NVIDIA L40S and X13 GPU Platforms

Vuong Luong

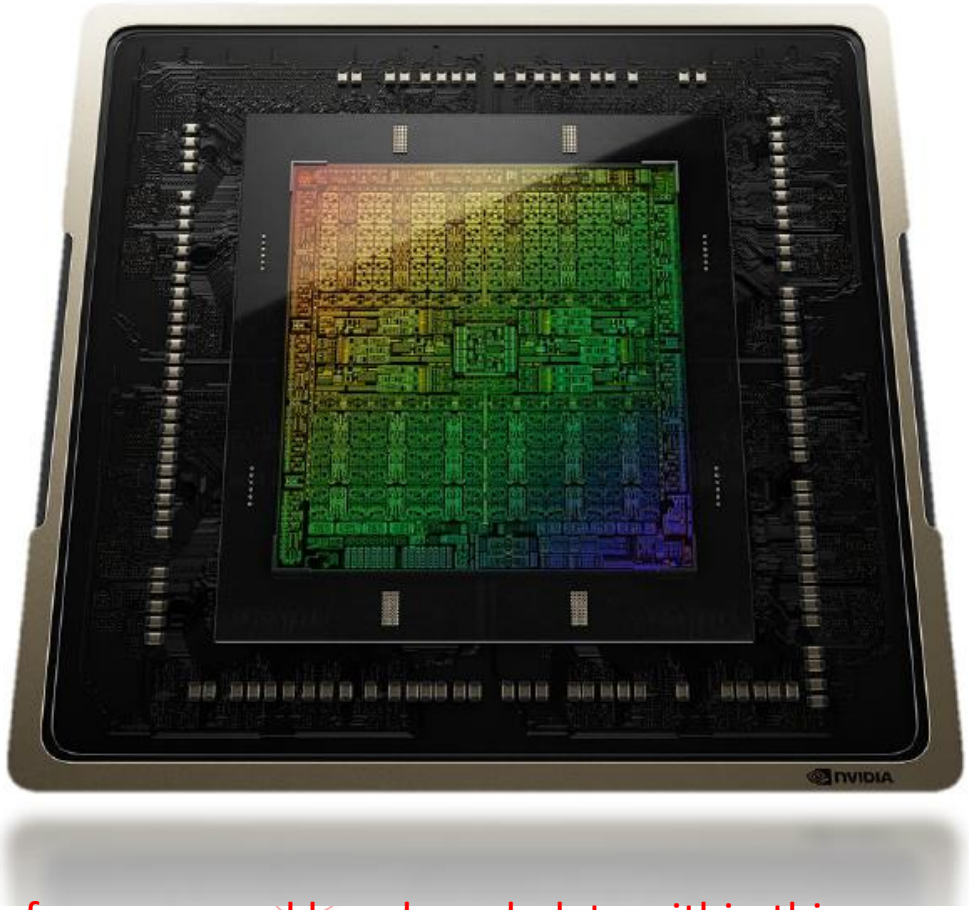
August 2023

Introducing NVIDIA L40S

The Most Powerful Universal GPU for AI and Graphics

Confidential

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

Performance and benchmark data within this presentation is *preliminary* and subject to change.

NVIDIA L40S

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

Fine Tuning LLM
4hrs
GPT-175B 860M Tokens[1]

LLM Inference
1.1X
Performance vs. HGX A100[2]

MLPerf Inference
1.1X
Performance vs. HGX A100[2]

GPT3 Training
<4 days
GPT-175 300B Tokens[3]

Image Gen AI
>82
Images per minute[4]

Full Video Pipeline
184
AV1 Encode Streams

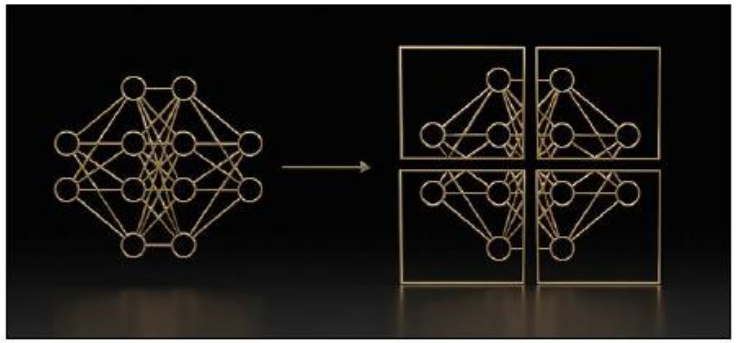
Preliminary performance specifications, subject to change

1. Retraining GPT-175B, 860M – 64 L40S GPUs
2. 8xL40S vs HGX A100, Projected MLPerf performance vs A100 submission- MLPerf inference v3.0, MLPerf Training v2.1
3. GPT 175B, 300B tokens, Foundational Training- 4K L40S GPUs
4. Image Generation, Stable Diffusion v2.1, 512 x 512 resolution
5. Concurrent Encoding Streams : 720P 30

Performance and benchmark data within this presentation is *preliminary* and subject to change.

Powerful Multi-Workload Acceleration

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



LLM Inference & Training

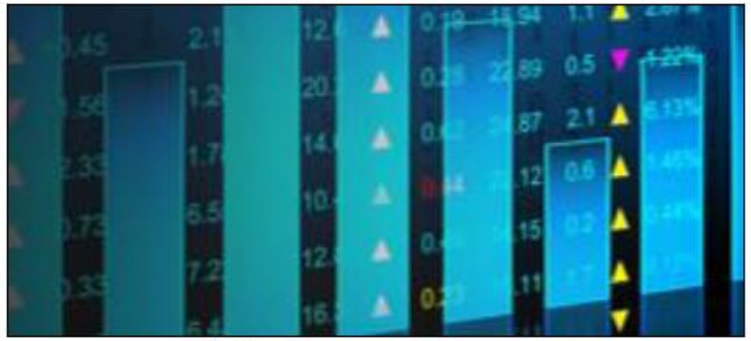
Accelerate AI training and inference workloads with 4th Gen Tensor Cores, Transformer Engine and support for FP8.

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 Delivers Higher Peak TFLOPS Than A100

Confidential

NVIDIA L40S Specifications

	NVIDIA L40S	NVIDIA HGX A100
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	QS: Started, PS: Aug	Longer Leadtime

3 Reasons To Transition from NVIDIA A100 to NVIDIA L40S

Superior Value and Availability

A100 Level Performance
+ Graphics and Video

Performance

1.2-2X Better Price-Performance than A100

Better Price-Performance

Fastest Time to Deployment

Shorter Lead Time

Intel DP 5U10 PCIE GEN5 GPU System: SYS-521GE-TNRT

Confidential

5U Up To 10x PCIE GEN5 GPU Intel® Sapphire Rapids Xeon® Scalable Processor System

- **Key Features**
  - Supports Up To 10x Double Wide GPUs 600W TDP
  - 13x PCIE GEN5 X16 Slots
  - 1x AIOM/OCP 3.0 Slot
  - Intel® Sapphire Rapids Xeon® Scalable Processor
  - Improved Thermal capability
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

CPU – Dual Socket Dual Intel® Sapphire Rapids Xeon® Scalable Processor Up to 56 Cores, CPU TDP up to 400W	**Memory –** 32 DIMM Slots 32 DIMMs Registered ECC DDR5 4800MHz SDRAM
Drives – 24 Hot-Swap Bays 8x 2.5" SATA 8x 2.5" U.2 NVMe SSD direct to CPU1 8x 2.5" U.2 NVMe SSD direct to CPU2 * 2x M.2 NVMe	**Expansion –** 13x PCI-E and 1x AIOM slot 6x PCI-E connect to CPU1/PLX 6x PCI-E connect to CPU2/PLX 1x PCI-E connect to CPU1 1x AIOM/OCP 3.0 connect to CPU2
Networking – Dual 10GbE 2x RJ45 10GbE 1x RJ45 1GbE IPMI	**Power Supply –** N+N Redundant 4x 2700W Titanium Level

* GPUDirect Storage in development

Intel DP 4U10 PCIE GEN5 GPU System: SYS-421GE-TNRT

Confidential

4U Up To 10x PCIE GEN5 GPU Intel® Sapphire Rapids Xeon® Scalable Processor System

- **Key Features**
  - Supports Up To 10x Double Wide GPUs 350W TDP
  - 13x PCIE GEN5 X16 Slots
  - 1x AIOM/OCP 3.0 Slot
  - Intel® Sapphire Rapids Xeon® Scalable Processor
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

CPU – Dual Socket Dual Intel® Sapphire Rapids Xeon® Scalable Processor Up to 56 Cores, CPU TDP up to 400W	**Memory –** 32 DIMM Slots 32 DIMMs Registered ECC DDR5 4800MHz SDRAM
Drives – 24 Hot-Swap Bays 8x 2.5" SATA 8x 2.5" U.2 NVMe SSD direct to CPU1 8x 2.5" U.2 NVMe SSD direct to CPU2 * 2x M.2 NVMe	**Expansion –** 13x PCI-E and 1x AIOM slot 6x PCI-E connect to CPU1/PLX 6x PCI-E connect to CPU2/PLX 1x PCI-E connect to CPU1 1x AIOM/OCP 3.0 connect to CPU2
Networking – Dual 10GbE 2x RJ45 10GbE 1x RJ45 1GbE IPMI	**Power Supply –** N+N Redundant 4x 2700W Titanium Level

* GPUDirect Storage in development

Intel DP 4U8 PCIE GEN5 GPU System: SYS-421GE-TNRT3

Confidential

4U Direct Connect 8x PCIE GEN5 GPU Intel® Sapphire Rapids Xeon® Scalable Processor System

- **Key Features**
  - Supports Up To 8x Double Wide GPUs 350W TDP
  - 8x PCIE GEN5 X16 Slots DIRECT CONNECT to CPUs
  - 1x AIOM/OCP 3.0 Slot
  - Intel® Sapphire Rapids Xeon® Scalable Processor
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

CPU – Dual Socket Dual Intel® Sapphire Rapids Xeon® Scalable Processor Up to 56 Cores, CPU TDP up to 400W	**Memory –** 32x DIMM Slots 32x DIMMs Registered ECC DDR5 4800MHz SDRAM
Drives – 24x Hot-Swap Bays and 2x M.2 8x 2.5" SATA SSD/HDD 4x 2.5" U.2 NVMe SSD direct to CPU 2x M.2 NVMe	**Expansion –** 8x PCI-E and 1x AIOM slot 4x PCI-E GEN5 X16 Direct Connect to CPU1 4x PCI-E GEN5 X16 Direct Connect to CPU2 1x AIOM/OCP 3.0 connect to CPU2
Networking – Dual 10GbE 2x RJ45 10GbE 1x RJ45 1GbE IPMI	**Power Supply –** N+N Redundant 4x 2700W Titanium Level

Next Gen 4U 4 GEN 5 GPU SystemSYS-741GE-TNRT

Confidential

Dual CPUs and 4 PCIe Gen5 GPUs

System Front View

System Rear View

Key Features

- Supports up to 4 Double Width GPUs
- Dual CPUs up to 350W TDP

Key Applications

- AI Compute/Model Training/Deep Learning, HPC
- Real-Time High Quality Multi-GPU Ray Tracing
- High Performance Simulation of Complex 3D Graphics

80 PLUS TITANIUM

Specifications

CPU – Dual Socket Dual Sapphire Rapids CPU (up to 350W TDP)	**Memory –**DIMM Slots 16x DIMM slots, ECC DDR5 Designed for up to 4800MT/s
Drives – 8 Hot-Swap Bays 8x 2.5" NVMe U.2 or 8x HS 3.5" SATA/SAS	**Expansion –** 7 PCIe Slots 7x PCIe 5.0 x16 (4 FHFL/DW & 3 FH)
I/O ports 2x RJ45 10GbE 1x RJ45 1GbE IPMI 1x VGA, 7x USB 3.01x COM Header	**Power Supply –** N+N Redundant 2x 2000W Titanium Level Efficiency Power Supplies 2x 2600W Titanium Level Efficiency Power Supplies (option)

Workload to GPU Mapping Best perf/$

Confidential

Multi-Workload (graphics, video, compute)	Best compute perf/$	L40**S**
	Best graphics perf/$	L40
	Best video perf/$	L4
Large Scale Training	Fastest time to solution	HGX H100
Small/Mid Scale Training	Best perf/$	8x L40**S**

H13 Servers with L40S

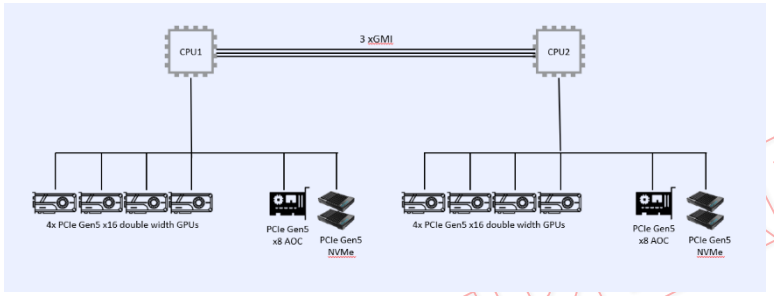
With All New AMD EPYC™ 9004 Series Processor

H13 - 4U GPU Servers

Confidential

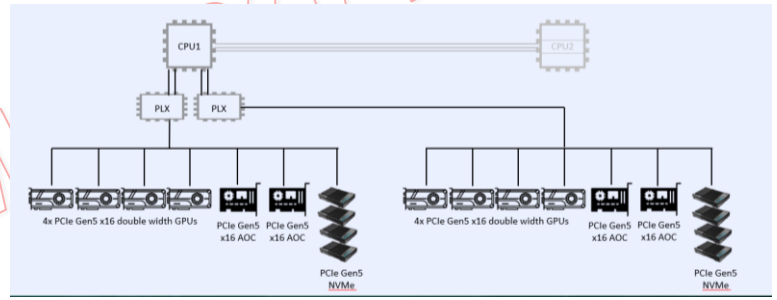
Direct-connect, GPU-direct, RDMA and Ultra-low Latency

Up to 10x L40S GPU !!



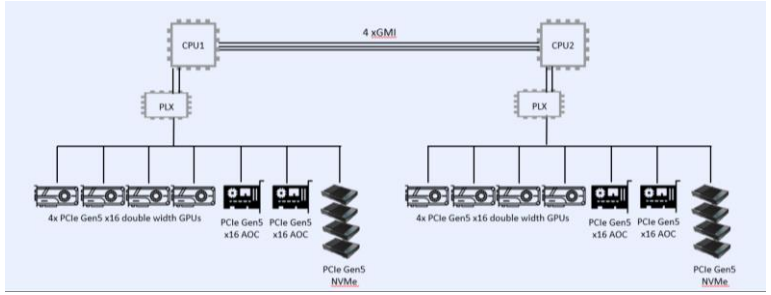
AS -4125GS-TNRT

Dual Root, Direct-connect



AS -4125GS-TNRT1

Single Root – Dual PLX Partitions



AS -4125GS-TNRT2

Dual Root – Dual Partitions

	4U 8GPU	4U 10GPU	4U 10GPU
Model Number	AS -4125GS-TNRT	AS -4125GS-TNRT1	AS -4125GS-TNRT2
CPU	Dual SP5 socket for AMD EPYC™ Series Processor, Up to **128 cores**	Single SP5 socket for AMD EPYC™ Series Processor, Up to **128 cores**	Dual SP5 socket for AMD EPYC™ Series Processor, Up to **128 cores**
GPU Support	NVIDIA® A100, H100, AMD Instinct™ MI210 PCIe Optional NVIDIA NVLink™ Bridge, AMD Infinity Fabric™ Link for GPU-to-GPU connectivity		
GPU Quantity	Up to 8x PCIe GEN5 FHFL	Up to 10x PCIe GEN5 FHFL	Up to 10x PCIe GEN5 FHFL
Memory	24 DIMM slots; Up to 6TB DDR5-4800	12 DIMM slots; Up to 3TB DDR5-4800	24 DIMM slots; Up to 6TB DDR5-4800
Expansion	8 PCIe 5.0 x16 slots for double-width GPU accelerators 1 PCIe 5.0 x16 or 2 x8 slots	Up to 10 PCIe 5.0 x16 slots for double-width GPU accelerators 1 PCIe 5.0 x16 slot	Up to 10 PCIe 5.0 x16 slots for double-width GPU accelerators up to 2x PCIe 5.0 x16 slots
Storage	Up to 4 hot-swap 2.5" NVMe drives 2x 2.5" hot-swap SATA drives 1 M.2 NVMe slot for boot drive	Up to 8 hot-swap 2.5" NVMe drives 2x 2.5" hot-swap SATA drives 1 M.2 NVMe slot for boot drive	Up to 8 hot-swap 2.5" NVMe drives 2x 2.5" hot-swap SATA drives 1 M.2 NVMe slot for boot drive
Networking	Up to 2x 10GbE BaseT	Up to 2x 10GbE BaseT	Up to 2x 10GbE BaseT
Power Supplies	4x Redundant 2000W Titanium level	4x Redundant 2000W Titanium level	4x Redundant 2000W Titanium level
Workloads	AI, Deep Learning, 3D Simulation, Cloud Gaming, HPC, Media/ Video Streaming, Research		

Hyper AS -1115HS-TNR

AMD EPYC Genoa ***Single processor***

Two Gen3 ***M.2***

128 cores

80 PLUS TITANIUM ***1200W*** Redundant Titanium

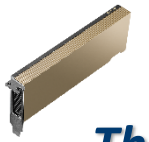
96%+ Power Efficiency

12-Channel ***DDR5-4800*** MHz

TWO DIMM per channel ***6TB***

Supports 1x L40S!

NVIDIA Mellanox TECHNOLOGIES



PCI EXPRESS 5.0 ***Up to Three PCIe Gen5 AOC***

PCIe Gen5 AIOMs

Flexible Networking Options*

1G, 10G, 25G, 50G, 100G, 200G, 400G

SERIAL ATA

nvm EXPRESS

Eight 2.5" Hot-Swap ***Optional*** *Extend up to* ***Twelve***

2.5"drive ***NVMe /SAS/SATA(290W CPU)***

Certain CPUs with high TDP may be supported only under specific conditions. Please contact Supermicro Technical Support for additional information about specialized system optimization

Hyper AS -2015HS-TNR

AMD EPYC Genoa **Single processor**

Two Gen3 **M.2**

128 cores

Up to 4x L40S!*

GPU Support **Upgrade Fan** & **2.5”**

1200W Redundant Titanium

96%+ Power Efficiency

12-Channel **DDR5-4800** MHz

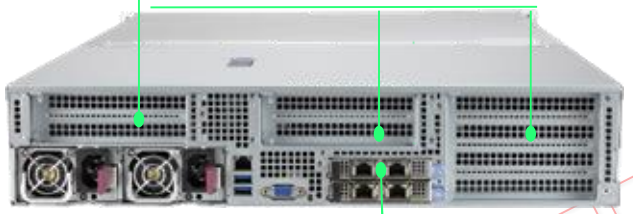
TWO DIMM per channel **6TB**

NVIDIA

Mellanox TECHNOLOGIES

PCI EXPRESS 5.0

Up to **Eight** PCIe **Gen5 AOC**



PCIe Gen5 AIOMs

Flexible Networking Options*

1G, 10G, 25G, 50G, 100G, 200G, 400G

SERIAL ATA

nvm EXPRESS

Twelve 3.5” Hot-Swap

NVMe /SAS/SATA

*Thermal testing in process, support conditions may change

*Certain CPUs with high TDP may be supported only under specific conditions. Please contact Supermicro Technical Support for additional information about specialized system optimization.

Hyper AS -2025HS-TNR

AMD EPYC Genoa ***Dual processor***

128 cores

Two Gen3 ***M.2***

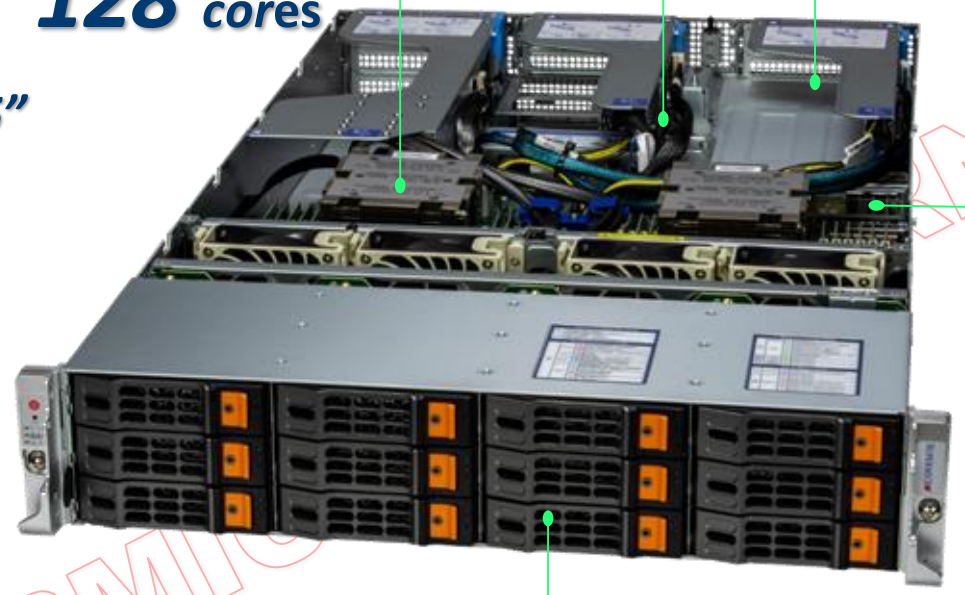
80 PLUS TITANIUM

1600W Redundant Titanium

96%+ Power Efficiency

Up to 2x L40S!

GPU Support ***Upgrade Fan*** & ***2.5”***



24-Channel ***DDR5-4800*** MHz

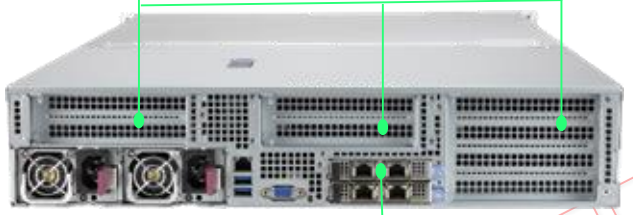
One DIMM per channel

NVIDIA

Mellanox TECHNOLOGIES

PCI EXPRESS 5.0

Up to ***Eight*** PCIe ***Gen5 AOC***



PCIe ***Gen5 AIOMs***

SERIAL ATA

nvm EXPRESS

Twelve 3.5” Hot-Swap ***NVMe /SAS/SATA***

3.5” do not support GPU

Flexible Networking Options*

1G, 10G, 25G, 50G, 100G, 200G, 400G

Certain CPUs with high TDP may be supported only under specific conditions. Please contact Supermicro Technical Support for additional information about specialized system optimization

CloudDC AS -2015CS-TNR

Single AMD EPYC™ 9004 Processor

128 cores

360W (cTDP 400W) *

860W Redundant

Titanium **96%** Power Efficiency

80 PLUS TITANIUM

TWO Gen3 **M.2**

12 -Channel **DDR5-4800** MHz

One DIMM per channel

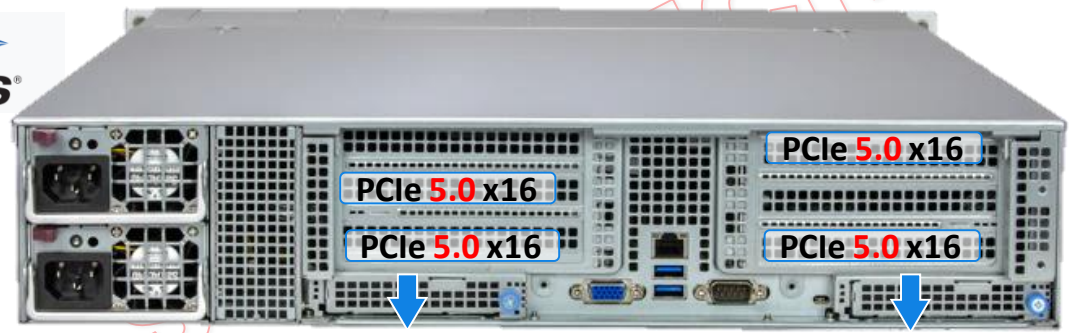
12 3.5" Hot-Swap **SATA Default** or

4 2.5" Hot-Swap **NVMe** option

Up to 4+2 PCI-E Gen5

2x AIOM + 4 FHHL (or 2 Double-width)

PCI EXPRESS

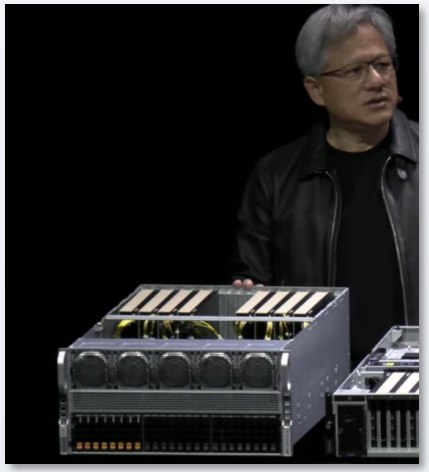


Up to 2x L40S!

Certain CPUs with high TDP may be supported only under specific conditions. Please contact Supermicro Technical Support for additional information about specialized system optimization

NVIDIA L40S Announcement (Today, Aug. 8th)

Promotions



NVIDIA Jensen Huang has announced L40S at SIGGRAPH Keynote
w/ Supermicro's 8-10 GPU System
SYS-521GE-TNRT

Promotion (Website, Social Media, Newsletters, Search/Display Ads, etc.)

Sales Assets & Trainings

Sales Assets (GPU Brochure, One-pager, email template, etc.)

Sales Lunch Training Channel Regional Training (in US, EMEA and APAC)

GPU Brochure – Large Language Models to the AI Edge

3

Enterprise AI Inferencing & Training

Generative AI Inference, Large Language Model Inference, Speech Recognition, Recommendation, Computer Vision

Workload Sizes

Extra Large

4U/5U 8-10 GPU PCIe
GPU-based Inference and Training

Large

Medium

SuperBlade®
High Density, Disaggregated

2U MGX System
Modular Building Block Platform Supporting Today's and Future GPUs, CPUs, and DPUs

2U Grace MGX System
(Codenamed: C2)
Modular Building Block Platform with Energy-efficient Grace CPU Superchip

16

Enterprise AI Inferencing & Training

Use Cases

- Content creation (image, audio, video, writing)
- AI-enabled office applications and services
- Enterprise business process automation

Opportunities and Challenges

- Total solution complexity
- Open architecture
- Vendor flexibility (CPU & GPU)
- GPU-based training and inference

Key Technologies

- NVIDIA H100 (NVL, PCIe), A100, L40S, L40, and L4 GPUs
- PCIe 5.0 storage and networking
- Intel and AMD CPU options
- NVIDIA Grace™ Superchip (2 Grace CPUs on one Superchip) with NVLink® Chip-2-Chip (C2C) interconnect
- Flexible rackmount servers from 1U to 6U to balance compute, storage, and networking for various enterprise AI workload needs

Solution Stack

- NVIDIA AI Enterprise software
- NVIDIA NGC™ catalog: containers, pre-trained models
- RedHat OpenShift, VMWare

H100 NVL
2 FHFW H100 GPU
with NVLink Bridge (4x faster than PCIe)
PCIe 5.0
400W per GPU
94GB HBM3 per GPU

H100 PCIE
FHFL DW PCIe 5.0 x16
300W per GPU
80GB HBM2e

L40S\L40
FHFL DW
PCIe 4.0 x16
350W (L40S)/300W (L40)
48GB GDDR6

L4
HHHL SW
PCIe 4.0 x16
72W

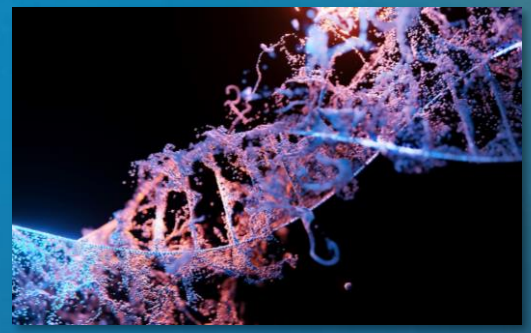
17

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

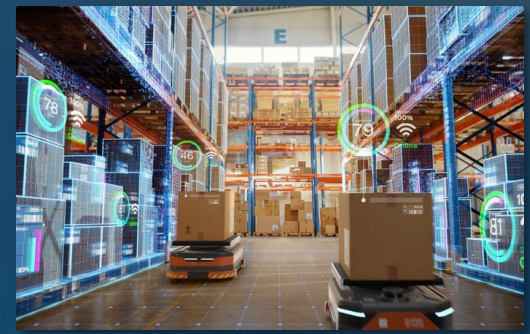
AI-enabled services/applications, chatbots, business automation

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

NVIDIA GPU Map and Supermicro Compatibility

Supermicro System GPU Compatibility

	H100 (SXM)	H100 (NVL)	H100 (PCIe)	L40S	L40	L4	RTX 6000 Ada
4U/5U/8U GPU	4 (4U/5U) 8 (8U)						
4U/5U 10-GPU		10 (4U/5U)	10 (4U/5U)	10 (4U/5U)	10 (4U/5U)	10 (4U/5U)	8 (4U/5U)
SuperBlade		20 (8U) 10 (6U)	20 (8U) 10 (6U)	20 (8U) 10 (6U)	20 (8U) 10 (6U)	40 (8U) 20 (6U)	
BigTwin			4 (2U2N)	4 (2U2N)	4 (2U2N)	4 (2U2N) 8 (2U4N)	2 (2U)
CloudDC			2 (2U)	2 (2U)	2 (2U)	4 (2U) 2 (1U)	
Hyper			4 (2U) 1 (1U)	4 (2U) 1 (1U)	4 (2U) 1 (1U)	4 (2U) 2 (1U)	
WIO			2 (2U) 1 (1U)	2 (2U) 1 (1U)	2 (2U) 1 (1U)	4 (2U) 2 (1U)	
Hyper-E			3	3	3	4	
Short-Depth Edge						2	
Compact Edge/IoT						3	
Workstation		4	4	4	6	4	

Download from Marketing Portal

Visualization and Omniverse Workloads

Omniverse Optimized Systems

Highest Performance, Tailored for NVIDIA Omniverse

Benefits & Advantages

- New next-generation purpose-built system for NVIDIA Omniverse™ Enterprise
- Optimized for power immersive, photorealistic 3D models, simulations, and digital twins
- Flexible storage configurations
- Up to 2x more storage and I/O flexibility

4U/SU8 GPU (PCIe)
8 NVIDIA L40S/L40 PCIe
3 NVIDIA ConnectX-7
16 U.2 NVMe drives
SYS-421GE-TNRT / AS -4125GS-TNRT / SYS-521GE-TNRT

Key Features

- 8 NVIDIA L40S/L40 PCIe GPUs
- 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+
- 3 NVIDIA ConnectX-7
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling.
- 16 U.2 NVMe drive bays

24

Visualization and Omniverse Workloads

2U Hyper Systems

Hyper - Flagship Performance Rackmount System
Designed for Ulimate Flexibility

Benefits & Advantages

- Highly flexible modular architecture
- Compute optimized design for maximum airflow
- Maximum availbility of PCIe lanes for GPUs and networking
- Tool-less platform for ease of configuration and servicing

2U Hyper
4 NVIDIA L40 PCIe
8 NVMe drives
32 DIMMs DDR5-4800
SYS-221H-TNR / AS -2115HS-TNR

Key Features

- Up to 4 NVIDIA L40S/L40 GPUs
- Dual 4th Gen Intel® Xeon® Scalable processors or AMD EPYC™ 9004 Series processors
- Optimized thermal capacity and airflow to support CPUs up to 350W with GPUs up to 350W with air cooling
- Supports PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1+
- Advanced I/O Module (AIOM) for flexible networking options - OCP 3.0 SFF compatible

25

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

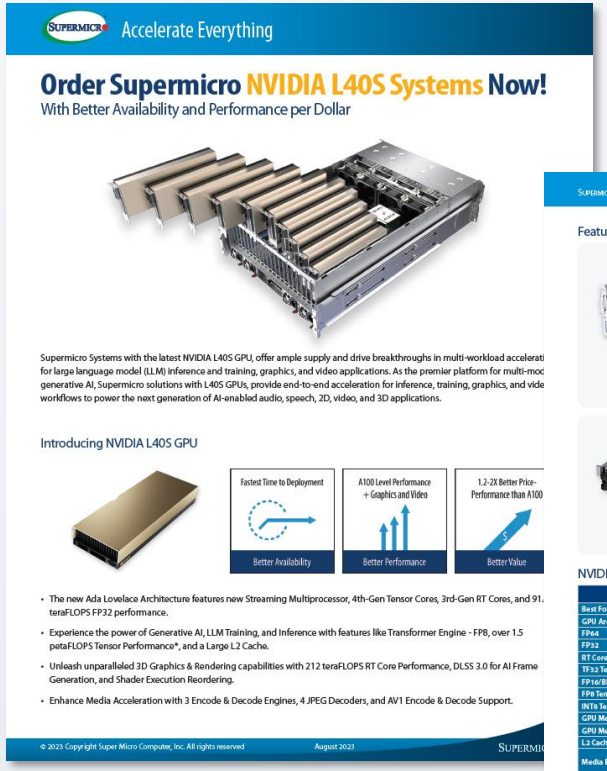
NVIDIA L40S Supported Supermicro Systems

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, AMD PCIe: Instinct MI210
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, AMD PCIe: Instinct MI210
AS -4125GS-TNRT2	NVIDIA PCIe: H100, H100 NVL, L40S, L40, A100, AMD PCIe: Instinct MI210
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, A101, AMD PCIe: Instinct MI210
AS -2025HS-TNR	NVIDIA PCIe: H100, L40S, L40, A100, AMD PCIe: Instinct MI210
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, AMD PCIe: Instinct MI210

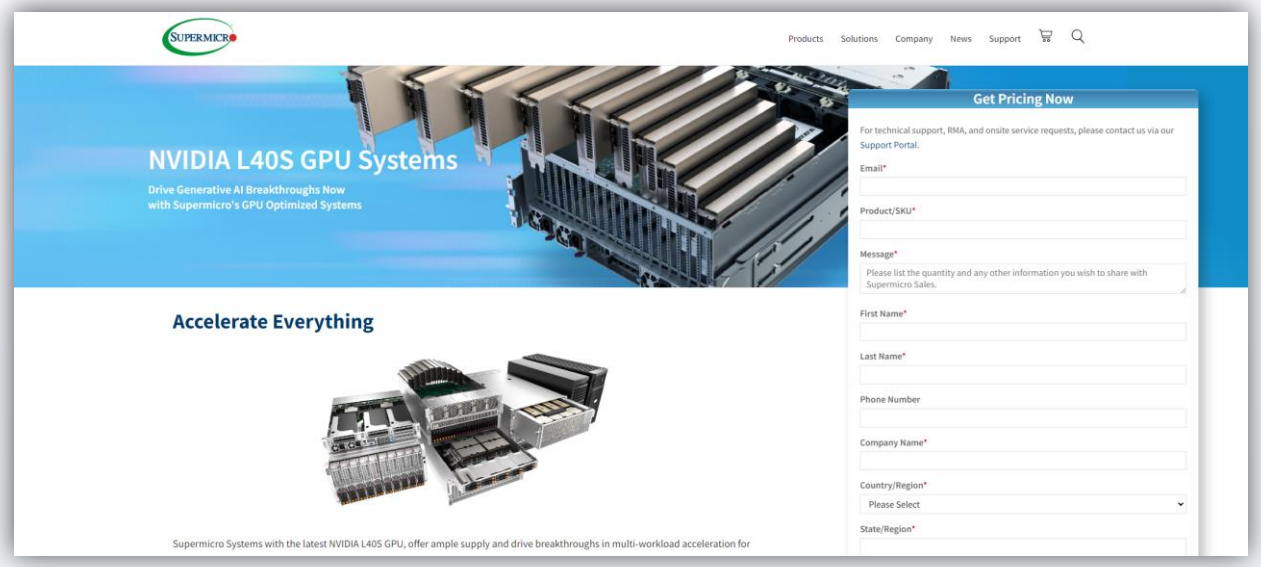
Updated all spec pages with L40S support

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)

Email Template and Flyer to Send Out to Your Customers



Supermicro L40S Launch Landing Page



- **Most Powerful Universal GPU for the Data Center**
- **Great for LLM Inference & Training, Graphics and Video Applications**

Improved Performance per Dollar and Availability

- 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.

Portal.Supermicro.com/Marketing

NEW! L40S Launch:

- AI GPU Brochure
- NVIDIA Solution Page
- L40S Launch Landing Page
- L40S Product Flyer
- L40S Sales Training Deck
- L40S Launch Customer Email Template
- L40S Webinar

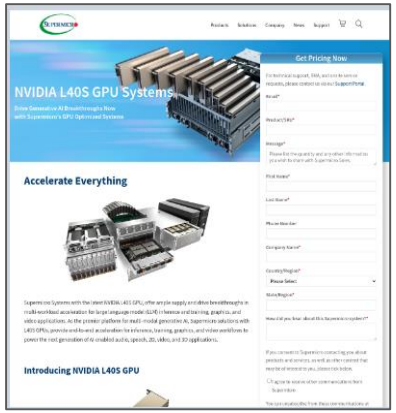
L40S Sales Assets

Tradeshows and Events
Find our events or request support for regional activities.

Marketing Collateral
The latest brochures, advertisements, eblasts and creative resources.

Sales Tools
Corporate templates, gifting options, COOP and logo guidelines and more.

New Product Info and Programs
New Product NDA Materials, Seeding Programs, JumpStart, ...



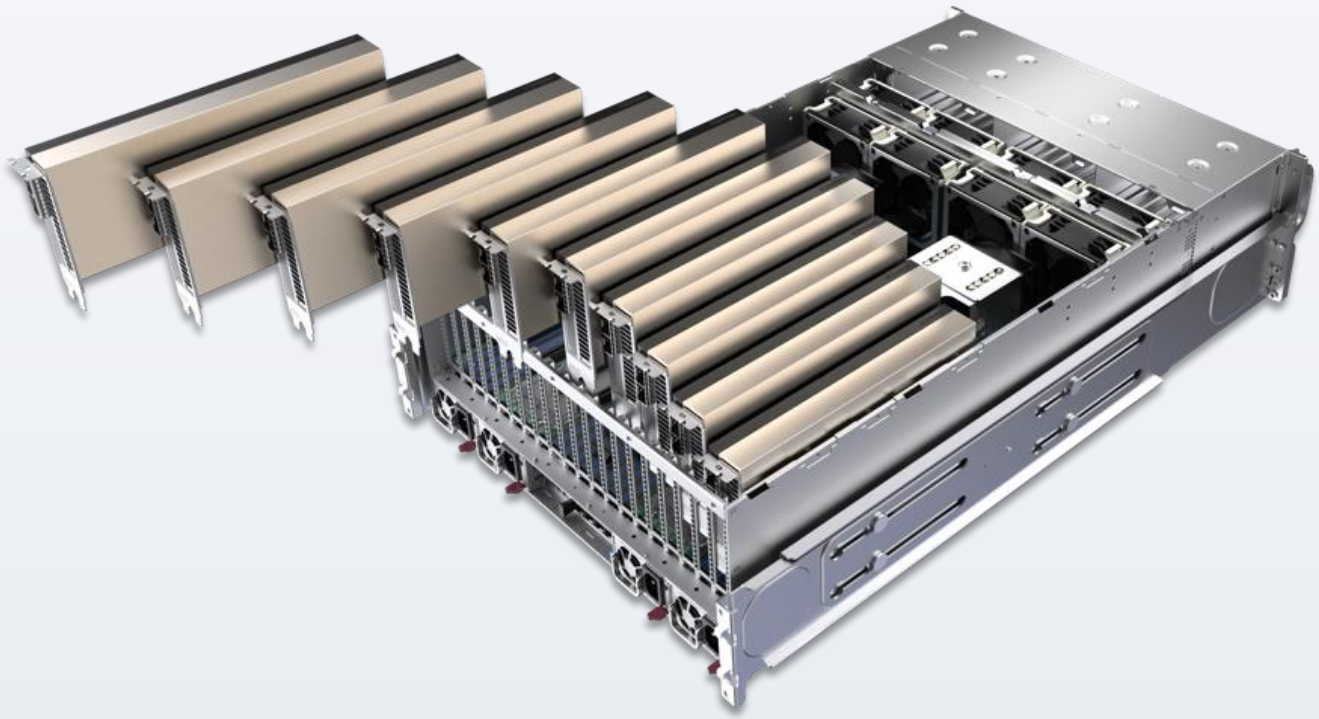
L40S Landing Page

Customer Email Template

Marketing_Portal > Marketing_Collateral > L40S Launch

Pre-order Now!

- Part Number: **GPU-NVL40S**
- Landing page: https://learn-more.supermicro.com/l40s
- Marketing Portal for assets download
  - Email template
  - GPU brochure
  - Datasheet
  - And more



©2023 Supermicro 29

DISCLAIMER

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

Thank You

www.supermicro.com