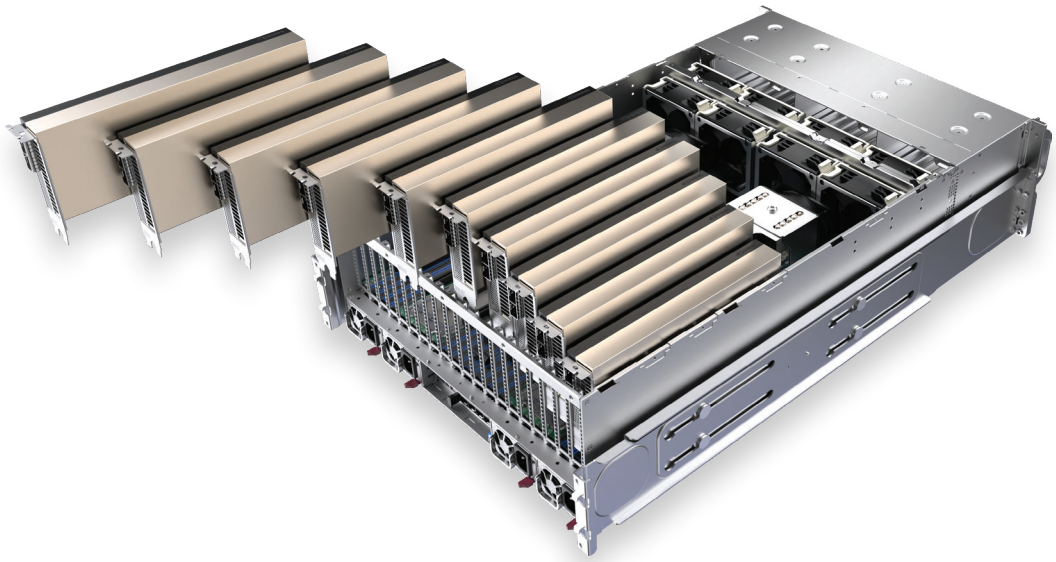
Order Supermicro NVIDIA L40S Systems Now!

With Better Availability and Performance per Dollar



Supermicro Systems with the latest NVIDIA L40S GPU, offer ample supply and drive breakthroughs in multi-workload acceleration for large language model (LLM) inference and training, graphics, and video applications. As the premier platform for multi-modal generative AI, Supermicro solutions with L40S GPUs, provide end-to-end acceleration for inference, training, graphics, and video workflows to power the next generation of AI-enabled audio, speech, 2D, video, and 3D applications.

Introducing NVIDIA L40S GPU

- The new Ada Lovelace Architecture features new Streaming Multiprocessor, 4th-Gen Tensor Cores, 3rd-Gen RT Cores, and 91.6 teraFLOPS FP32 performance.
- Experience the power of Generative AI, LLM Training, and Inference with features like Transformer Engine - FP8, over 1.5 petaFLOPS Tensor Performance*, and a Large L2 Cache.
- Unleash unparalleled 3D Graphics & Rendering capabilities with 212 teraFLOPS RT Core Performance, DLSS 3.0 for AI Frame Generation, and Shader Execution Reordering.
- Enhance Media Acceleration with 3 Encode & Decode Engines, 4 JPEG Decoders, and AV1 Encode & Decode Support.

Featured Products

SYS-421GE-TNRT/ SYS-521GE-TNRT
(Up to 10 L40S GPUs)

8U SuperBlade
(Up to 20 L40S in 8U)

2U Hyper SYS-221H-TNR
(Up to 4 L40S GPUs)

ARS-221GL-NR
(Up to 4 L40S GPUs)

2U CloudDC
(Up to 2 L40S GPUs)

2U Hyper-E
(Up to 3 L40S GPUs)

NVIDIA L40S Specifications Comparison

	NVIDIA L40S	NVIDIA HGX A100	NVIDIA H100 NVL
Best For	Universal GPU for Gen AI	Highest Perf Multi-Node AI	Generative AI performance
GPU Architecture	NVIDIA Ada Lovelace	NVIDIA Ampere	NVIDIA Hopper
FP64	N/A	9.7 TFLOPS	68 TFLOPS
FP32	91.6 TFLOPS	19.5 TFLOPS	134 TFLOPS
RT Core	212 TFLOPS	N/A	N/A
TF32 Tensor Core*	366 TFLOPS	312 TFLOPS	1,979 TFLOPS
FP16/BF16 Tensor Core*	733 TFLOPS	624 TFLOPS	3,958 TFLOPS
FP8 Tensor Core*	1466 TFLOPS	N/A	7,916 TFLOPS
INT8 Tensor Core*	1466 TOPS	1248 TOPS	7,916 TOPS
GPU Memory	48 GB GDDR6	80 GB HBM2e	188GB HBM3 w/ ECC
GPU Memory Bandwidth	864 GB/s	2039 GB/s	7.8TB/s
L2 Cache	96 MB	40 MB	100 MB
Media Engines	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	0 NVENC 5 NVDEC 5 NVJPEG	14 NVDEC 14 NVJPEG
Power	Up to 350 W	Up to 400 W	2x 350-400 W
Form Factor	2-slot FHFL	8-way HGX	2x 2-slot FHFL
Availability	QS: Started, PS: Aug	Longer Leadtime	Longer Leadtime

Go to https://learn-more.supermicro.com/L40S or scan the QR code to visit the Supermicro NVIDIA L40S Systems web page: