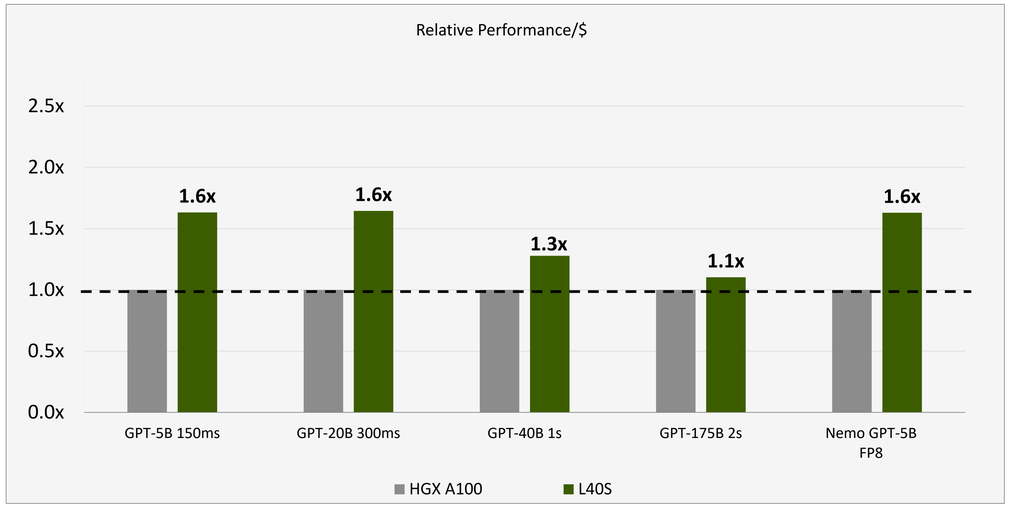
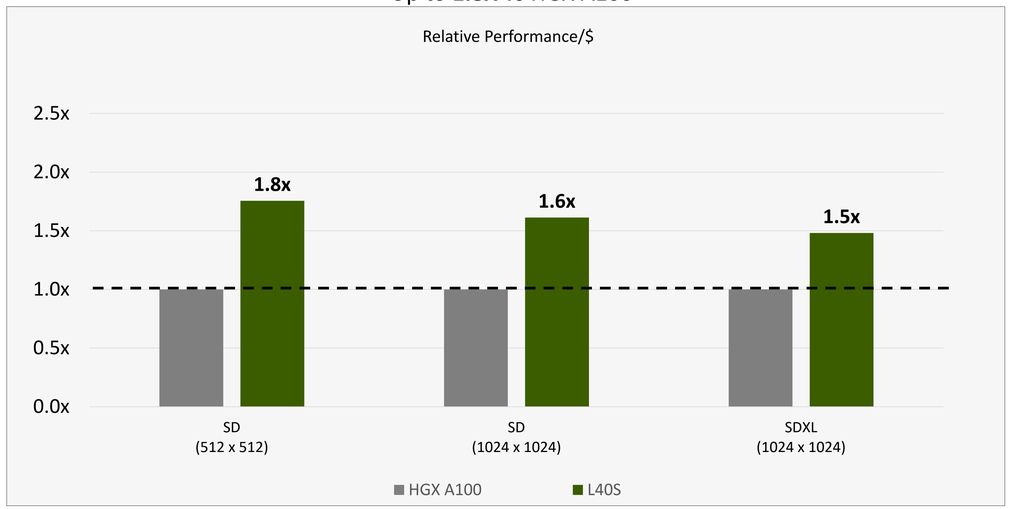


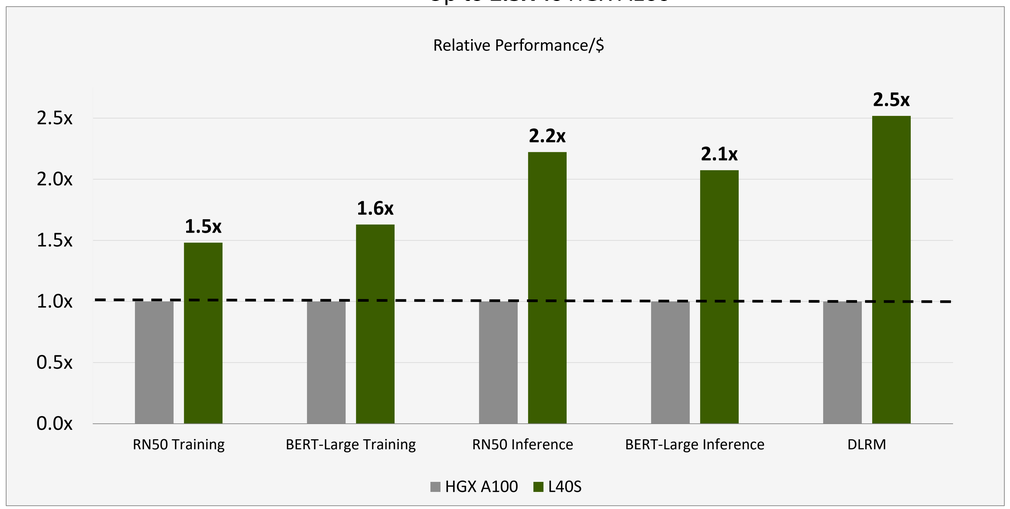
Image Generative AI

Up to **1.8X** vs HGX A100



Traditional DL Inference & Training

Up to **2.5X** vs HGX A100



TCO shown for representation only (includes 3yrs Opex).
System configuration: HGX A100 (4U chassis, 2S 64C CPU) | Two L40S servers w 4 GPU (2U chassis, 2S 32C CPU)
HGX A100 : HGX A100 8 GPU : 8x A100 80GB SXM

Unified Architecture for AI & Graphics Acceleration

NVIDIA L40S

AI TRAINING & DEVELOPMENT

INFERENCE

GENERATIVE AI

COLLABORATION

SIMULATION

INDUSTRIAL DIGITALIZATION

NVIDIA AI ENTERPRISE

NVIDIA Omniverse Enterprise

NVIDIA AI Enterprise **NVIDIA Omniverse**

NVIDIA-Certified Systems

100+ Global Systems ranging from 1-10 L40S

NVIDIA OVX L40S

Reference Architecture Configuration Limited Partners

*** CX7 PN:** MCX755106AC-HEAT / MCX755106AS-HEAT (2x200)

L40S Increases Features and Performance

Accelerating Generative AI, LLM Training, Inference, Omniverse & Rendering

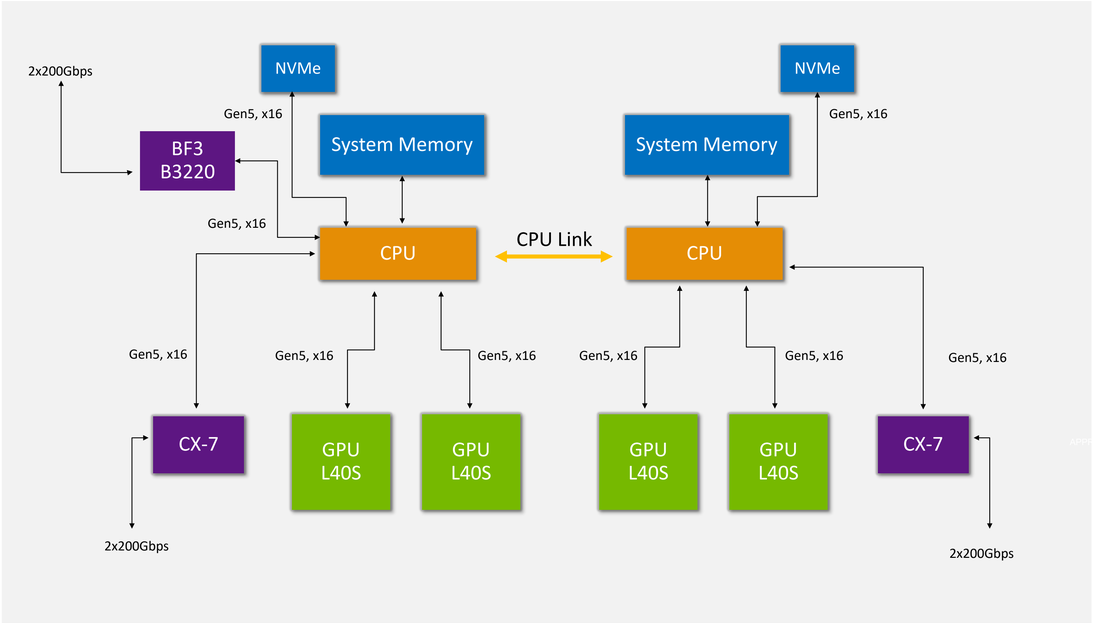
	A100 80GB SXM	L40S
Best For	Highest Perf Multi-Node AI	Universal GPU for Gen AI
GPU Architecture	NVIDIA Ampere	NVIDIA Ada Lovelace
FP64	9.7 TFLOPS	N/A
FP32	19.5 TFLOPS	91.6 TFLOPS
TF32 Tensor Core*	312 TFLOPS	366 TFLOPS
FP16/BF16 Tensor Core*	624 TFLOPS	733 TFLOPS
FP8 Tensor Core*	N/A	1466 TFLOPS
INT8 Tensor Core*	1248 TOPS	1466 TOPS
RT Core	N/A	212 TFLOPS
GPU Memory	80 GB HBM2e	48 GB GDDR6
GPU Memory Bandwidth	2039 GB/s	864 GB/s
L2 Cache	40 MB	96 MB
Media Engines	0 NVENC 5 NVDEC 5 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG
Power	Up to 400 W	Up to 350 W
Form Factor	8-way HGX	2-slot FHFL
Interconnect	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s
Availability	Longer Leadtime	Production

- 5X Increase FP32 for HPC
- FP8 Support for GenAI, LLM Training & Inference
- Supports Ray Tracing for Rendering & Graphics, DLSS3.0 for AI Frame Generation and SER
- Increased L2 Cache
- AV1 for Compressing and Streaming Video
- 350W for Maximum Performance
- Shortest Leadtime Availability

* Specifications with sparsity.

Optimized Architecture for AI Enterprise & Omniverse Workflows

Optimized **2-4-3** Server Configuration



NVIDIA OVX L40S - SERVER	
CPUs	**2x 32c Intel Xeon Gold 6448Y** **2x 32c AMD EPYC 9354**
GPUs	**4x NVIDIA L40S**
Networking - E/W	**2x ConnectX-7 (2x200G)** **(Switch: Either Ethernet w/ Spectrum-3, 200G or InfiniBand w/ Quantum-1, HDR)**
Networking - N/S	**1x BlueField-3, B3220 (2x200G)** **(Switch: Ethernet w/ Spectrum-3, 200G)**
Host Memory	**Min. 384GB total memory w/ one DIMM per channel**
Host Boot Drive	**1x 2TB NVMe**
Host Storage	**2x 4TB NVMe**

2 **CPUs – Intel or AMD options**

4 **GPUs – For graphics and compute**

3 **Networking Adapters**

2x ConnectX-7 for GPU-GPU communication ("east-west")
1x BlueField-3 for management/storage/security ("north-south")

NVIDIA will **Build**, **Tune** and **Optimized** Software Based on this Reference Configuration

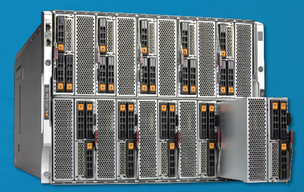
SMC Systems: SYS-521GE, SYS-421GE, SYS-4125GS

***Alternate configuration of 2:8:5 is possible if all PCIe slots are PCIe x16**

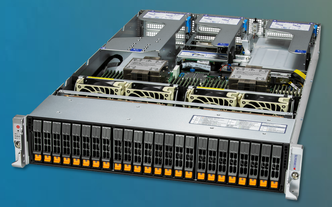
The Industry's Broadest Portfolio of Servers

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

PCIe GPU
High Performance and Flexibility for AI, 3D Simulation and the Metaverse

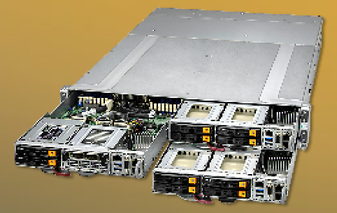


SuperBlade®
Highest Density Multi-Node Architecture for HPC Applications



Hyper
Best-in-class Performance and Flexibility Rackmount Server

BigTwin®
Industry-leading Multi-node Architecture



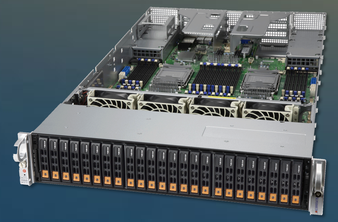
GrandTwin™
Multi-Node Architecture Optimized for Single-Processor Performance

FatTwin®
Multi-Node 4U Advanced Twin Architecture with 8 or 4 Nodes

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

WIO
Industry's Widest Variety of I/O Optimized Servers

Petascale All-Flash
Revolutionary Petascale NVMe for Unprecedented Density and Capacity



Multi-Processor
Highest Performance and Flexibility for Enterprise Applications

Hyper-E
Best-in-class Performance and Flexibility for Edge Data Centers

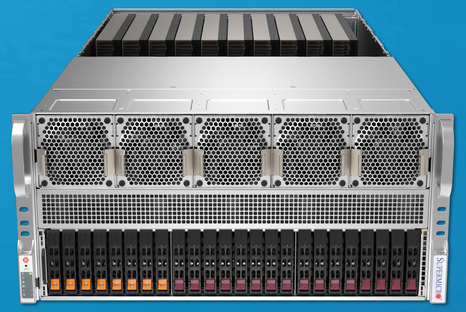
SuperEdge
High-Density Computing and Flexibility at the Intelligent Edge

IoT/5G
Compact Form Factors for 5G and Edge computing

SuperWorkstation
For High Performance Desktop Workloads

Mainstream
Versatile Entry-Level for Enterprise Applications

L40S Optimized Broadest Portfolio of Servers



8-10 PCIe GPU Systems

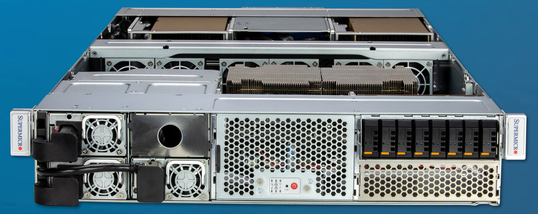
High Performance and Flexibility for AI, 3D Simulation and the Metaverse

SuperBlade®

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

Hyper

Best-in-class Performance and Flexibility Rackmount Server



MGX Systems

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

CloudDC

All-in-one Rackmount Platform for Cloud Data Centers

Hyper-E

Best-in-class Performance and Flexibility for Edge Data Centers

Networking Adapters	3	5	5
GPUs	4	8	8
CPUs	2	2	2

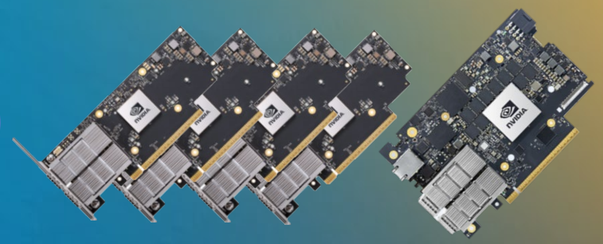
SYS-741GE-TNRT

SYS-521GE-TNRT

SYS-421GE-TNRT

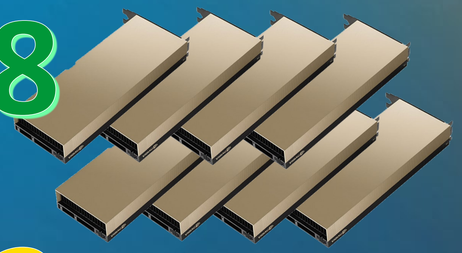
Networking Adapters

5



GPUs

8



CPUs

2

AS -4125GS-TNRT2

NVIDIA L40S Supported Supermicro Systems

SKU	Supported GPUs (under "GPU Section" of spec page)
SYS-421GE-TNRT	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100
SYS-521GE-TNRT	NVIDIA PCIe: H100, L40S, L40, A100
AS -4125GS-TNRT	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100,
SYS-741GE-TNRT	NVIDIA PCIe: H100, L40S, L40, A100
SYS-221GE-NR	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100
ARS-221GL-NR	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100
AS -4125GS-TNRT1	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100
AS -4125GS-TNRT2	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100
SYS-421GE-TNRT3	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100
SBI-611E-1C2N	NVIDIA PCIe: H100, L40S, L40, A100
SBI-611E-1T2N	NVIDIA PCIe: H100, L40S, L40, A100
SBI-611E-5T2N	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100
SBI-411E-1G	NVIDIA PCIe: H100, L40S, L40, A100
SBI-411E-5G	NVIDIA PCIe: H100, L40S, L40, A100
SYS-121H-TNR	NVIDIA PCIe: H100, L40S, L40, A100
SYS-221H-TNR	NVIDIA PCIe: H100, L40S, L40, A100
SYS-221H-TN24R	NVIDIA PCIe: H100, L40S, L40, A100
SYS-241H-TNRTTP	NVIDIA PCIe: H100, L40S, L40, A100
AS -2015HS-TNR	NVIDIA PCIe: H100, L40S, L40, A100
AS -2025HS-TNR	NVIDIA PCIe: H100, L40S, L40, A100
SYS-221HE-FTNR	NVIDIA PCIe: H100, L40S, L40, A100
SYS-221HE-FTNRD	NVIDIA PCIe: H100, L40S, L40, A100
SYS-521C-NR	NVIDIA PCIe: H100, L40S, L40, A100
SYS-621C-TN12R	NVIDIA PCIe: H100, L40S, L40, A100
AS -2015CS-TNR	NVIDIA PCIe: H100, L40S, L40, A100

Updated all spec pages with L40S support

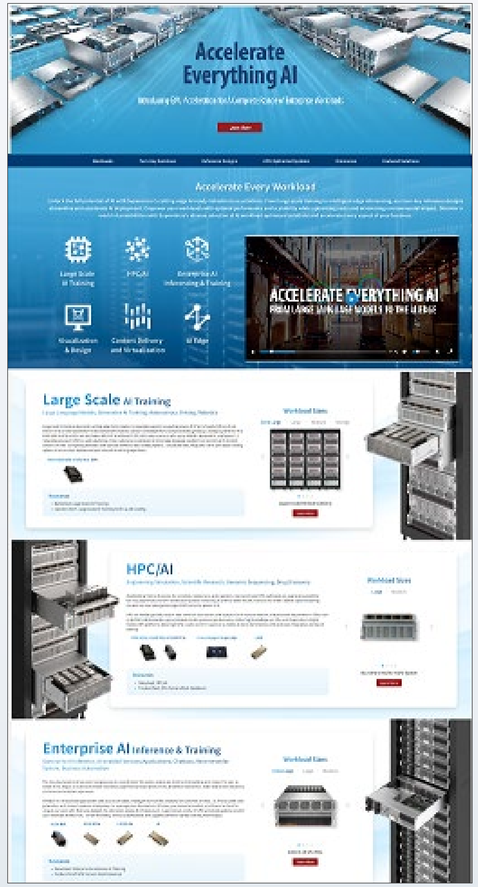
Processor	
CPU	Dual Socket E (LGA-4677) 4th Gen Intel® Xeon® Scalable processors
Note	Supports up to 350W TDP CPUs (Air Cooled) Supports up to 350W TDP CPUs (Liquid Cooled)
GPU	
Supported GPU	NVIDIA PCIe: H100, L40S, L40, A100
CPU-GPU Interconnect	PCIe 5.0 x16 Switch Dual-Root
GPU-GPU Interconnect	NVIDIA® NVLink™ Bridge (optional)

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!

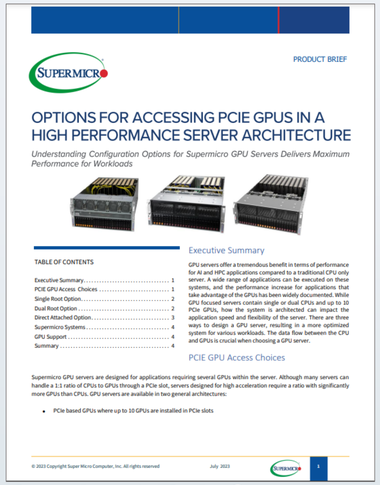
Leverage Sales Assets

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