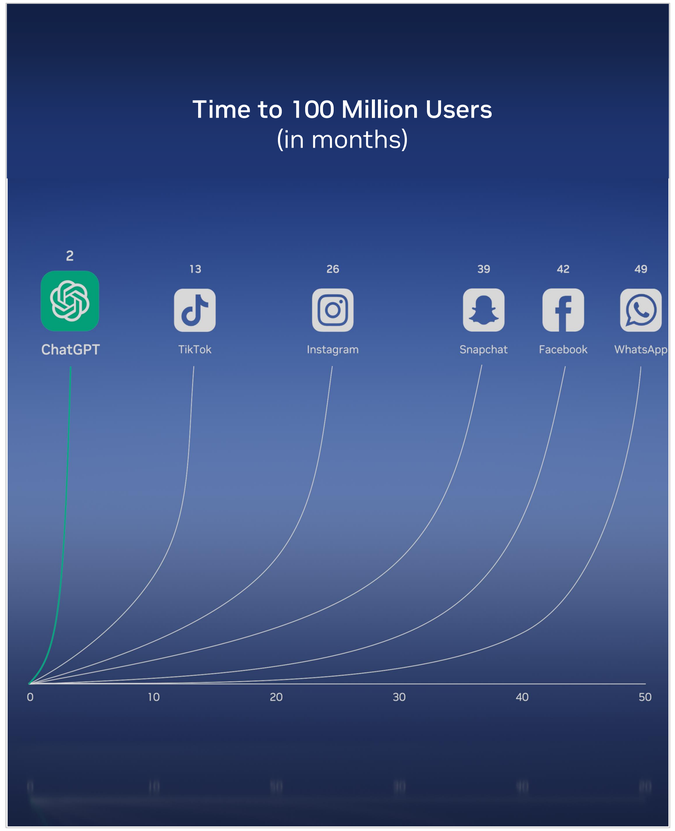
NVIDIA L40S Product Deck

August 2023

The iPhone Moment of AI is Here

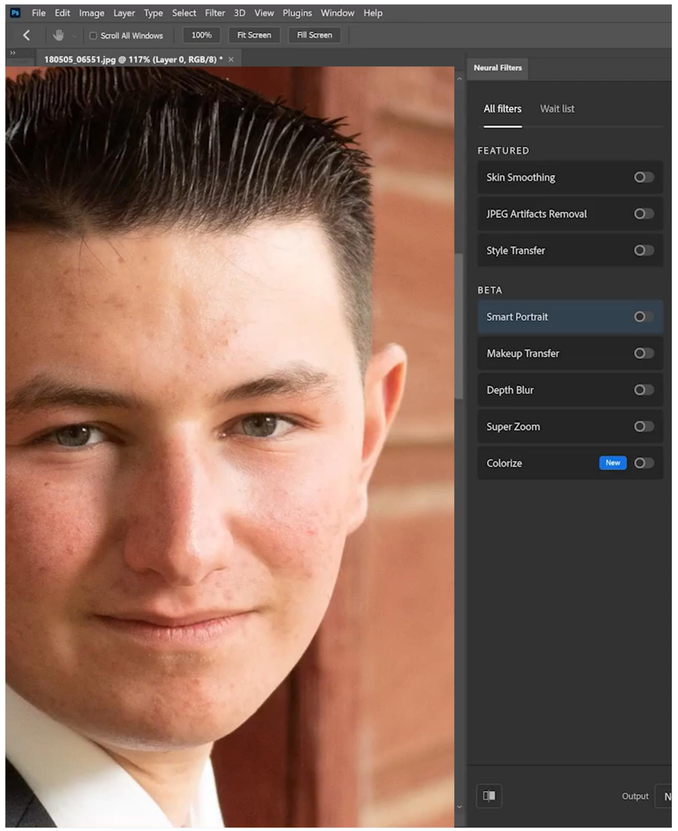
Every major application and workflow is going to include AI

CHATBOTS
Fastest Growing Application Ever

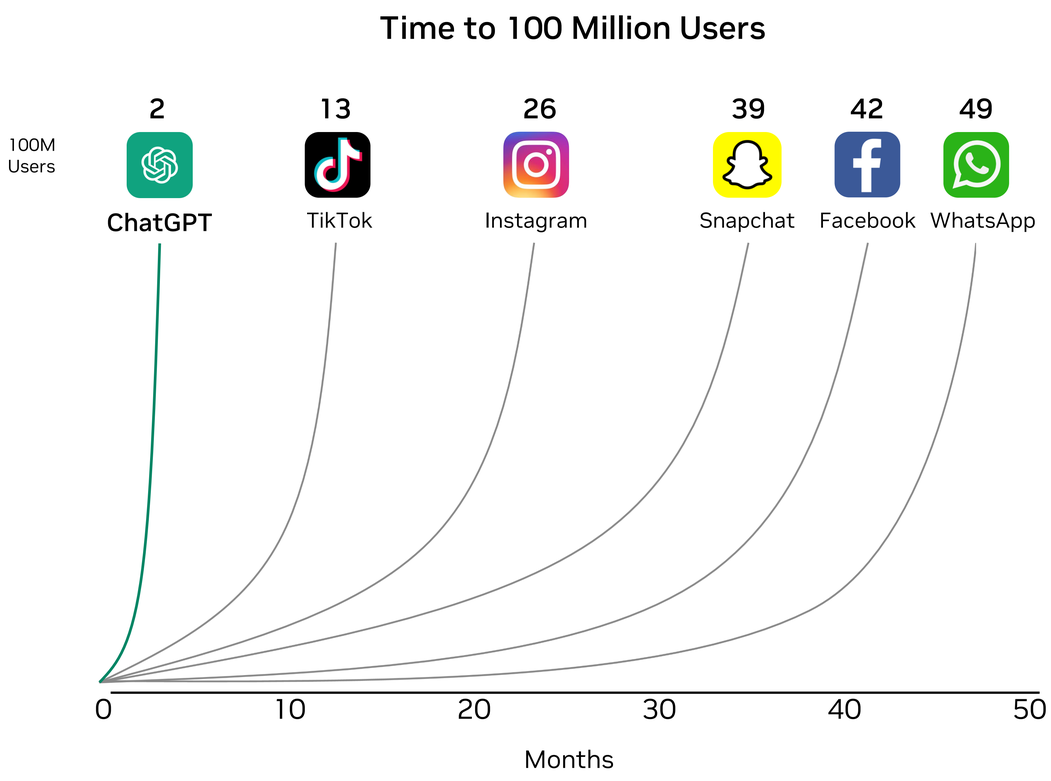


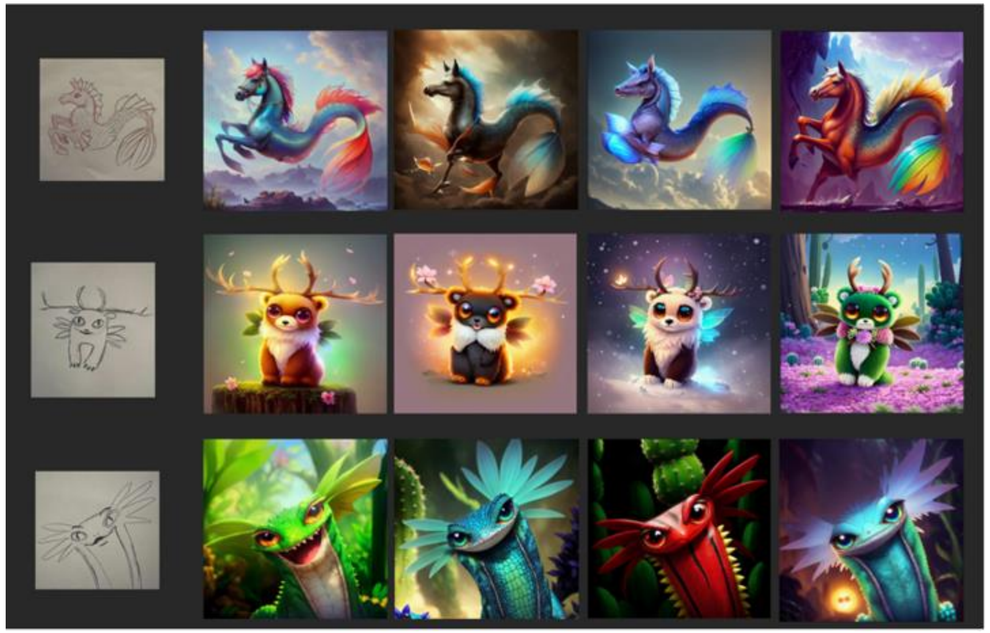
GENERATIVE ART
Over 200M+ Users

AI-AUGMENTED APPLICATIONS
ISVs Accelerating AI Integration



Rise of Generative AI





"ChatGPT managed to beat the popularity of TikTok"

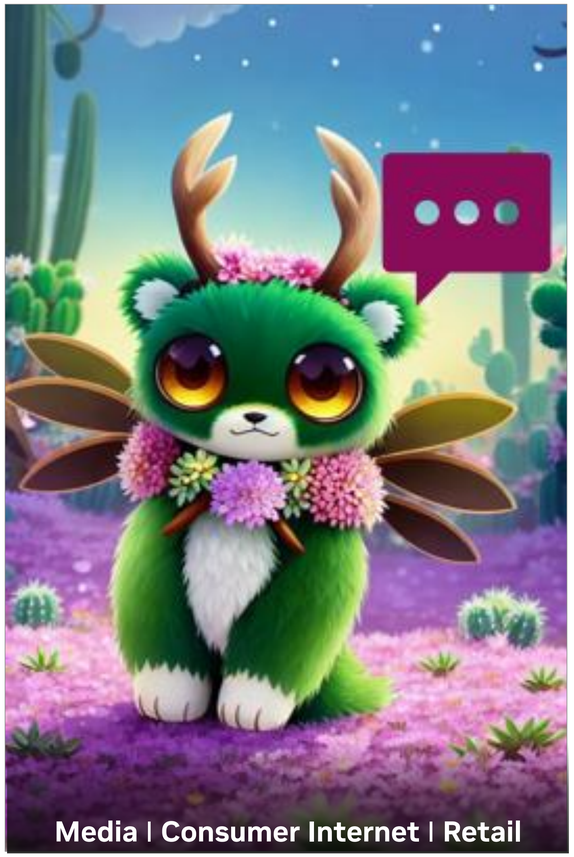
Forbes

"Stable Diffusion has more than 10 million daily users across all channels"

Bloomberg, Oct 2022

Diverse Workloads Require Flexible Compute

Enterprises need universal accelerator to handle various workloads

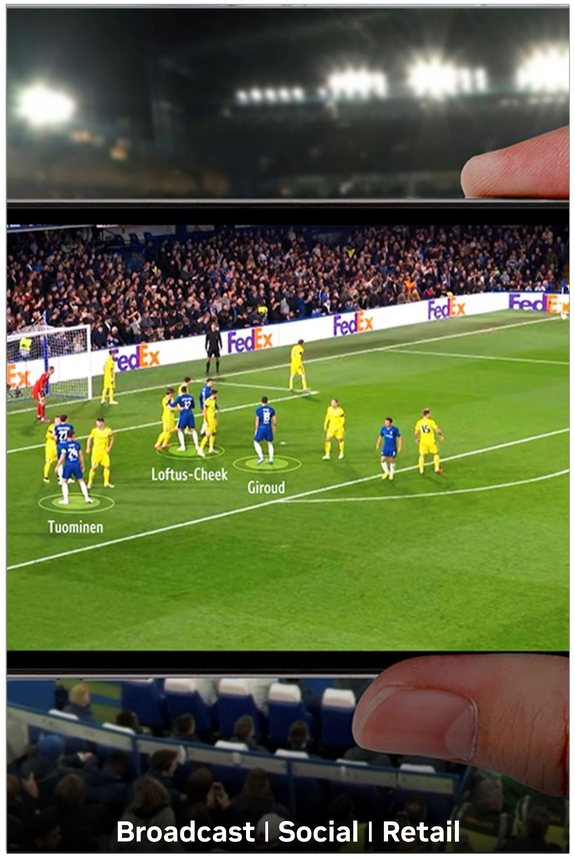


Generative AI

Mainstream Language Models
Creative AI – Image Generation

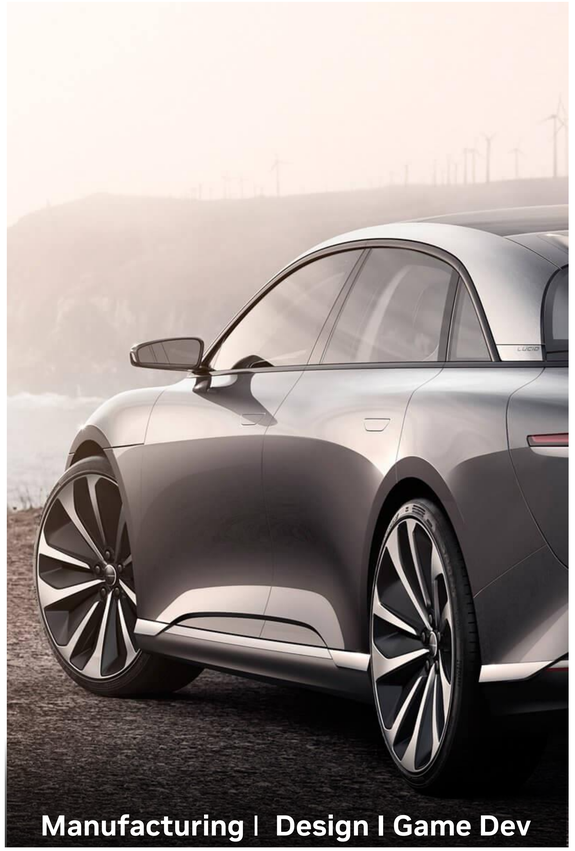
Mainstream Compute

AI training & Inference
Algorithmic Analysis



AI Video

Encoding/Decoding
Live streaming



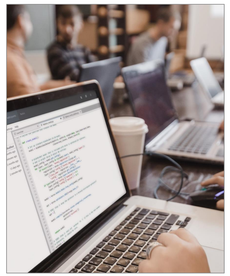
Premium Graphics

AR/VR
Omniverse

NVIDIA RTX

Built for AI and graphics-intensive workflows

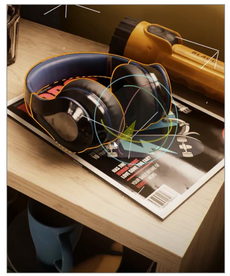
AI TRAINING & DEVELOPMENT



INFERENCE

GENERATIVE AI

CONTENT CREATION



COLLABORATION

SIMULATION

INDUSTRIAL DIGITALIZATION

NVIDIA AI Enterprise

NVIDIA Omniverse Enterprise

Workstation

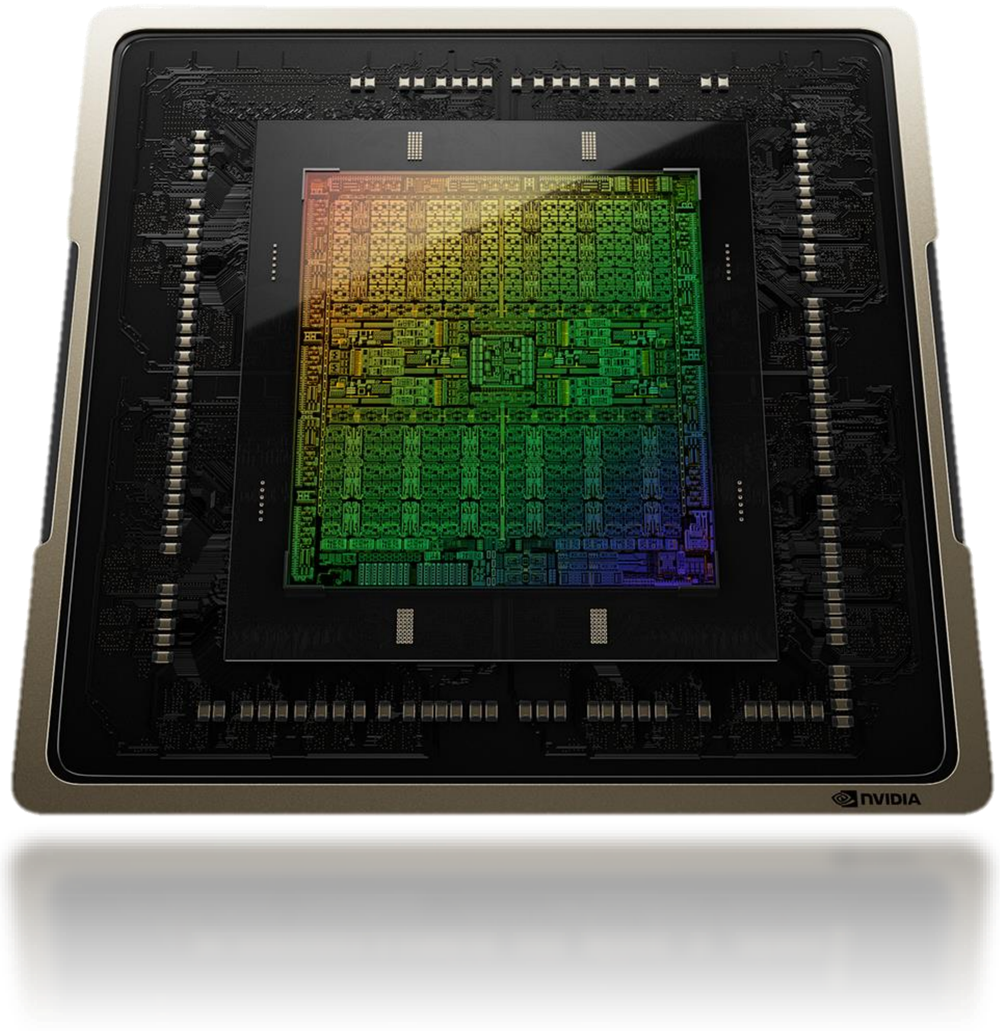
ovx

Cloud

Introducing NVIDIA L40S

Unparalleled AI and Graphics Performance for the Data Center.

NVIDIA L40S
Based on the Ada Lovelace Architecture



New Ada Architecture Features

- New Streaming Multiprocessor
- 4th-Gen Tensor Cores
- 3rd-Gen RT Cores
- 91.6 teraFLOPS FP32

Gen-AI, LLM Training, & Inference

- Transformer Engine - FP8
- 1.5 petaFLOPS Tensor Performance*
- Large L2 Cache

3D Graphics & Rendering

- 212 teraFLOPS RT Core Performance
- DLSS 3.0, AI Frame Generation
- Shader Execution Reordering

Media Acceleration

- 3 Encode & 3 Decode Engines
- 4 JPEG Decoders
- AV1 Encode & Decode Support

*Peak teraFLOPS, sparsity enabled

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

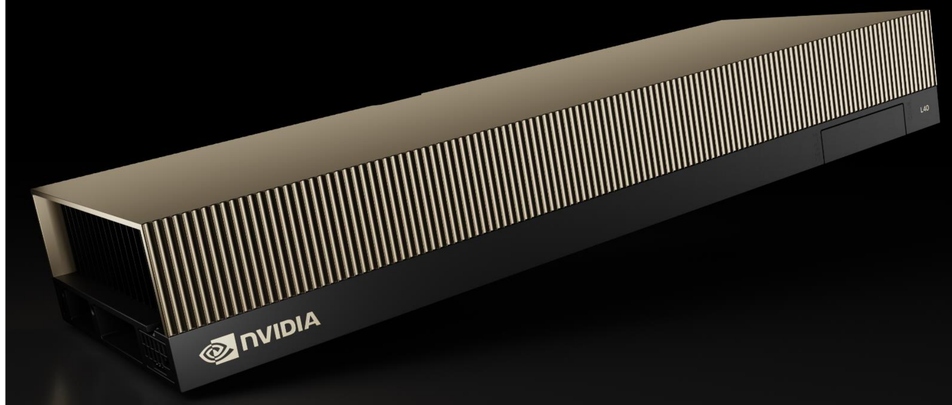
GPT-175 300B Tokens[4]

Image Gen AI
>82

Images per minute[5]

Full Video Pipeline
184

AV1 Encode Streams[6]

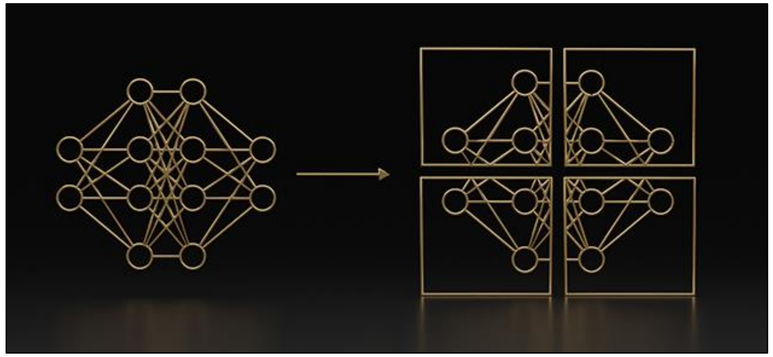


Dual-Slot I FHFL I 350W

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

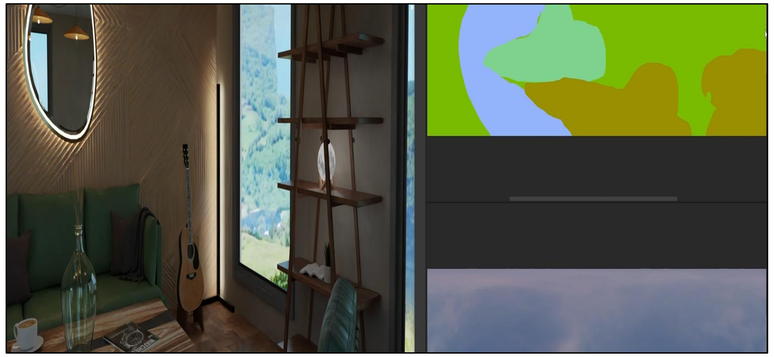
Powerful Multi-Workload Acceleration

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

L40S Value Proposition

Powerful AI & Graphics, Data Center Ready, Available in August

Performance	Value	Availability
Powerful AI + Graphics	Better Price-Performance	Short Lead Time
Data Center Scale	Accelerate many workloads	Fast deployment

Price-performance L40S vs A100 **

NVIDIA L40S Specifications

	L40S	A100 80GB SXM
Best For	Universal GPU for Gen AI	Highest Perf Multi-Node AI
GPU Architecture	NVIDIA Ada Lovelace	NVIDIA Ampere
FP64	N/A	9.7 TFLOPS
FP32	91.6 TFLOPS	19.5 TFLOPS
RT Core	212 TFLOPS	N/A
TF32 Tensor Core*	366 TFLOPS	312 TFLOPS
FP16/BF16 Tensor Core*	733 TFLOPS	624 TFLOPS
FP8 Tensor Core*	1466 TFLOPS	N/A
INT8 Tensor Core*	1466 TOPS	1248 TOPS
GPU Memory	48 GB GDDR6	80 GB HBM2e
GPU Memory Bandwidth	864 GB/s	2039 GB/s
L2 Cache	96 MB	40 MB
Media Engines	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	0 NVENC 5 NVDEC 5 NVJPEG
Power	Up to 350 W	Up to 400 W
Form Factor	2-slot FHFL	8-way HGX
Interconnect	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s
Availability	August 2023	Longer Leadtime

* Specifications with sparsity.

LLM Training & Inference

NVIDIA L40S For LLM Training

Great solution for Fine Tuning, Training Small Models and Small/Mid Scale Training up to 4K GPU

Fine-Tuning Existing Models[1] (Time to Train 860M Tokens)			
		Expected Speedup w TE/FP8	
	HGX A100	**L40S**	**HGX H100**
GPT-40B LoRA (8 GPU)	12 hrs.	**1.7x**	**4.4x**
GPT-175B LoRA (64 GPU)	6 hrs.	**1.6x**	**4.3x**

Training Small Models[2] (Time to Train 10B Tokens)			
		Expected Speedup w TE/FP8	
	HGX A100	**L40S**	**HGX H100**
GPT-7B (8 GPU)	17 hrs.	**1.3x**	**3.4x**
GPT-13B (8 GPU)	32 hrs.	**1.2x**	**3.6x**

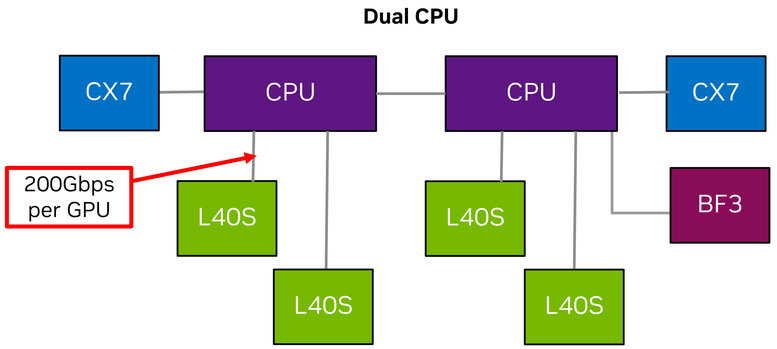
Training Foundation Models[3] (Time to Train 300B Tokens)			
		Expected Speedup w TE/FP8	
	HGX A100	**L40S**	**HGX H100**
GPT-175B (256 GPU)	64 days	**1.4x**	**4.5x**
GPT-175B (1K GPU)	16 days	**1.3x**	**4.6x**
GPT-175B (4K GPU)	4 days	**1.2x**	**4.1x**

Preliminary performance projections, subject to change.

HGX A100 8 GPU : 8x A100 80GB SXM

1. Fine-Tuning LoRA (GPT-40B) 8 GPUs, Fine-Tuning LoRA (GPT-175B), 64 GPUs: global train batch size: 128 (sequences), seq-length: 256 (tokens)
2. Small model Training (GPT-7B, GPT-13B) 8 GPUs: global train batch size: 512 (sequences), seq-length: 2048 (tokens)
3. Large model Training (GPT-175B) 256-4K GPUs: global train batch size: 2048 (sequences), seq-length: 2048 (tokens)

Recommended System Configuration

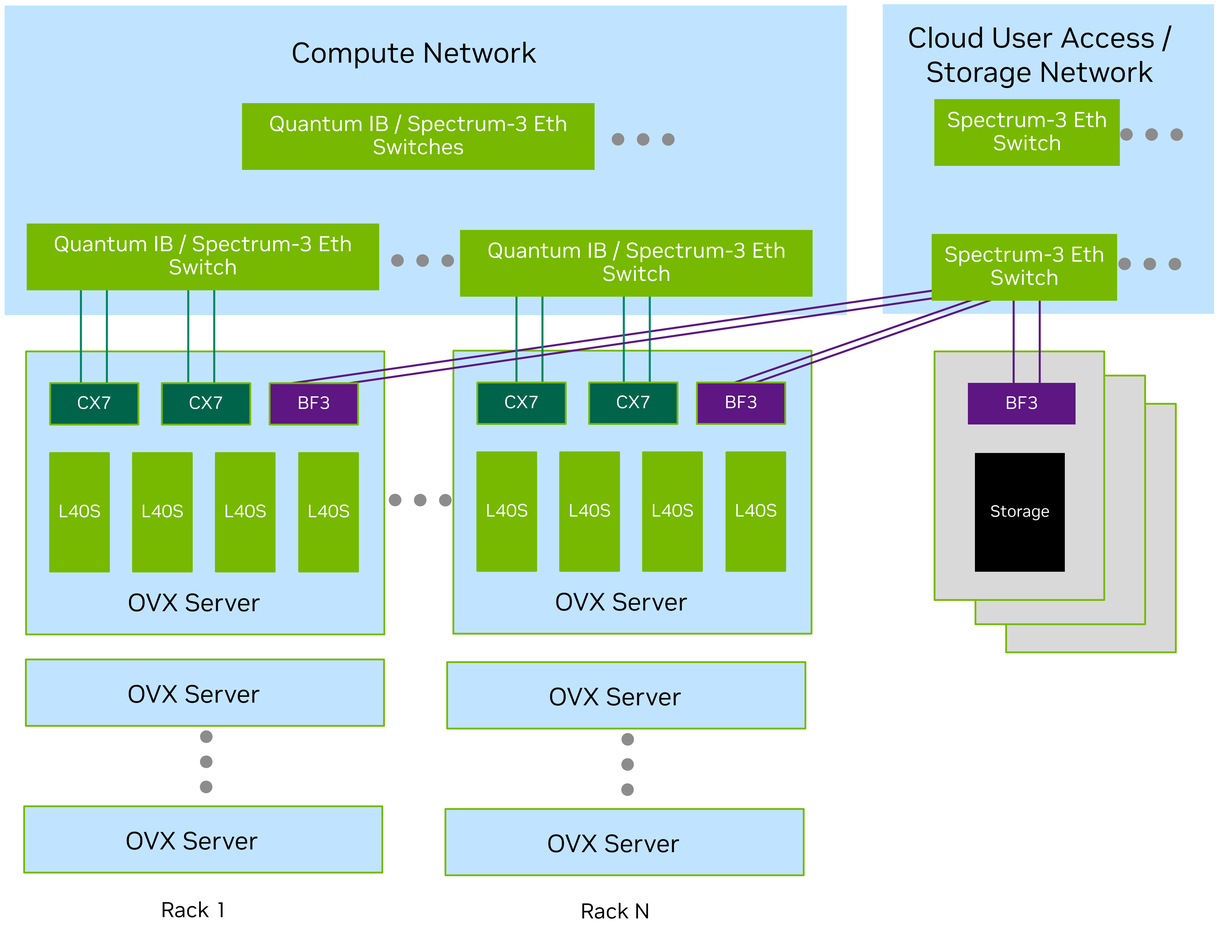


E/W Traffic: 200Gbps network bandwidth per L40S is recommended. Dual-Port 200Gbps CX-7; Ethernet or InfiniBand

N/S Traffic- Bluefield-3 DPU recommended

NVIDIA Full Stack-Optimized Reference Architecture

With Quantum 200G InfiniBand or Spectrum 200G Ethernet



200G Reference Architecture Supports the Following Devices:
ConnectX-7 NIC (2x200G ports) and BlueField-3 DPU (2x200G ports)
Quantum 8700 Series, 16Tb/s, 1U rack mount, 40-ports, 200G per port
Spectrum-3 SN4600 Series, 12.8Tb/s, 2U rack mount, 64-ports, 200G per port

NCCL-optimized AI Training and Inference Platform

Full solution (compute and network) Available Now

Performance-Driven 200G Computing Network

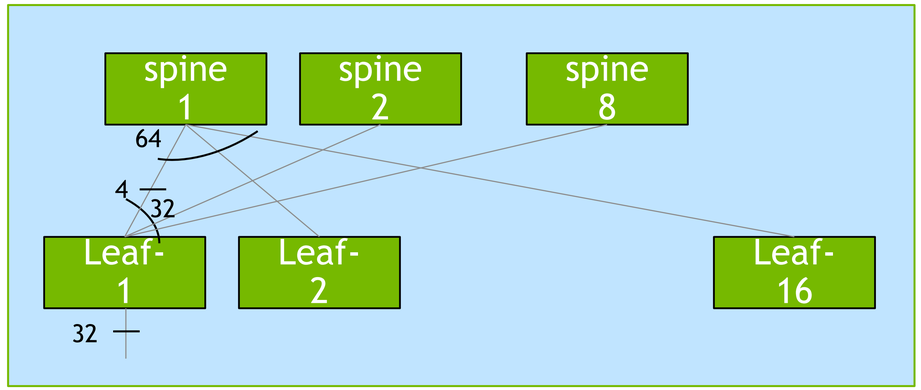
- Quantum InfiniBand with In-Network Computing for optimized performance at scale and under load
- Spectrum Ethernet delivering Ethernet for AI networking infrastructure with RoCE

Enhanced Cloud Management and User Access

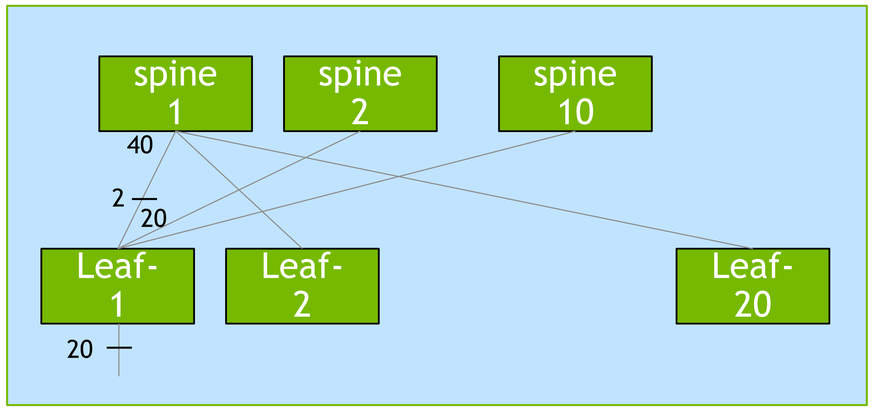
- BlueField-3 DPU Enables Secure, Multi-Tenant Environments
- Streamline Operations with SDN Acceleration, Software-Defined Storage, and Hardware Offloads
- Programmable Arm Cores

Compute Network Examples For Reference Architecture

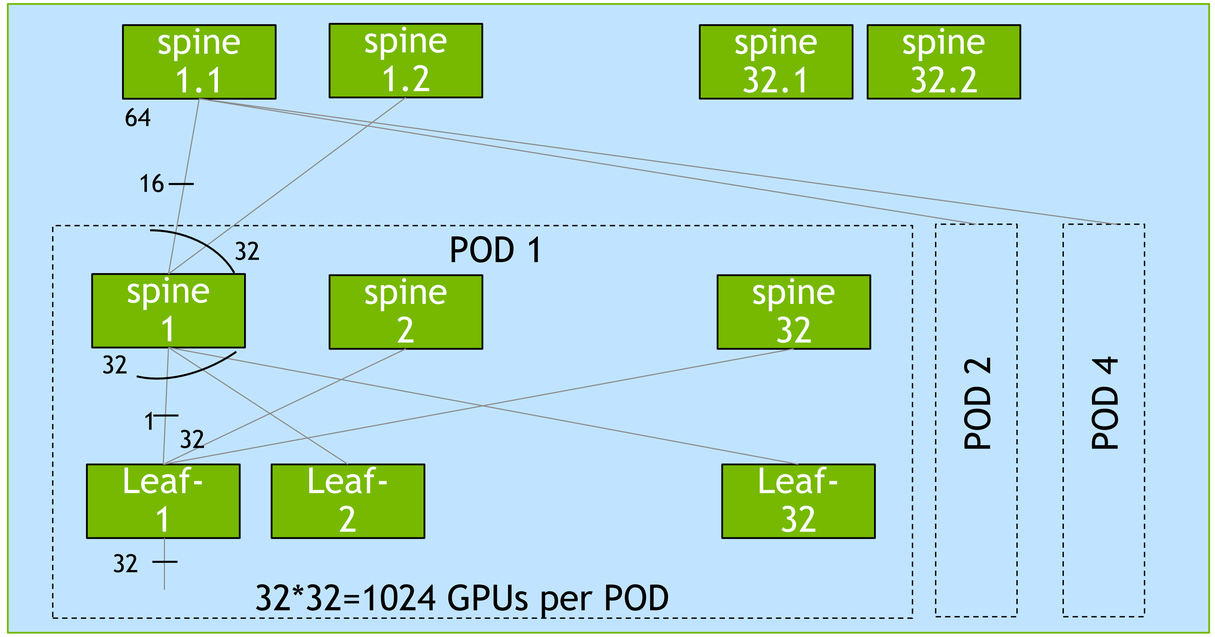
512 L40S System using Spectrum 200G Ethernet



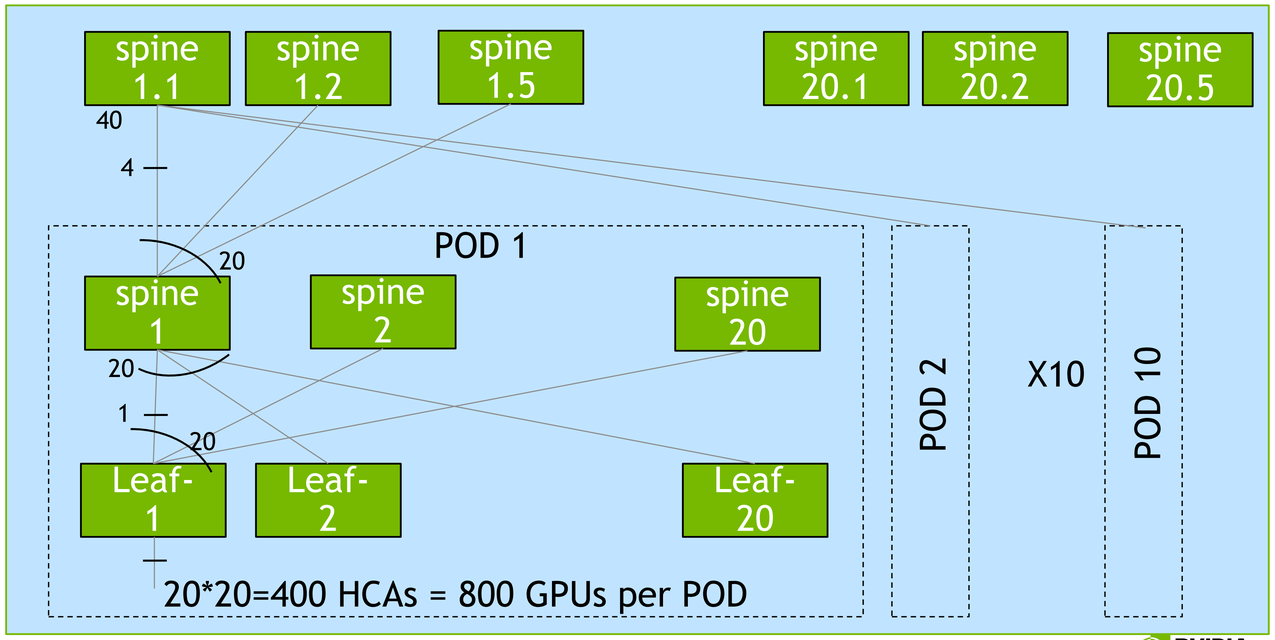
400 L40S System using Quantum 200G InfiniBand



4096 L40S System using Spectrum 200G Ethernet



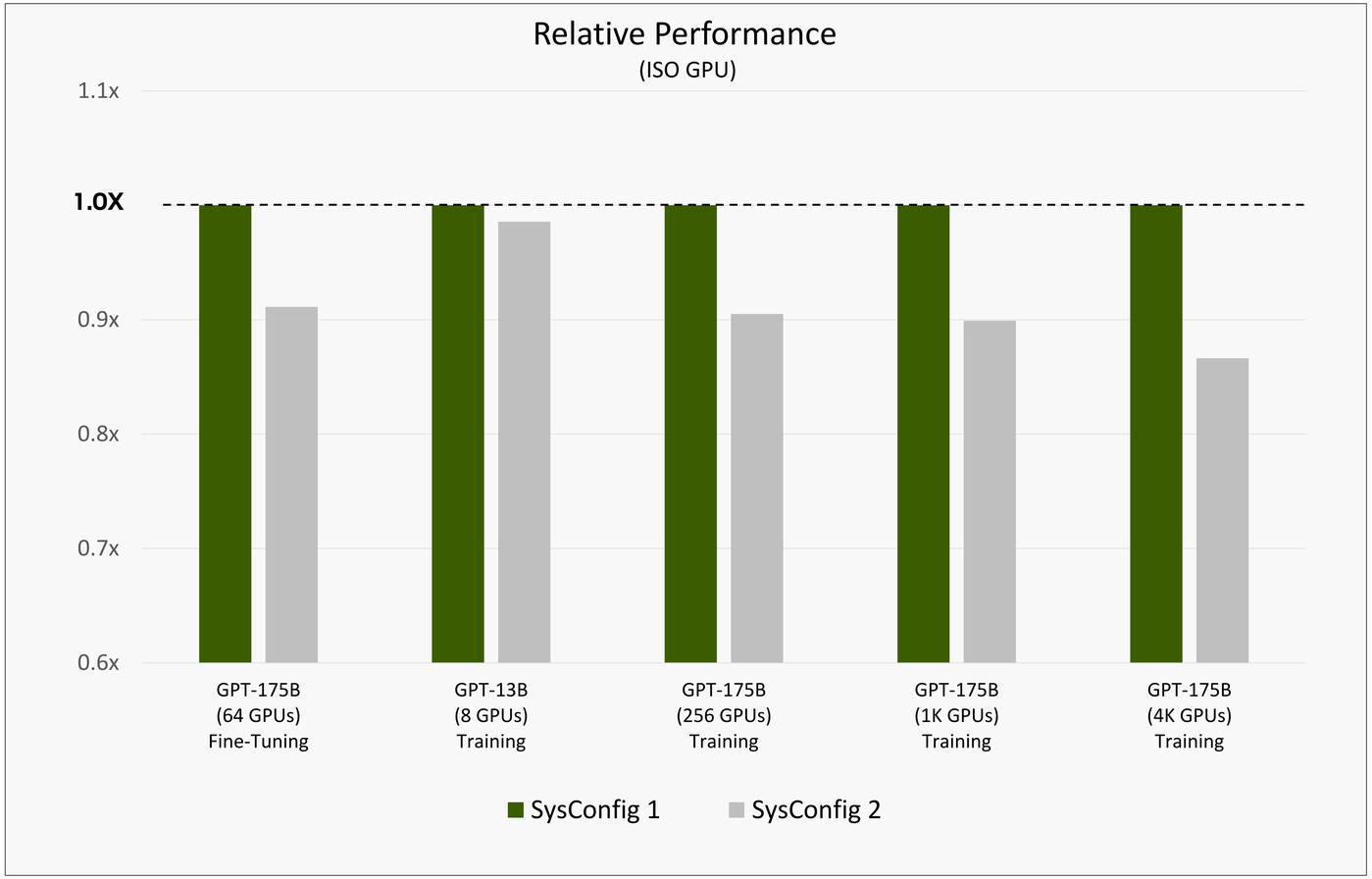
4000 L40S System using Quantum 200G InfiniBand



Please work with NVIDIA on the specific networking kit for each example.

Recommended Configuration for High Performance

4 GPU Configurations Deliver More Performance with Shorter Lead Time

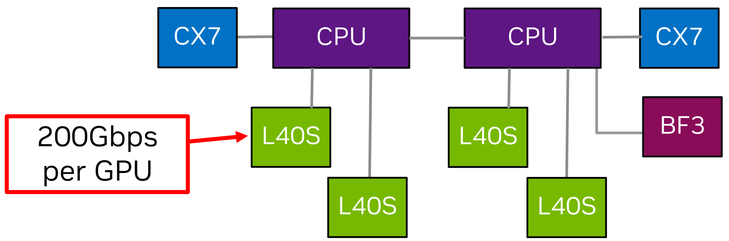


Preliminary performance projections, subject to change

1. Fine-Tuning LoRA (GPT-175B): global train batch size: 128 (sequences), seq-length: 256 (tokens)
2. Small model Training (GPT-7B, GPT-13B): global train batch size: 512 (sequences), seq-length: 2048 (tokens)
3. Large model Training (GPT-175B): global train batch size: 2048 (sequences), seq-length: 2048 (tokens)

System Configuration 1: **4x L40S**

Recommended

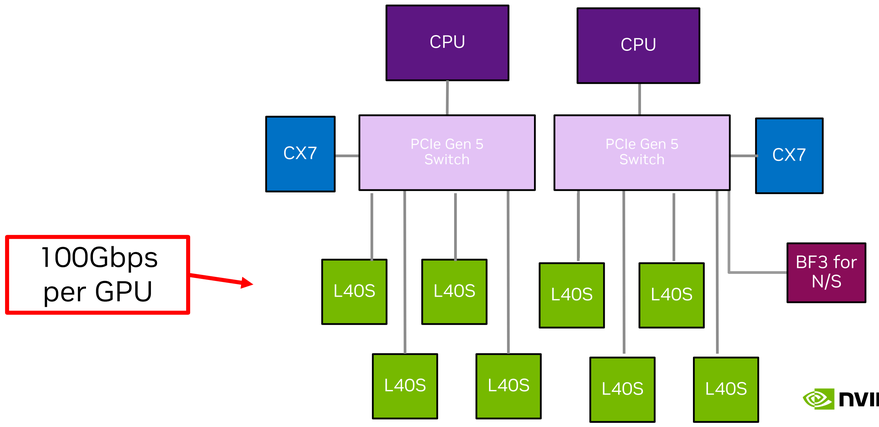


E/W Traffic: 200Gbps network bandwidth per L40S is recommended. Dual-Port 200Gbps CX-7; Ethernet or Infiniband

N/S Traffic- Bluefield-3 DPU recommended

System Configuration 2: **8x L40S**

Consideration: Long lead-times for PCIe Gen5 Switch



Measured Performance

L40S vs. A100 80GB SXM

	Benchmarks	# GPUs	Precision	Metric	A100[1]	L40S	Relative Performance L40S/A100
DL Training	GPT 7B[2](GBS=512)	**8**	**FP16/FP8**	**Samples/sec**	13.5	15.9	**1.2x**
	ResNet-50 V1.5 Training (BS=32)	**1**	**FP16**	**Images/sec**	2707	2748	**1.0x**
	BERT Large Pre-Training Phase 1 (BS=128, seq 512)	**1**	**FP16**	**Sequences/sec**	579	472	**0.8x**
	BERT Large Pre-Training Phase 2 (BS=8, seq 512)	**1**	**FP16**	**Sequences/sec**	152	161	**1.1x**
DL Inference	ResNet-50 V1.5 Inference (BS=32)	**1**	**INT8**	**Images/sec**	23439	34588	**1.5x**
	BERT Large Inference (BS=8, seq 128)	**1**	**INT8**	**Sequences/sec**	3011	4090	**1.3x**
	BERT Large Inference (BS=8, seq 384)	**1**	**INT8**	**Sequences/sec**	1116	1598	**1.4x**
	BERT Large Inference (BS=128, seq 128)	**1**	**INT8**	**Sequences/sec**	5065	5273	**1.0x**
	BERT Large Inference (BS=128, seq 384)	**1**	**INT8**	**Sequences/sec**	1445	1558	**1.1x**
Stable Diffusion	Demo Diffusion 2.1 Inference (BS=1, 512x512)	**1**	**FP16**	**Pipeline Latency (ms)**	827	743	**1.1x**
	Demo Diffusion 2.1 Inference (BS=1, 1024x1024)	**1**	**FP16**	**Pipeline Latency (ms)**	4186	3582	**1.2x**
	Stable Diffusion XL (BS=1, PyTorch native)	**1**	**FP16**	**Pipeline Latency (ms)**	10450	11194	**0.9x**
	Stable Diffusion XL (BS=1, PyTorch optimized)	**1**	**FP16**	**Pipeline Latency (ms)**	7353	7382	**1.0x**
	Stable Diffusion XL (BS=1, TRT optimized)	**1**	**FP16**	**Pipeline Latency (ms)**	5251	5547	**1.0x**

Preliminary performance projections, subject to change.
1. A100 80GB SXM
2. GPT-7B mapping. L40S : MBS=1, TP=1, PP=2, DP=4 | A100 : MBS =4, TP=2, PP=1, DP=4

Measured Performance

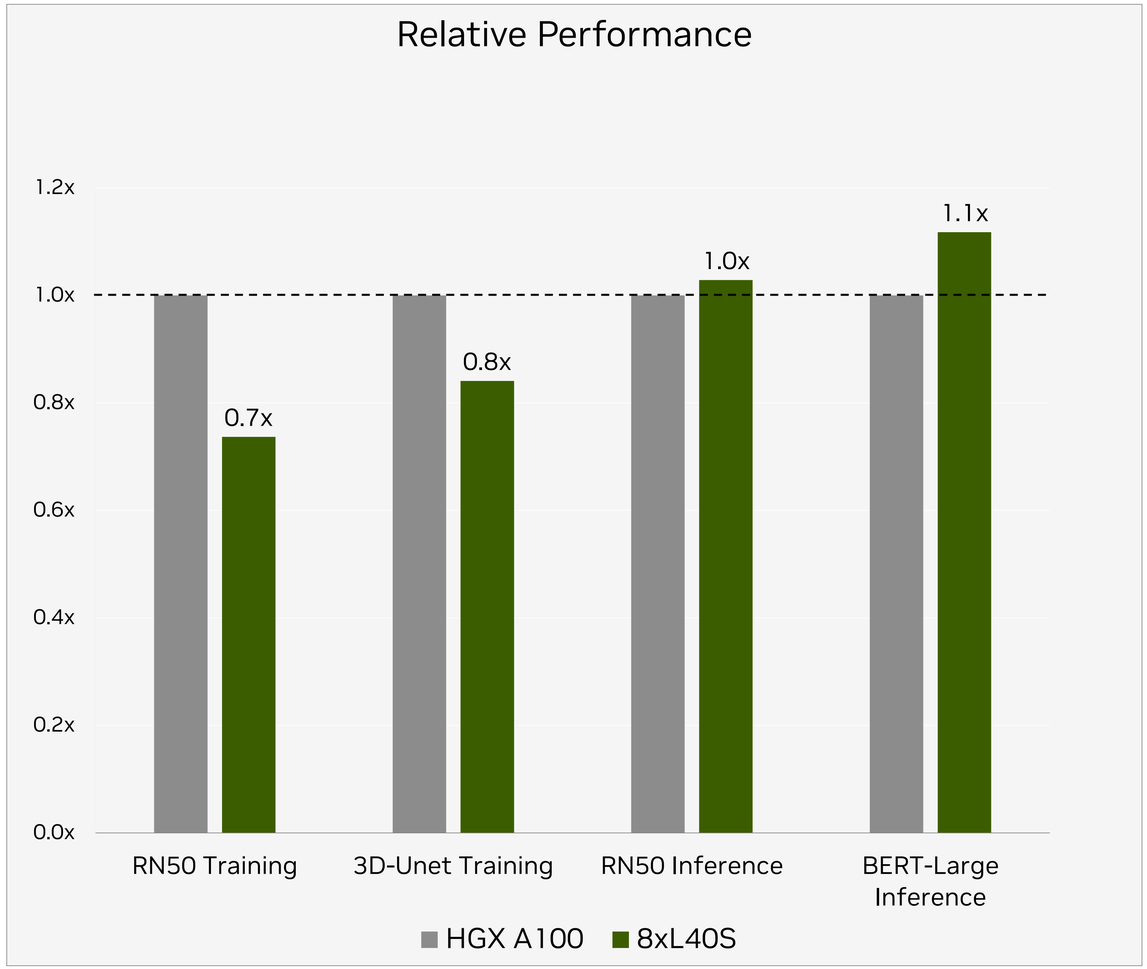
L40S vs. A100 80GB SXM

	Benchmarks	# GPUs	Precision	Metrics	A100[1]	L40S	Relative Performance L40S/A100
DL Inference	GPT2 Inference (BS=1)	1	FP16	Samples/sec	1333	1828	1.4x
	GPT2 Inference (BS=32)	1	FP16	Samples/sec	6502	7578	1.2x
	GPT2 Inference (BS=128)	1	FP16	Samples/sec	6850	6701	1.0x
	DLRM (BS=1)	1	TF32	Records/sec	6495	9458	1.5x
	DLRM (BS=64)	1	TF32	Records/sec	319131	517072	1.6x
	DLRM (BS=2048)	1	TF32	Records/sec	4668287	6980429	1.5x
	ViT Inference (BS=32, seq 224)	1	FP16	Samples per Second	1556	1477	1.0x
	ViT Inference (BS=32, seq 384)	1	FP16	Samples per Second	501	404	0.8x
	HF Swin Base Inference (BS=1,Seq 224)	1	INT8	Samples per Second	633	920	1.5x
	HF Swin Base Inference (BS=32,Seq 224)	1	INT8	Samples per Second	2998	3564	1.2x
	HF Swin Large Inference (BS=1,Seq 384)	1	INT8	Samples per Second	345	411	1.2x
	HF Swin Large Inference (BS=32,Seq 384)	1	INT8	Samples per Second	570	478	0.8x

Preliminary performance projections, subject to change.
1. A100 80GB SXM

L40S Delivers A100-Level Performance for AI

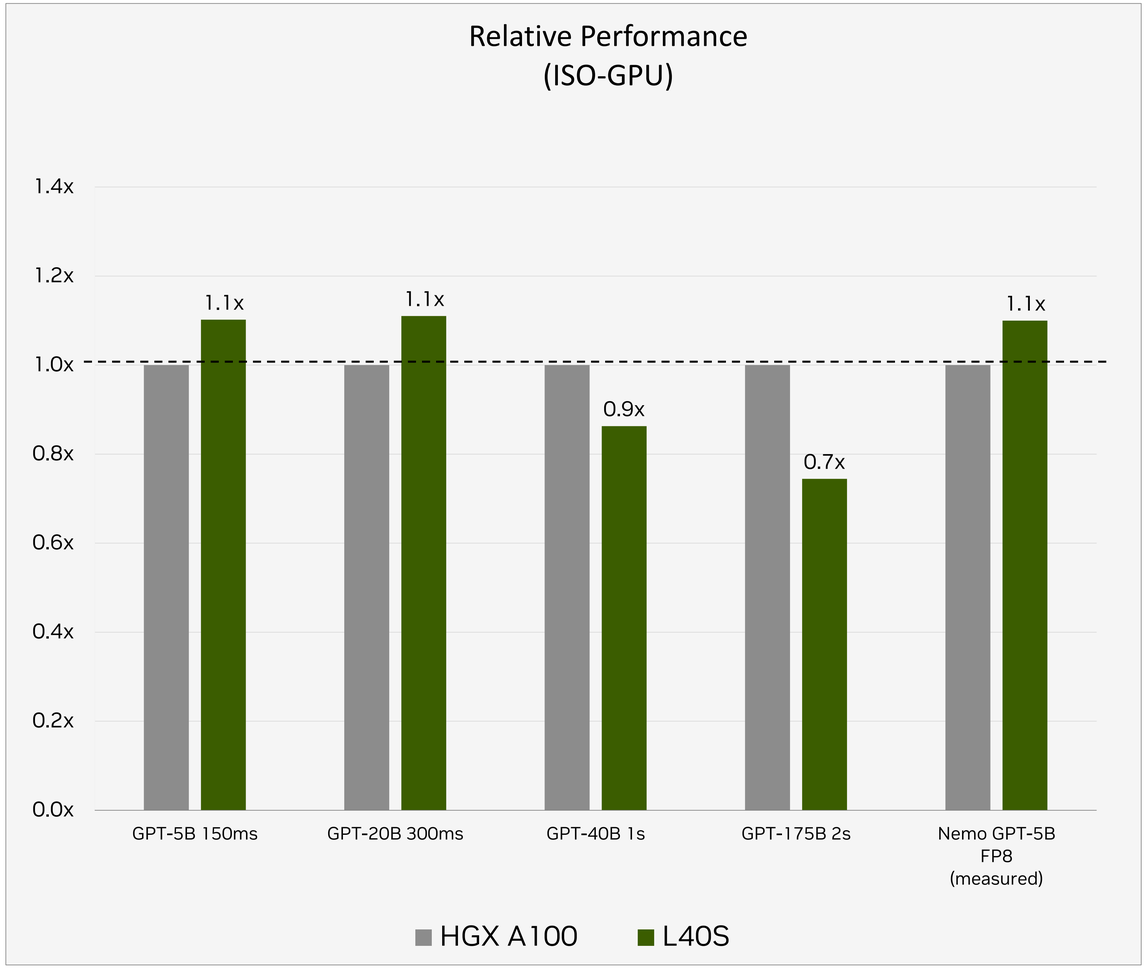
Across Variety of Training and Inference Workloads Found within MLPerf Benchmark

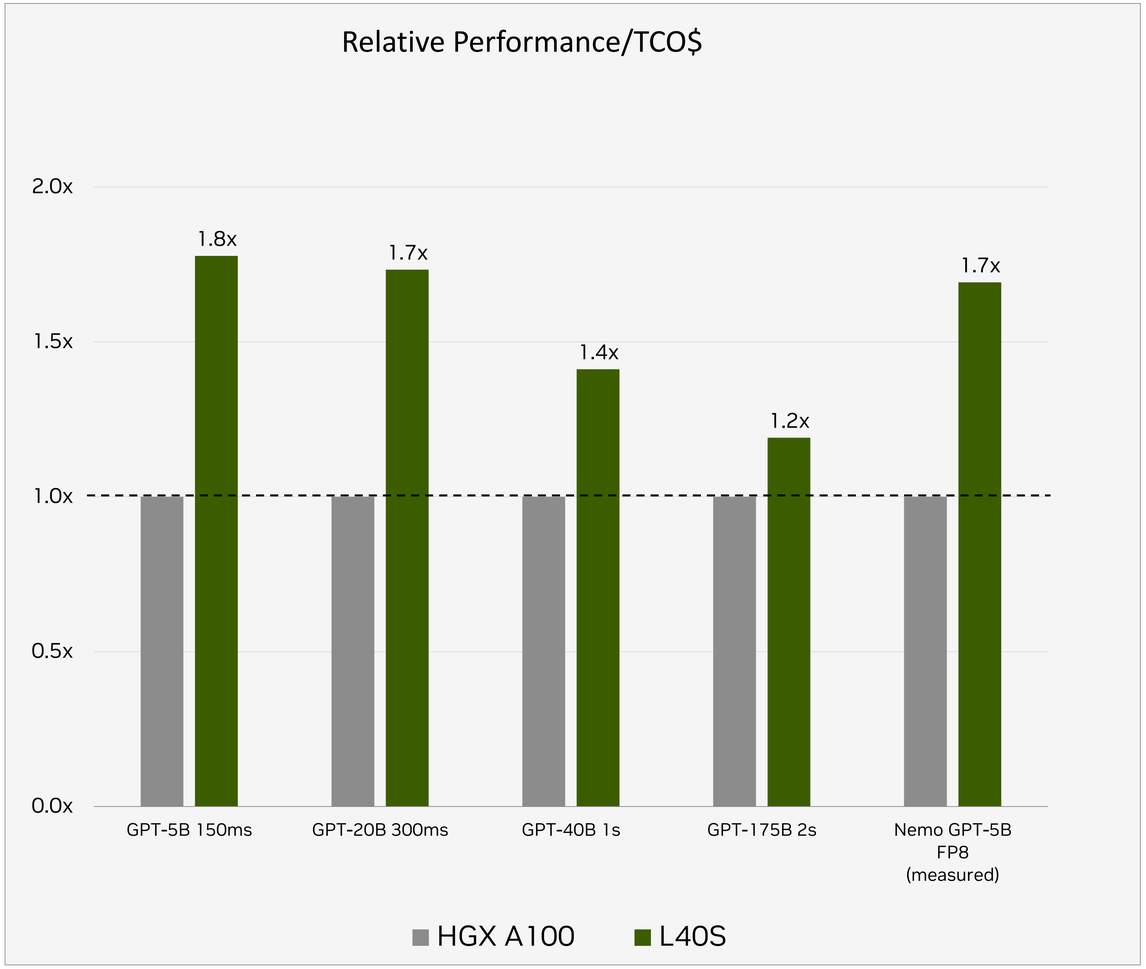


Preliminary performance projections, subject to change
1. Two systems with 4xL40S, vs HGX A100 8 GPU

L40S Delivers A100-Level Performance for LLM Inference

Variety of Sizes and Latency Targets: GPT3 5B-175B, FP8



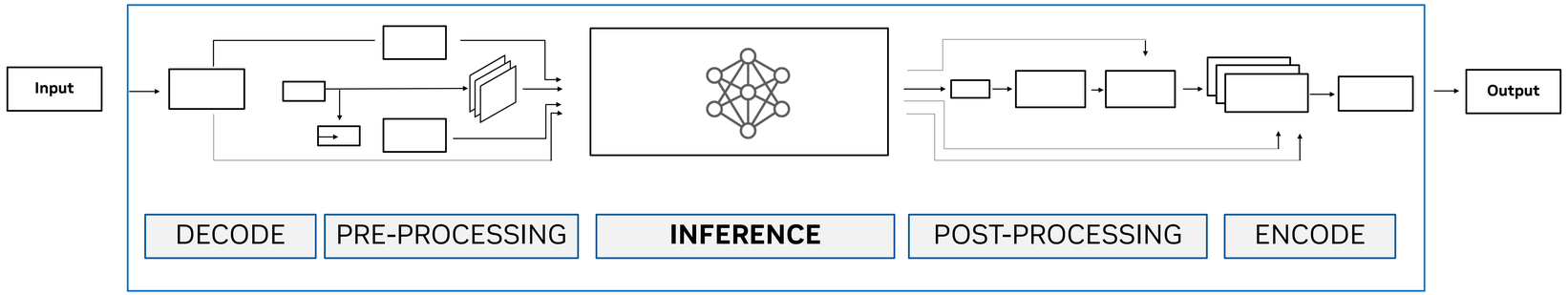


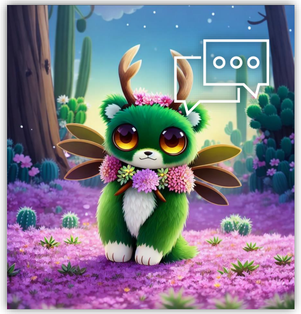
Preliminary performance projections, subject to change
1. Two systems with 4xL40S, vs HGX A100 8 GPU

Generative AI and Omniverse

Generative-AI Pipelines are Multimodal

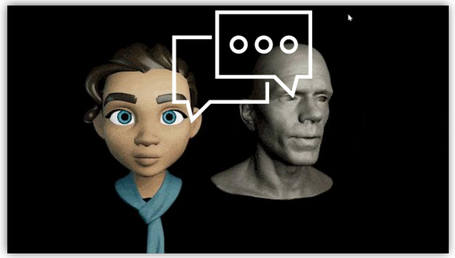
L40S Delivers Versatile Capabilities for End-to-End Acceleration





End-to-End Generative AI Pipeline

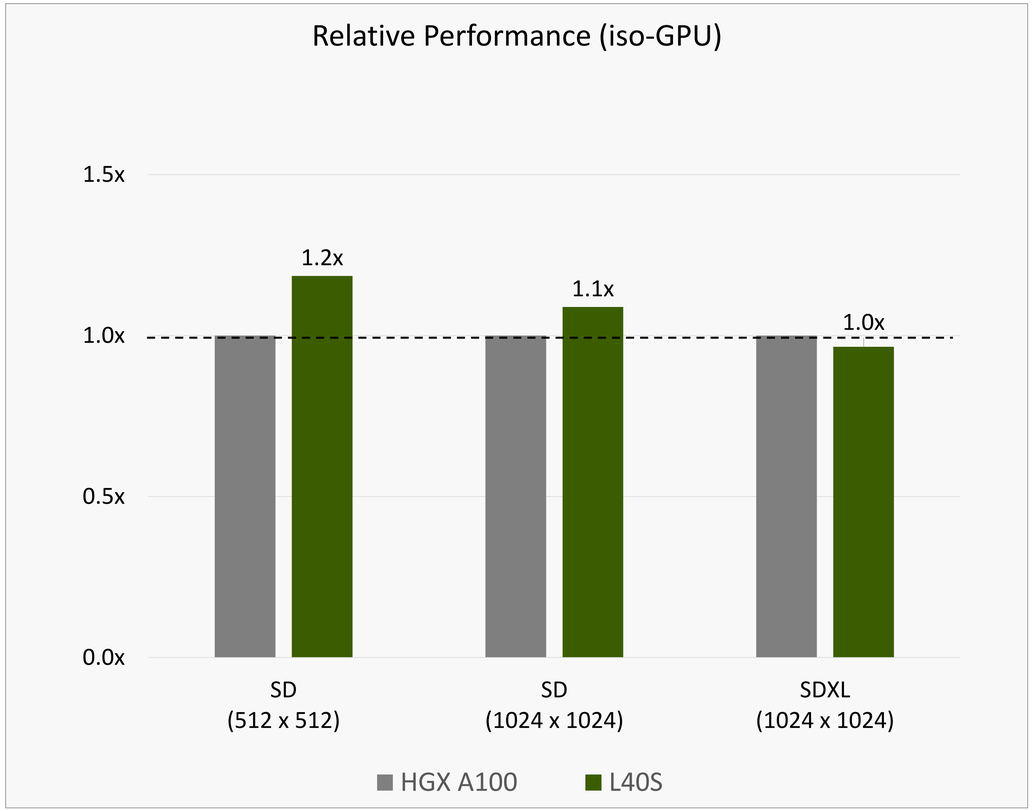
Speech Recognition	LLM	Text-to-Speech	Text-to-Image	Audio to Face	Rendering

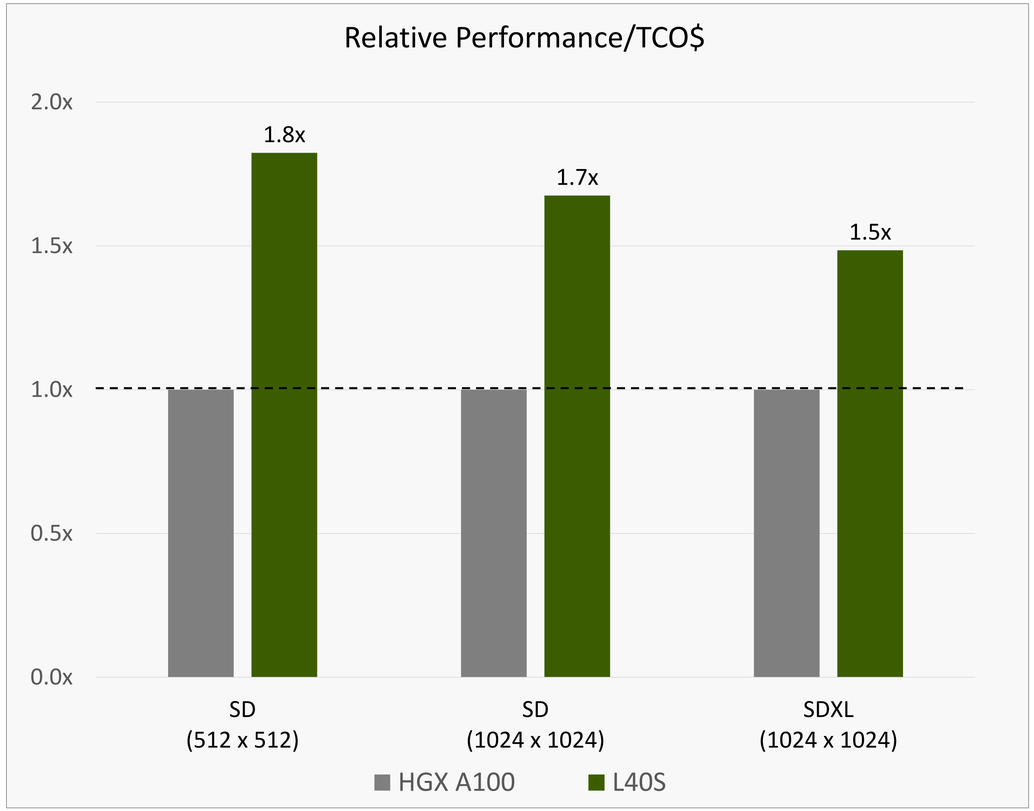


< 1s

L40S Delivers Better Performance vs A100 for Image Gen AI – FP16

Across Different Image Sizes and Resolutions





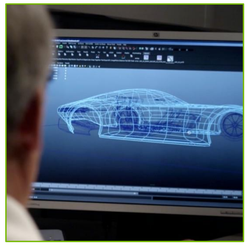
Measured performance; Iso-GPU
Two systems with 4xL40S vs HGX A100 8 GPU
Stable Diffusion v2.1, TRT 8.6.1, BS:1, FP16 | Stable Diffusion XL 1.0, TRT 8.6.1, BS:1, FP16

Omniverse Enterprise Unlocks Unified Digitalization

Computing Platform for End-to-end Industrial Digitalization, Digital Twin and Metaverse Applications

Full Fidelity Visualization

Unified Data Pipeline

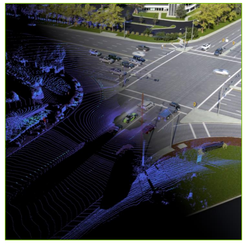


HIL/SIL Simulation

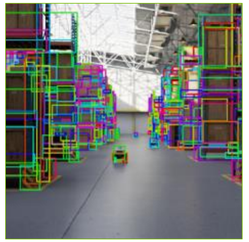
Factory Planning & Operations

Sustainable Content Creation Supply Chain

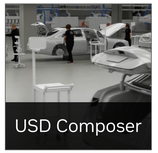
Autonomous Vehicle Testing & Validation



Synthetic Data Generation



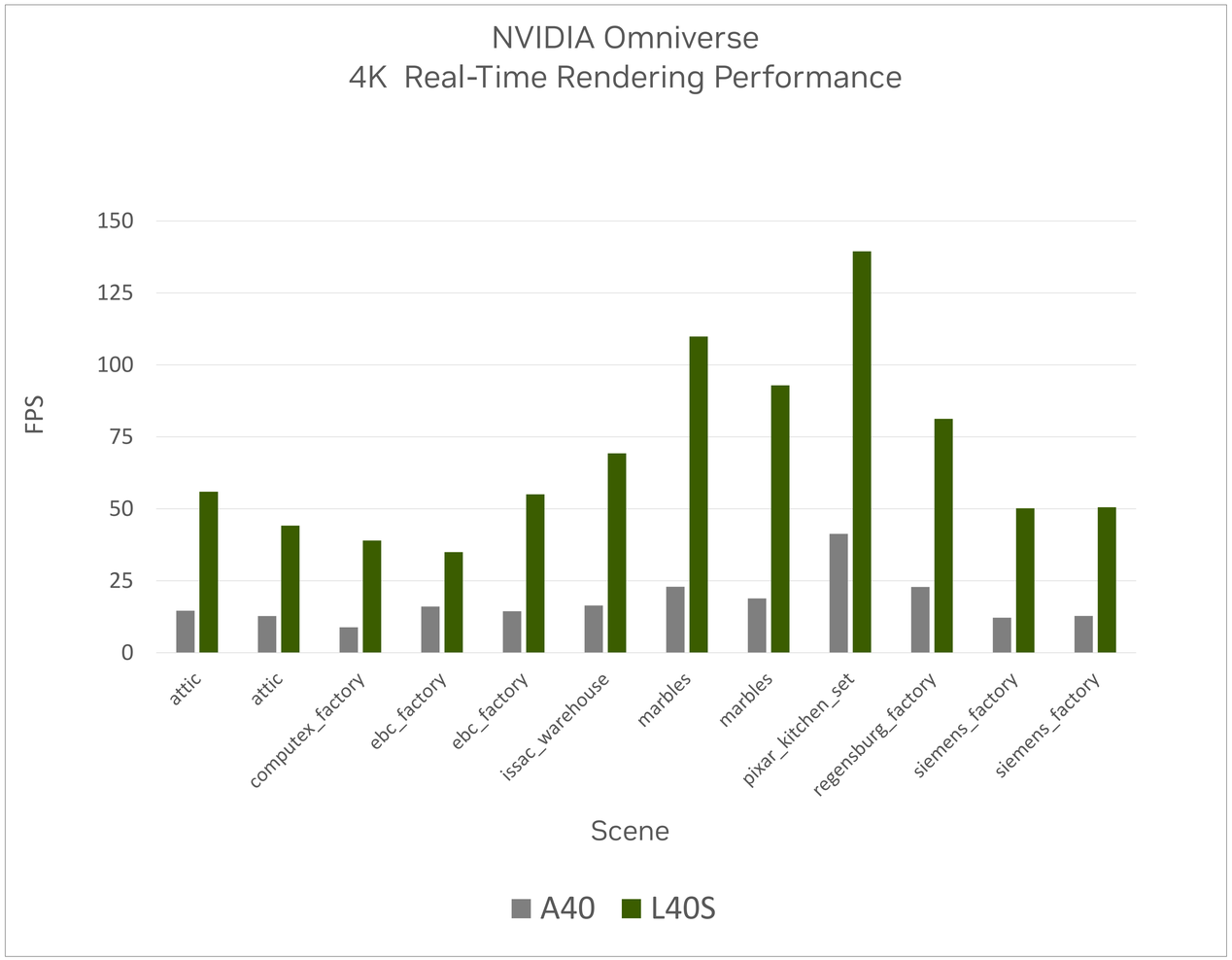
Framework Applications



NVIDIA OVX

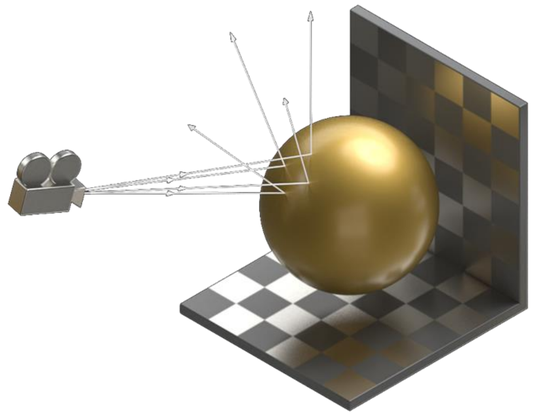
L40S Delivers Incredible Performance for Omniverse

Ada Lovelace delivers a multi-fold increase in performance



Powerful Capabilities for Visual Computing

- 3rd Gen RTX
- Real-time ray tracing w/ DLSS 3
- Powerful virtual workstation graphics
- Batch path tracing



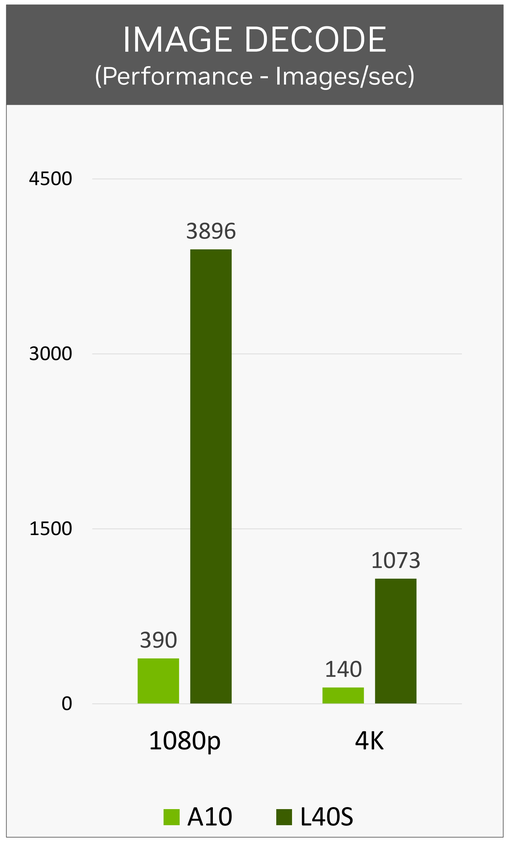
Preliminary performance projections, subject to change.
1. Omniverse RTX Renderer: NVIDIA A40, NVIDIA L40 as proxy for L40S, NVIDIA L40; DLSS3 enabled, Real time Render Mode, 4K Resolution

L40S VIDEO Performance

Modern Applications Seek High Performance for Interactive User Experience

		A10	L40S
NVENC	H.264	✓	✓
	H.265	✓	✓
	AV1	-	✓
NVDEC	H.264	✓	✓
	H.265	✓	✓
	AV1	✓	✓
	VP8	✓	✓
	VP9	✓	✓
	MPEG4	✓	✓
OFA		✓	✓
NVJPEG	Decode	-	✓

AV1 + NVJPEG
(NEW FEATURES)



10x more vs A10

Concurrent Video Streams

# of streams	ENCODE			DECODE			
	H.264	HEVC	AV1	H.264	HEVC	AV1	VP9
8K30	5	6	6	5	10	7	10
4K30	21	23	24	20	40	27	38
1080p30	81	88	94	70	141	82	127
720p30	171	179	184	156	285	131	309

Up to 4x more Encode-streams vs A10
Up to 2x more Decode-streams vs A10

Summary

Product Line Up Specifications Comparison

	L4	L40	L40S[1]	H100 NVL[3]
GPU Architecture	NVIDIA Ada Lovelace	NVIDIA Ada Lovelace	NVIDIA Ada Lovelace	NVIDIA Hopper
FP64	N/A	N/A	N/A	68 TFLOPS
FP32	30 TFLOPS	90.5 TFLOPS	91.6 TFLOPS	134 TFLOPS
RT Core	73.1 TFLOPS	209 TFLOPS	212 TFLOPS	N/A
TF32 Tensor Core[2]	121 TFLOPS	181 TFLOPS	366 TFLOPS	1,979 TFLOPS
FP16/BF16 Tensor Core[2]	242 TFLOPS	362 TFLOPS	733 TFLOPS	3,958 TFLOPS
FP8 Tensor Core[2]	484 TFLOPS	724 TFLOPS	1466 TFLOPS	7,916 TFLOPS
INT8 Tensor Core[2]	485 TOPS	724 TOPS	1466 TOPS	7,916 TOPS
GPU Memory	24 GB GDDR6 w/ ECC	48 GB GDDR6 w/ ECC	48 GB GDDR6 w/ ECC	188GB HBM3 w/ ECC
GPU Memory Bandwidth	300 GB/s	864 GB/s	864 GB/s	7.8TB/s[4]
L2 Cache	48 MB	96 MB	96 MB	100 MB
Media Engines	2 NVENC (+AV1) 4 NVDEC 4 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	14 NVDEC 14 NVJPEG
Power	Up to 72 W	Up to 300 W	Up to 350 W	2x 350-400 W
Form Factor	1-slot LP	2-slot FHFL	2-slot FHFL	2x 2-slot FHFL
Interconnect	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s	PCIe Gen5 x16: 128 GB/s
Availability	Shipping	Shipping	QS: Started, PS: Aug	Longer Leadtimes

1. Preliminary specifications, subject to change.
2. Specifications with sparsity.
3. Specifications shown for 2x H100 NVL PCIe cards paired with NVLink Bridge, also available in single PCIe card configuration option.
4. Aggregate 2x H100 NVL PCIe card bandwidth

NVIDIA L40S in the Data Center Portfolio

Specialized accelerators for a broad range of use cases

AI & COMPUTE WORKLOADS		GRAPHICS & GENERAL-PURPOSE WORKLOADS	
DL Training & DA	Language Processing	Graphics & Rendering	Mainstream Acceleration
DL Inference	Conversational AI	Omniverse	DL Inference
HPC	Recommenders	Virtual Desktops	Media Processing
Limited Availability, Long Lead Times			
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

NVIDIA L40S

The Most Powerful Universal Data Center GPU for AI and Graphics

NVIDIA L40S Availability and CTA

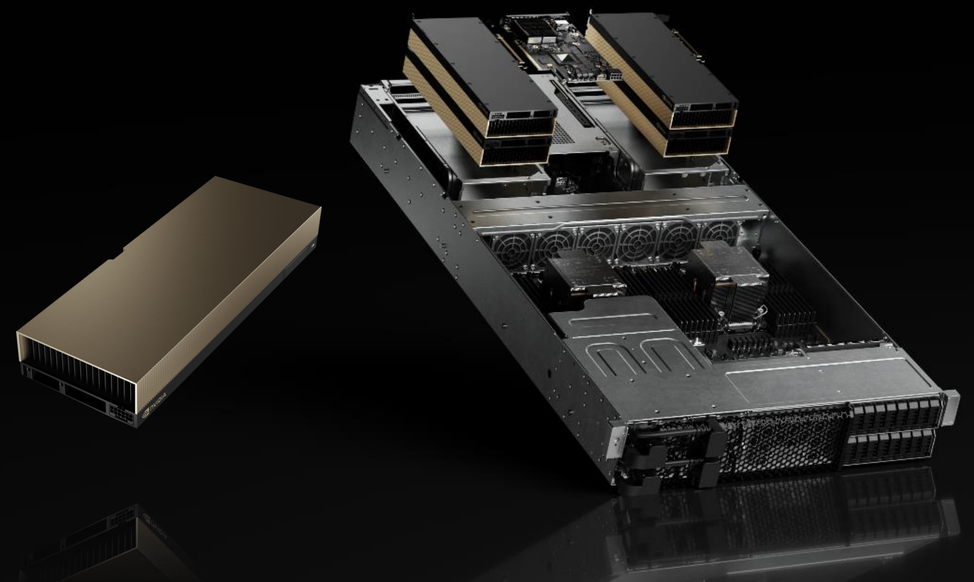
NVIDIA-Certified Systems with L40 and OVX available soon

1. **Identify customers looking to expand AI capacity**
   - For customers who cannot transition to H100, or cannot wait for H100, position L40S
   - Lead Times
     - HGX-H100- *Long*
     - H100 PCIe – *Longer*
     - A100 – *Longest*
2. **Brief key customers about the L40S**
3. **Position OVX systems for NVIDIA AI Enterprise and Omniverse use cases at scale**
4. **Share customer feedback with PM/PMM teams**

First availability in <u>September</u>

NVIDIA CERTIFIED

ASUS BOXX CoreWeave DELL Technologies GIGABYTE Hewlett Packard Enterprise Lenovo QCT SUPERMICRO



NVIDIA®