Workstation for the Next Generative AI Solution

Solution Manager – Alan CL Huang
Product Management Team
alanhuang@supermicro.com

Agenda

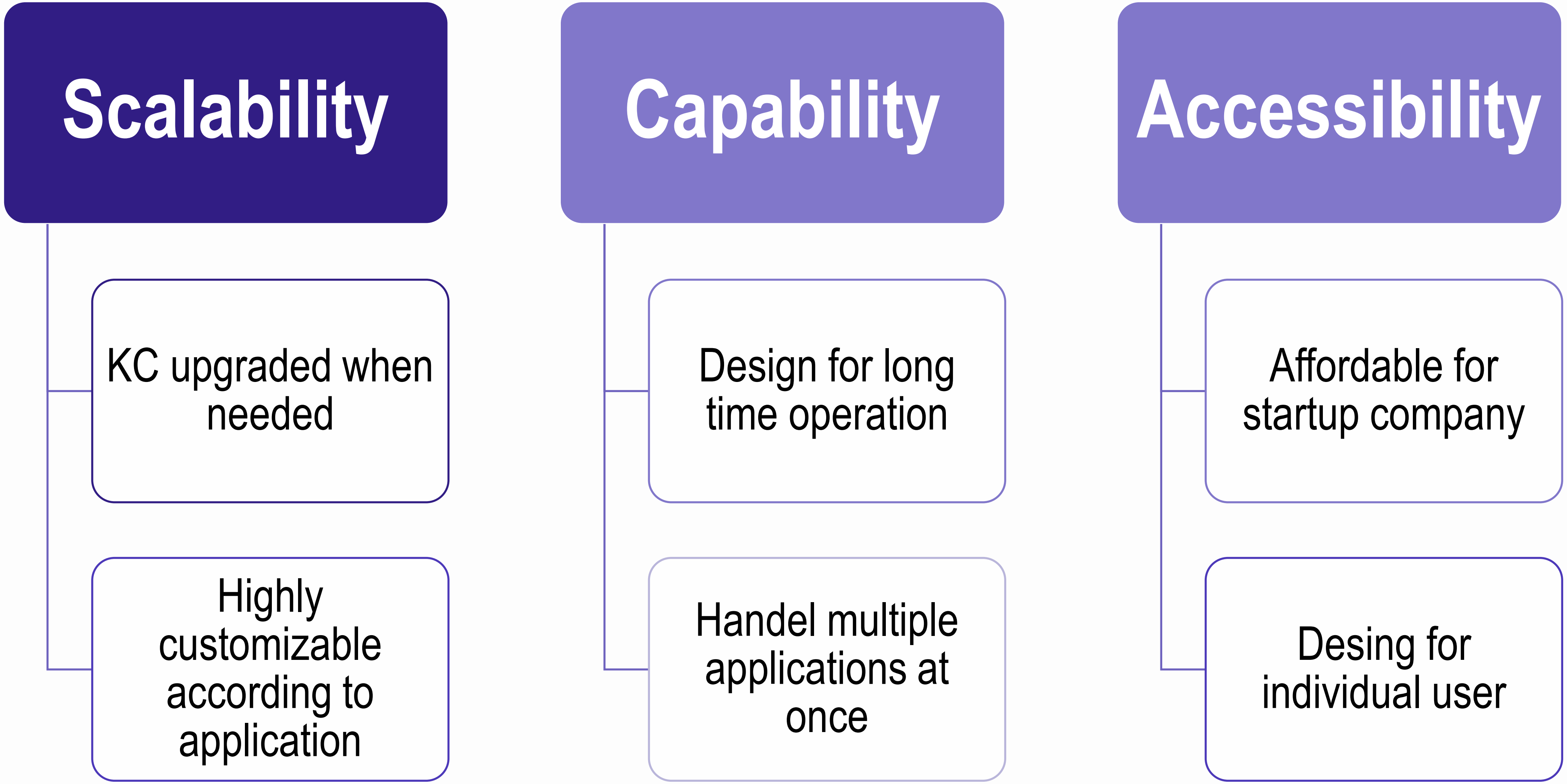
- Server or Workstation
- X13 DP WS Solutions
- New GPU Implementation
- Workstation Application
- New CPU Implementation

Role of High-Performance Workstation

- Performance is in between Personal Computer & Server
- It is designed for a single user usage with advance graphic & large storage capabilities
- Workstations are used primarily to perform computationally intensive scientific and engineering tasks, also in some complex financial and business applications.
- High-end workstations often serve a network of attached “client” PCs, which use resident tools and applications to access and manipulate data stored on the workstation.

Advantage of Workstation

Confidential



Scalability

- KC upgraded when needed
- Highly customizable according to application

Capability

- Design for long time operation
- Handel multiple applications at once

Accessibility

- Affordable for startup company
- Desing for individual user

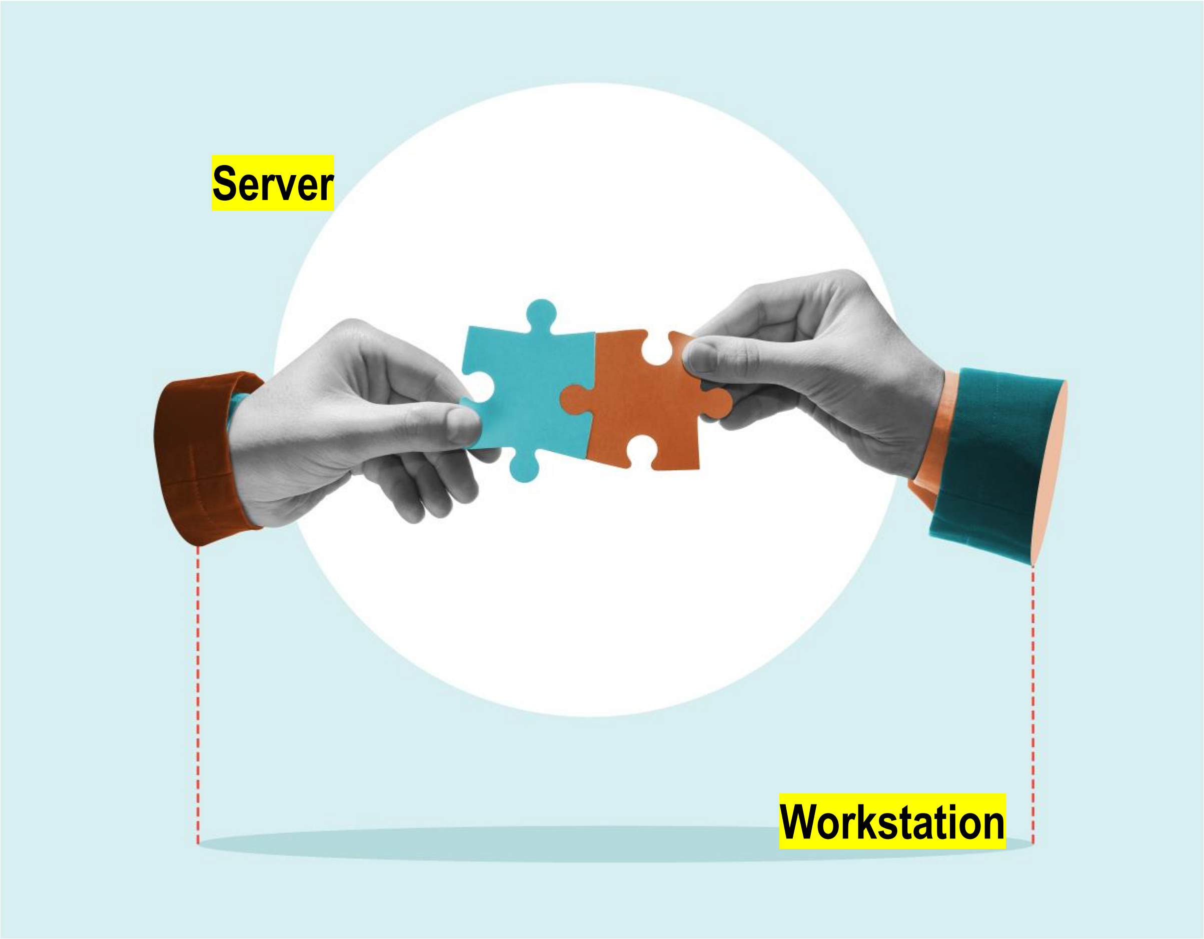
Comparison Table

	Server	Workstation
Definition	A server is an application or device that performs service for connected clients as part of client server architecture.	A computer that is used to power applications such as graphic art, 3-D design, Video Editing, or other CPU/RAM intensive software
Function	Internet, Office, Education, Home Networks	Business, Design, Engineering, Multi-Media Production
Operating systems	Free BSD, Solaris, Linux, Windows server	Unix, Linux, Windows workstation
GUI (Graphic User Interface)	Optional	Installed
Examples	Web servers, application servers etc	Video and audio workstation.
Application	Hosting, Intranet	Professional, Individual
Reliability	Often comes with error correcting (ECC) DDR modules, storage disks are typically in RAID and often have more than one power supply unit along with more than one Network port. Can be run in multiple-CPU setups.	No error correcting DDR modules (non-ECC), RAID storage disks aren't typically used. Only one power supply unit and very often only one network port.

SUPERMICRO

Confidential

Last puzzle of AI workload



Individual data scientists, data engineers, and AI researchers often use a personal AI or data science workstation in the process of building and maintaining AI applications.

GPU-accelerated workstations make it possible to build complete model prototypes using an appropriate subset of a large dataset. This is often done in hours to a day or two.

Certified hardware compatibility along with seamless compatibility across AI tools is very important. SMC is the only organization certified 3 different program by Nvidia

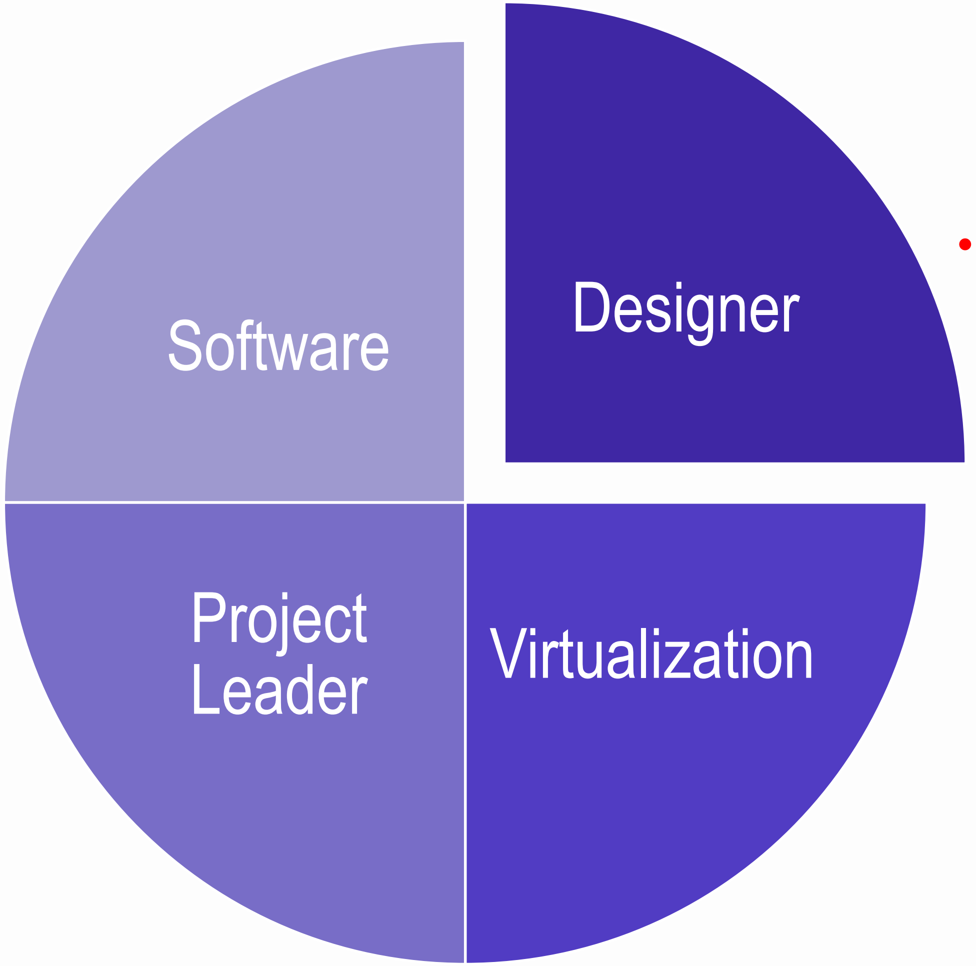
- Data Center
- Workstation
- Edge Computing

NVIDIA. CERTIFIED

Workstation Target Verticals

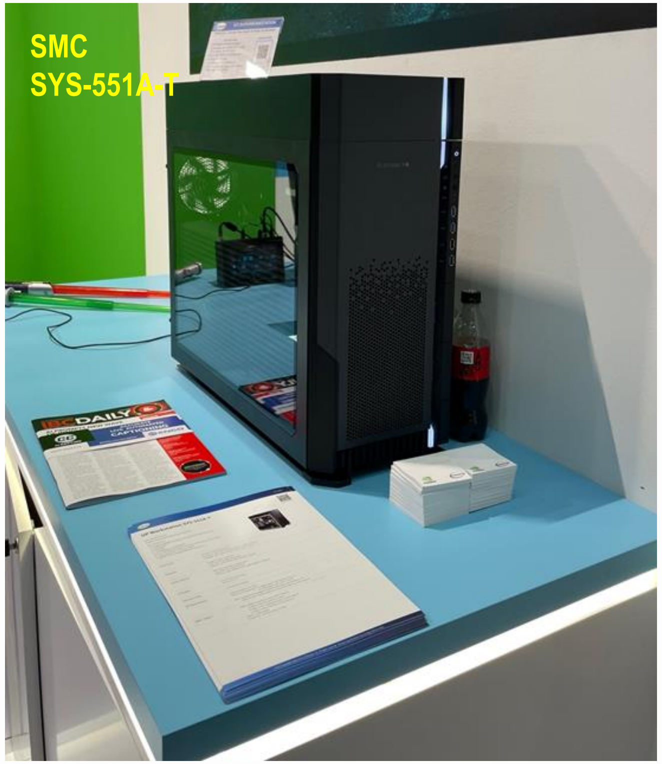
- **AI TRAINING & INFERENCE**
  - Software Compiling
  - AI Training
  - ChatGPT-like AI
  - LLaMA (Lama glama)
  - Alpaca

- **MANUFACTURING & ENGINEERING**
  - CAD Design
  - 3D Modeling
  - Rendering
  - Design Engineering
  - Virtual Reality



- **MEDIA & ENTERTAINMENT**
  - Video Editing
  - Media transcoding
  - Rendering
  - Lighting & Look Development
  - Animation

IBC 2023


SMC
SYS-551A-T

ASUS
ProArt

DELL
DELL Technologies
NVIDIA
AMD

HPE
Editing & Motion Graphics

X13 DP WS Solutions

X10-X11-X12-X13 DP MB Transition Chart

Confidential

Verticals	X10 DP	X11 DP	X12 DP	X13 DP (planning)
			Q2/2020 ~ Q2/2027	~Q1/2023 ~ TBD

Mainstream	X10DRI(-T)
	X10DRL-I(LN4)/C(T)
		X11DPI-N(T)
		X11DPL-i
			X12DPi-N(T)6
			X12DPL-i6/NT6
				X13DEi-(T)

Workstation	X10DAI/C/X X10DRG-Q X10DAL-I	X11DAi-N X11DAC X11DPG-QT	X12DAi-N6 X12DPG-QT6	**X13DAI-T** **X13DEG-QT**
Ultra	X10DRU-I+/X(LL)	X11DPU(-V) X11DPU-Z(E)+ X11DPU-XLL X11DPU-R	X12DPU-6 X12DHM-6	X13DEM (Hyper)
Twin Series	X10DRT-L(IBQ/IBF) X10DRT-H/HIBQ/HIBF X10DRT-P(T)/PIBQ/PIBF X10DRT-B+	X11DPT-PS X11DPT-BR X11DPT-B(H) X11DPT-L	X12DPT-PT6 X12DPT-B6	X13DET-B(BigTwin)
GPU Optimized	X10DGQ X10DRG-H(T) X10DRG-O(T)+-CPU	X11DGQ X11DPG-OT-CPU X11DGO X11DPG-SN	X12DPG-OA6 X12DPG-AR X12DGO-6 X12DGQ-R	X13DEG-OA (4U10GPU) X13DEG-QR (Redstone) X13DGO (Delta)
CouldDC/MegaDC	X10DRW-I(T) X10DDW-I(N) X10DRW-E/N(T)	X11DDW-L/NT X11DPD-L/M25	X12DDW-A6 X12DPD-A6/M25	X13DDW-A (CloudDC)
Data Center Optimized	B10DRC/-N B10DRi B10DRG-IBF/IBF2/TP X10DRD-I(N)TP/LTP X10DRD-L/I(N)(T)	B11DPT-P B11DPE	B12DPT-6 B12DPE-6	B13DET (SuperBlade) B13DEE
Resource Optimized	X10DRC/I-LN4+/T4+ X10DRH-C/I(T) X10DRH-ILN4/CLN4 X10DRX	X11DPH-T(q) X11DPX-T	X12DPi-N(T)6	X13DEi-(T)

FatTwin	X10DRFF-C/I(T)G
	X10DRFR(-T)
	X10DRFR-N(T)
		X11DPFR-S(N)
		X11DPFF-SN(R)
			X12DPFR-AN6	N/A (UP FatTwin)

Storage	X10DSC+ X10DSN-TS X10DSC-TP4S X10DRS	X11DSC X11DSN-TS(q) X11DSF-E X11DPS-RE X11DSC+	X12DSC-6	X13DSF-A (NVMeAll flash)

Supermicro DP Workstation Lineup

Expert 2S GPU WS

Intel® 4th Gen Xeon® SP and Xeon Max series

Liquid Cooling

SYS-751GE-TNRT

Validated Accelerated GPUs:

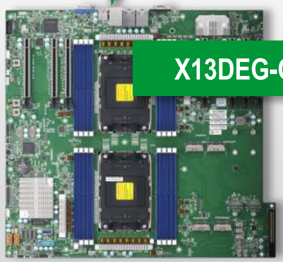
- 4x A100 w/ Liquid Cooler
- 4x H100 w/ Liquid Cooler(planning)

Air Cooling

SYS-741GE-TNRT

Validated Accelerated GPUs:

- **Nvidia**:
  - 2x H100 NVL, 4x H100, 4X A100
  - 4x RTX 6000 Ada, 4x RTX A5500
  - L40S(testing), 4x L40, 7x L4
- **AMD** MI210
- **Intel** Data Center GPU Flex series



X13DEG-QT

Key Features

DDR5 | PCIe 5.0 | CXL 1.1 | NVMe VROC

Built-in Accelerators (IAA, DSA, QAT, DLB, AMX) | MAX CPU

Expert 2S WS

Intel® 4th Gen Xeon® SP

Air Cooling

SYS-751A-I

Optional Accessory

P/N: DVM-TEAC-DVDRW24-HBT

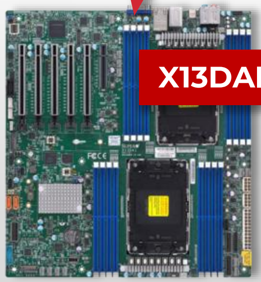
DVD/RW Drive Kit

P/N:
1x MCP-290-GS706-0N
1x MCP-290-00057-0N

Rack Mount Kit

Validated Accelerated GPUs:

- RTX 6000 Ada, RTX 5000 Ada, RTX 4000 Ada, RTX 4000 Ada SFF(testing)
- RTX A6000, RTX A5500, RTX A4500, RTX A2000(testing)
- Quadro RTX T1000, RTX T400



X13DAI-T

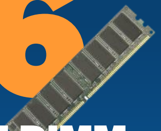
Key Features

DDR5 | PCIe 5.0 | CXL 1.1 | VROC

Built-in Accelerators (IAA, DSA, QAT, DLB, AMX) | EATX

Expert 2S Workstation – SYS-751A-I

SYS-751A-I



16 DDR5 DIMM

2 M.2 NVMe

6 PCIe 5.0 Slots (5x PCIe x16 + 1x PCIe x8)

3 x dual-width GPUs or 6 x single-width GPUs

8 Hybrid Storage

8 x SATA 3.0 or 4 x SATA 3.0 + 4 x NVMe SSD

10Gps 1 x USB 3.2 Gen 2

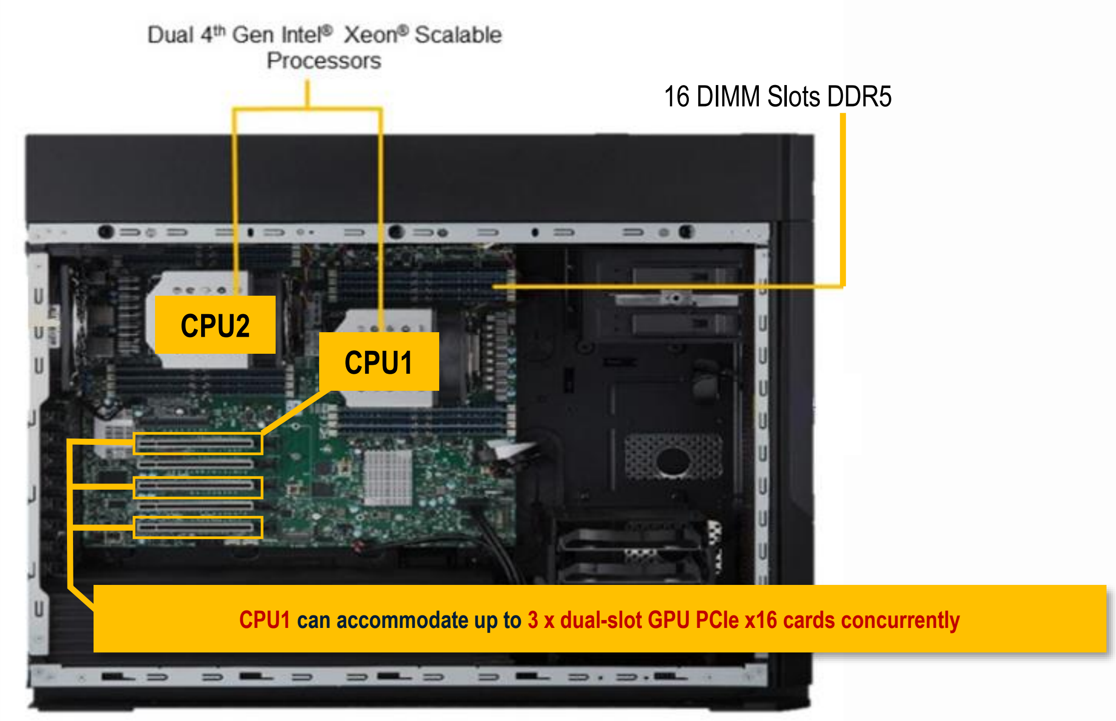
10G Dual Network Ports

- Design for 24/7 & 365 Operations
- Intel® Dual 4th Xeon® CPU support Higher CPU core frequency when running single threaded workloads
- Fast storage have faster launch times, faster loading times and caching
- Fast USB Transfer Speed when accessing larger video files
- A variety of professional graphic cards with optimized thermal design
- **Compatible with DVD/RW Drive Kit to back up your data and images, or to access DVD disk with license key**
- **Rack Mount Kits provided to install on your server rack**

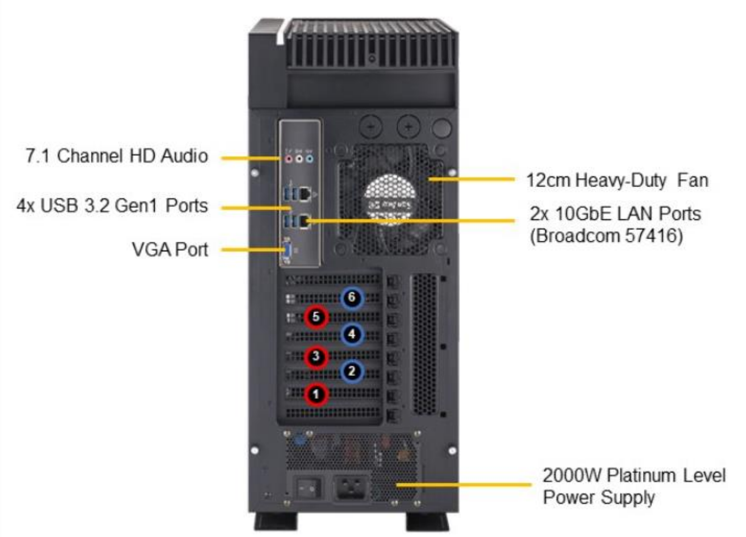
Expert 2S WS - Intel® 4th Gen Xeon® SP

Confidential

SYS-751A-I



CPU1 can accommodate up to 3 x dual-slot GPU PCIe x16 cards concurrently



	Slot Description
1 2 3 4 5	PCIe 5.0 x16
6	PCIe 5.0 x8

Feature Details:

- Intel Dual 4th Gen Xeon SP (XCC/MCC)
- 16 x DIMM Slot, 1DPC ECC DDR5 designed for up to 4800 MT/s
- 256, 128, **96 (only XCC SKU – competitive price)**, 64, 32, 16 GB Memory support
- 5x PCI Gen5 x16 slots
- PCI-E Dual Root for maximum bandwidth
- Dual 10GbE RJ-45 LAN

- USB 3.2 10Gbps support
- 4x NVMe/3.5” SATA drive bays
- up to 8x 2.5” NVMe/SATA drive bays by cage
- 2000W PS/2 power
- **Operating Temperature:**
  - Support 350W(2P) at 30˚C
  - Supports 350W (2P) with 1 or 2 GPU PCIe cards at 25°C

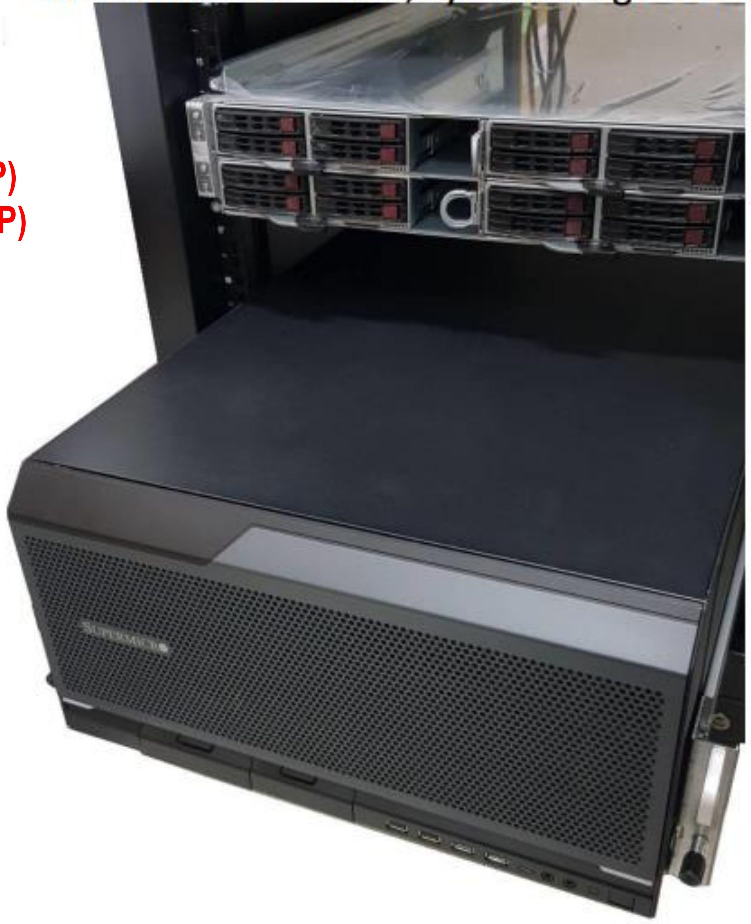
Workstation Deployment in Rack

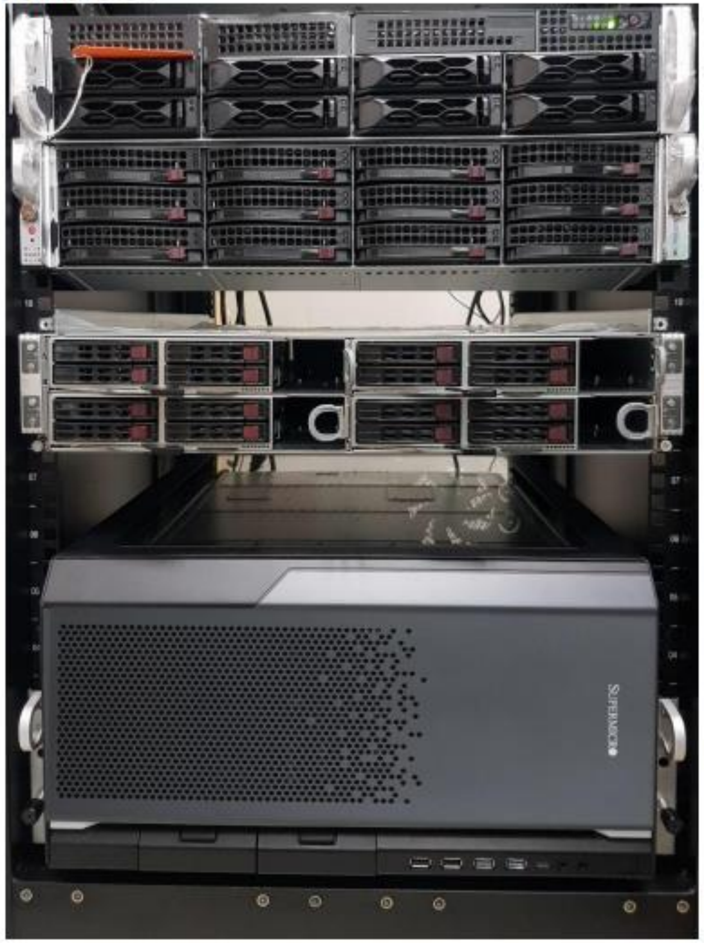
Confidential

SYS-751A-I (DP)
SYS-551A-T (UP)

Rackmount Conversion Kit
+ MCP-290-GS706-0N
and rail (6U):
+ MCP-290-00057-0N

In standard 19" Rack, system height is around 5U.





Expert 2S WS - Intel® 4th Gen Xeon® SP

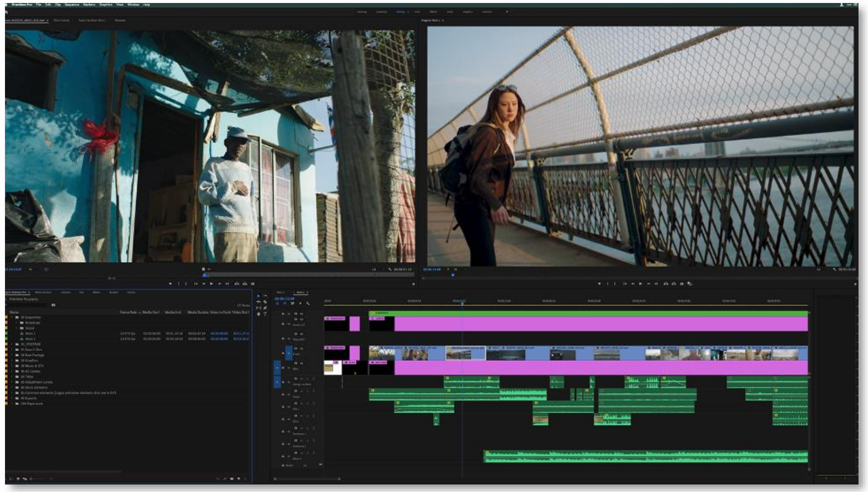
SUPERMICRO

Confidential

SOLUTIONS FOR MEDIA & ENTERTAINMENT

Best for

- Video streaming & Editing
- Media transcoding
- Rendering
- Lighting & Look Development
- Animation



Avid Media Composer

DaVinci Resolve

Adobe Premiere Pro

Foundry Nuke

Adobe After Effects

Autodesk Flame

SYS-751A-I

Higher clock-speed & Higher core-count of CPU and the **large memory (at least 32GB)** would be required for high quality video such as 4K or 8K that provide the best results.

Primary drive would be **NVMe or SATA SSD** for the operating system and all applications to have faster launch times, faster loading times and caching

Some workflows such Media Composers render & playback would be affected what GPU card you select. **Our offerings range from high level to entry level**. In addition to **RTX Nvidia** solution, **Intel GPU solution** has been also validated to be provide the hardware encoder/ decoder capabilities

Expert 2S WS - Intel® 4th Gen Xeon® SP

Confidential

SOLUTIONS FOR MEDIA & ENTERTAINMENT

Best for

- Video streaming & Editing
- Media transcoding
- Rendering
- Lighting & Look Development
- Animation

SYS-751A-I

Render Engine hardware Combability list

	NVIDIA GPU (CUDA / OptiX)	AMD GPU (OpenCL)	CPU support	CPU+GPU Hybrid
V-Ray (Some Versions)	🟢	🟢	🟢	🟢
V-Ray NEXT	🟢	🔴	🟢	🟢
Redshift	🟢	🔴	🔴	🔴
Octane	🟢	🔴	🔴	🔴
Arnold	🟢	🔴	🟢	🔴
Maxwell	🔴	🔴	🟢	🔴
MentalRay	🔴	🔴	🟢	🔴
Enscape	🟢	🟢	🔴	🔴
Lumion	🟢	🟢	🔴	🔴
Twinmotion	🟢	🟢	🔴	🔴
Twilight Render	🔴	🔴	🟢	🔴
F-Storm	🟢	🔴	🔴	🔴
RenderMan	🔴	🔴	🟢	🔴
AMD ProRender	🟢	🟢	🟢	🔴
TheaRender	🟢	🟢	🟢	🟢
Corona	🔴	🔴	🟢	🔴
Cinema 4D (physical)	🔴	🔴	🟢	🔴
Cinema 4D (standard)	🔴	🔴	🟢	🔴
Cinema 4D (prorender)	🟢	🟢	🔴	🔴
Blender (Internal)	🔴	🔴	🟢	🔴
Blender (Cycles)	🟢	🟢	🟢	🟢

Depends on applications, 3D Modeling, Animation, Rendering can work in GPU/CPU/Hybrid modes.

Under CPU mode

As many cores as possible with large system RAM capacity.

Under GPU mode

VRAM is the important factor when doing the complex projects(at least 16GB VRAM). Our GPU VRAM offering ranges **from 4GB to 48GB**.

Primary drive would be **NVMe or SATA SSD** for the operating system and all applications to have faster launch times, faster loading times and caching

Autodesk Maya

Autodesk 3ds Max

Autodesk Arnold

Pixar Renderman

Sidefx Houdini

REDSHIFT

Expert 2S WS - Intel® 4th Gen Xeon® SP

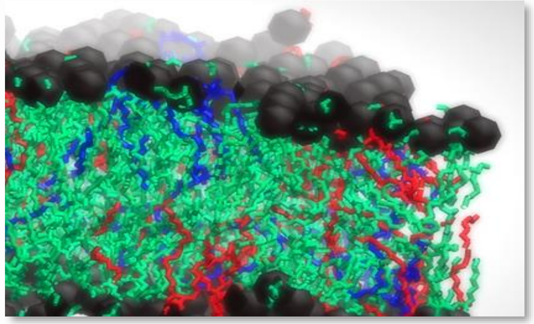
SOLUTIONS FOR LIFE SCIENCES

Best for

- Molecular Dynamics
- Quantum Chemistry
- Molecular Visualization and docking
- Bioinformatics
- Microscopy

SYS-751A-I

Molecular Dynamics



Great multi-GPU performance

Single precision (FP32) dominated

Applications:
ACEMD*, AMBER*, HOOMD-Blue*, Lattice Microbes*, SOP-GPU*, BAND, CHARMM, DESMOND, ESPResso, GROMACS, HALMD, LAMMPS, mdcore, MELD, miniMD, NAMD,.. etc.

blue* = application where > 90% of workloads is on GPU

Quantum Chemistry

Focus on using **GPU-accelerated** math libraries, OpenACC directives.

Double precision (FP64) is important.

Active GPU acceleration projects:
CASTEP, GAMESS, Gaussian, ONETEP, Quantum Supercharger Library*, VASP,... etc.

blue* = application where > 90% of workloads is on GPU

Running simulations required **higher hardware configurations with parallel computing** that we can offer the high-end GPU acceleration solutions, such Nvidia RTX 6000 or 5000 Ada architecture

Expert 2S WS - Intel® 4th Gen Xeon® SP

Confidential

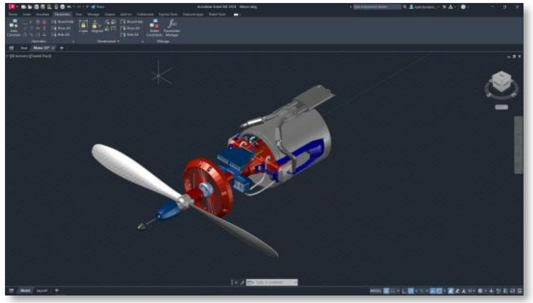
SOLUTIONS FOR MANUFACTURING & ENGINEERING

Across industries ranging from automotive to aerospace to consumer electronics

NVIDIA-based Workstations are conceived, developed, and manufactured.

SYS-751A-I

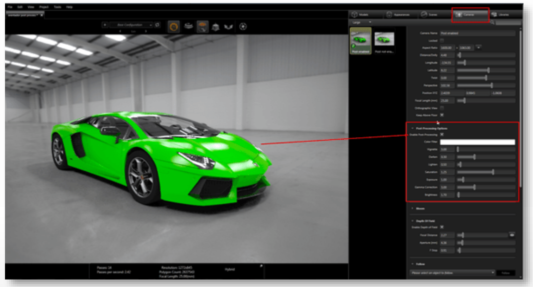
CAD Design



Applications

Autodesk AutoCAD, Fusion 360, Generative Design, Inventor, Dassault Systèmes CATIA, SOLIDWORKS, Rhino, PTC Creo, Siemens NX, Solid Edge, ESI IC.IDO, Virtalis VR4CAD

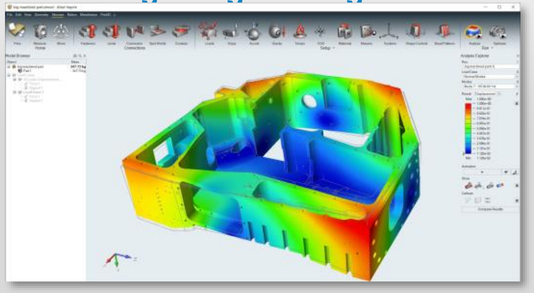
Rendering



Applications

SOLIDWORKS Visualize, Autodesk 3ds Max, Chaos V-Ray for Rhino, Allegorithmic Substance, Designer/Paint

Design Engineering



Applications

Autodesk AutoCAD, Fusion 360, Generative Design, Inventor, Dassault Systèmes CATIA, SOLIDWORKS , PTC Creo, Siemens NX, Solid Edge, ESI IC.IDO, Virtalis VR4CAD, Altair FluiDyna, HyperWorks, ANSYS Discovery Live, Fluent, Mechanical

Expert 2S WS - Intel® 4th Gen Xeon® SP

SOLUTIONS FOR AI TRAINING & INFERENCE

NVIDIA-based AI Development Workstations

SYS-751A-I

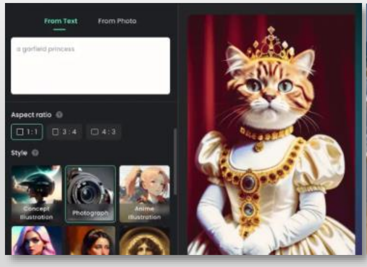
AI Training & Development

AI developers for prototyping, developing, and refining generative AI models in an on-premises environment, giving them the flexibility to experiment and calibrate AI workloads without racking up costs

Optimized for **Stable Diffusion, LLaMA, Alpaca, ChatGPT-like AI**

AI Inference

Deploy your trained models confidently, as have the capability to run parallel inference



Text-to-Image
(SDXL/SD models)

Text-to-Video
(Zeroscope model)

Text-to-Speech
(SpeechT5 model)

Nvidia RTX Accelerating Solutions for the different Workloads

Confidential

Quadro RTX T1000
Quadro RTX T400

Entry Level

- Video/Graphic Editing
- 3D CAD
- Gaming Development

RTX 5000 Ada, RTX 4500 Ada
RTX 4000 Ada, RTX 4000 Ada SFF
RTX A6000, RTX 5500
RTX 4500, RTX A2000

Medium Level

- Animation Development
- Big Data Analytics
- Real-Time rendering
- Content Creation

RTX 6000 Ada
RTX 5000 Ada

High Level

- Small/Medium Scale Simulation
- Virtual Reality Application Development
- AI/MI Development
- Virtual GPU and VDI solution such for Engineering & Sciences

Details of Nvidia RTX Accelerating Solutions

Confidential

"-" indicates Not Support

Performance (based on Tensor core)

	GPU PCIe	Arch.	NVLINK Bridge	PCIe Form factor	TDP	Memory	Decode encoder	Display	vGPU	Tensor Core (TFLOPS)	RT Core (TFLOPS)
Verified	**RTX 6000 Ada (NEW)**	Ada	-	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	350W	48GB GDDR6, 960GB/s	3x NVENC 3x NVDEC (+AV1 encode and decode)	4 x DisplayPort 1.4a	Yes	1457	210.6
Verified	**RTX 5000 Ada (NEW)**	Ada	-	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	250W	32GB GDDR6, 576GB/s	2x NVENC 2x NVDEC (+AV1 encode and decode)	4 x DisplayPort 1.4a	Yes	1044.4	151
Verified	**RTX 4500 Ada (NEW)**	Ada	-	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	210W	24GB GDDR6, 432GB/s	2x NVENC 2x NVDEC (+AV1 encode and decode)	4 x DisplayPort 1.4a	-	634	91.6
Verified	**RTX 4000 Ada (NEW)**	Ada	-	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	130W	20GB GDDR6, 360GB/s	2x NVENC 2x NVDEC (+AV1 encode and decode)	4 x DisplayPort 1.4a	-	327.6	61.8
Verified	**RTX A6000**	Ampere	Yes	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	300W	48GB GDDR6, 768GB/s	1x NVENC 2x NVDEC (+AV1 encode and decode)	4 x DisplayPort 1.4a	Yes	309.7	75.6
Testing	**RTX 4000 Ada SFF (NEW)**	Ada	-	• PCIe 4.0 x16 • dual-slot air cooling • ????	70W	20GB GDDR6, 280GB/s	2x NVENC 2x NVDEC (+AV1 encode and decode)	4 x Mini DisplayPort 1.4a	-	306.8	44.3
Verified	**RTX A5500**	Ampere	Yes	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	230W	24GB GDDR6, 768GB/s	1x NVENC 2x NVDEC (+AV1 encode and decode)	4 x DisplayPort 1.4a	Yes	272.8	66.6
Verified	**RTX A4500**	Ampere	Yes	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	200W	20GB GDDR6, 640GB/s	1x NVENC 1x NVDEC (+AV1 encode and decode)	4 x DisplayPort 1.4a	-	189.2	46.2
Testing	**RTX A2000**	Ampere	-	• PCIe 4.0 x16 • dual-slot air cooling • FHFL	70W	6GB GDDR6, 288GB/s	1x NVENC 1x NVDEC (+AV1 encode and decode)	4 x Mini DisplayPort 1.4a	-	63.9	15.6
Verified	**RTX T1000**	Turing	-	• PCIe 3.0 x16 • dual-slot air cooling • FHFL	50W	4GB GDDR6, 160GB/s	-	4 x Mini DisplayPort 1.4a	-	-	-
Verified	**RTX T400**	Turing	-	• PCIe 3.0 x16 • dual-slot air cooling • FHFL	30W	2GB GDDR6, 80GB/s	-	4 x Mini DisplayPort 1.4a	-	-	-

SUCCESSFUL CASES

X13DAI-T E-ATX adoption to Video Production Workstation

A Japan customer, A provider for the Live Media and Entertainment Market, is in pursuit of high-quality motherboard that requires the powerful CPU computing, higher memory bandwidth, the fast SSD with RAID and the large-scaled Storage with RAID, that has been seamlessly integrated into custom Chassis.

X13DAI-T E-ATX adoption to SIEMENS NX Workstation

A USA company provides AutoCAD workstation serves the vast majority of design needs for large sections of the design and engineering industry. This configuration is a dual socket system with high core-count and high clock-speed, fast M.2 NVMe storage and Nvidia Quadro T1000, to keep its performance as high as possible.

titan
WORKSTATION
intel
xeon
titan
W799
DUAL 4TH GEN INTEL XEON SCALABLE CPUS
UP TO 112 CORES FOR DEEP LEARNING, AI WORKSTATION PC

What we can offer:

- Building Blocks Solutions: Motherboard, Chassis, Power Supplies, Accessories, … etc.
- A variety of acceleration GPU Solutions for rendering special effects, color grading, and even video decoding and encoding
- One Stop Shop

MAX GPU 2S Workstation – SYS-741GE-TNRT

HIGH PERFORMANCE WORKSTATIONS

SYS-741GE-TNRT

16 DDR5 DIMM

2 M.2 NVMe

7 PCIe 5.0 x16 Slots
4 x dual-width GPUs or 7 x single-width GPUs

8 Hybrid Storage
8 x hot-swap NVMe/SATA

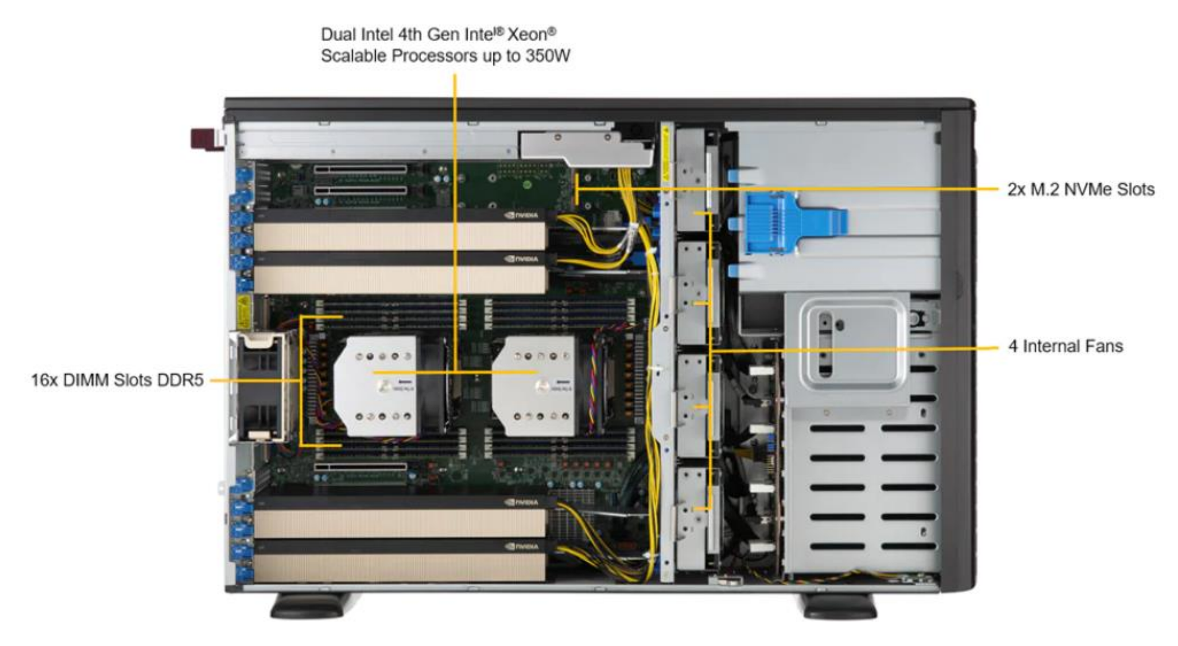
10G Dual Network Ports

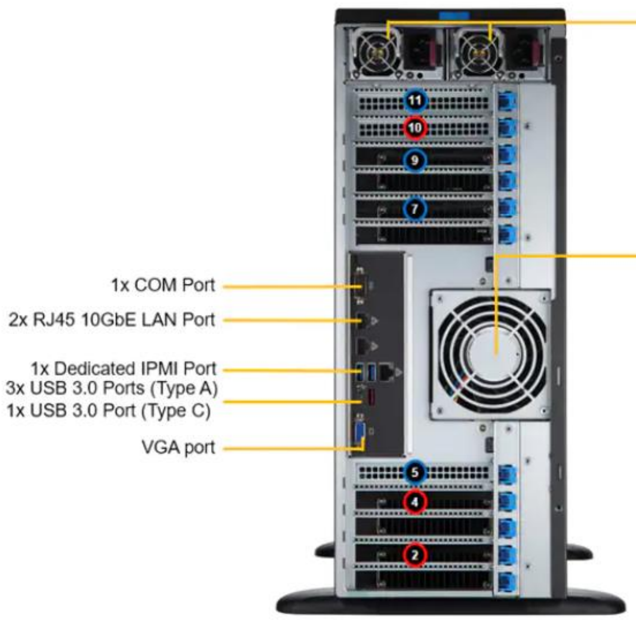
7x USB 3.2 Gen 1 Ports
(3 Type A, 1 rear Type C, 1 internal Type A, 2 via header)

- Design for 24/7 & 365 Operations
- Hardware Balance design & optimized thermal design
- **Intel Xeon® MAX CPU series support for HPC application**
- Fast storage have faster launch times, faster loading times and caching
- A variety of professional graphic cards support with optimized thermal design

intel xeon MAX SERIES | intel xeon PLATINUM | intel xeon GOLD | intel xeon SILVER | intel xeon BRONZE

X13 4U 4 GPU Workstation - SYS-741GE-TNRT


Dual Intel 4th Gen Intel® Xeon®
Scalable Processors up to 350W
2x M.2 NVMe Slots
4 Internal Fans
16x DIMM Slots DDR5



	Slot Name	Slot Description
2	Slot 2	PCIe 5.0 x16 (For Double-Width GPU)
4	Slot 4	PCIe 5.0 x16 (For Double-Width GPU)
10	Slot 10	PCIe 5.0 x16
5	Slot 5	PCI-E 5.0 x16
7	Slot 7	PCI-E 5.0 x16 (For Double-Width GPU)
9	Slot 9	PCI-E 5.0 x16 (For Double-Width GPU)
11	Slot 11	PCI-E 5.0 x16

CPU1 🟥 CPU2 🟦

Feature Details:

- Intel Dual 4th Gen Xeon SP (XCC/MCC) & **Intel Xeon Max series CPU**
- 16 x DIMM Slot, 1DPC ECC DDR5 designed for up to 4800 MT/s
- 256/128/**96 (only XCC SKU)**/64/32/16 GB Memory support
- Intel® 3rd Optane Persistent Memory Not available (EOL)
- **4x PCIe 5.0 x16 /CXL 1.1 (double-width)**, 3x PCIe 5.0 x16 /CXL 1.1 (single-width)
- Dual 10GbE RJ-45 LAN
- BMC AST2600 with RoT2.0 supports, 1x Dedicated BMC LAN port

Support 8 drives without additional storage PCIe card

- 2x M.2 NVMe for boot drive only
- **8x 3.5" Hot-swap SATA/NVMe/SAS drive bays**
- 3x 2.5" Fixed drive Bays
- 1 x VGA D-Sub connector(from BMC AST2600)
- 7 USB3.2 Gen 1 ports (3 Type A, 1 rear Type C, 1 internal Type A, 2 via header)
- 2x 2000W (1+1) Redundant Power Supplies, Titanium Level
- **Trusted Platform Module (TPM) onboard and SMC IPMI with RoT 2.0**

Windows Linux vmware ESXi

Confidential

MAX GPU 2S Workstation – SYS-741GE-TNRT

Accelerating Solutions for the different Workloads

Nvidia L40

Focus on workloads:

- Generating image AI inference
- 2D/3D content generation
- Video content moderation
- Real-time language translation
- Virtual GPU and VDI solution

Nvidia L40S

Focus on workloads:

- Generating image AI inference
- Large language model(LLM) inference and training.
- 2D/3D content generation
- Video content moderation

Nvidia H100 NVL

- Used for deploying large-scale LLMs training such as GPT-2 and support 188GB memory w/ 7.8TB/s BW
- Transformer Engine Acceleration capabilities to enable the faster training times and significant improvements

Nvidia L4

Focus on workloads:

- Generative Video AI inference
- Speech AI (ASR + NLP + TTS)
- Augmented Reality
- Virtual Workstations

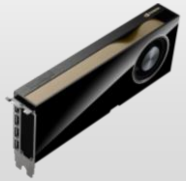
SYS-741GE-TNRT

Intel AMX accelerator

Small model Generative AI training/inference

- Fine tune Stable Diffusion model, it can be completed **within 2 hours**.
- Inference Workload of Stable Diffusion (under 32 cores with 512GB RAM), the image was inferred **in 5 secs.**

Nvidia 6000 Ada



Focus on workloads:

- Generating AI inference
- Virtual GPU and VDI solution

AMD MI210

Focus the workloads:

- HPC (Scientific Field)
- AI DL/ML training/inference
- Universities/Geoscience/Life Science

Intel Data Center GPU

Focus the workloads:

- HPC (Scientific Field)
- AI DL/ML training/inference

oneAPI

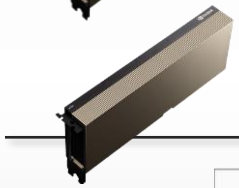
oneAPI can also be interoperable with Fortan Meta **Llama**

Workstation GPU PCIe Solutions

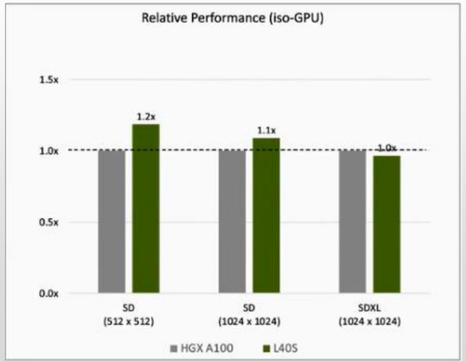
Compute Performance – AI/ML & Generative AI Training, Data Analytics, HPC

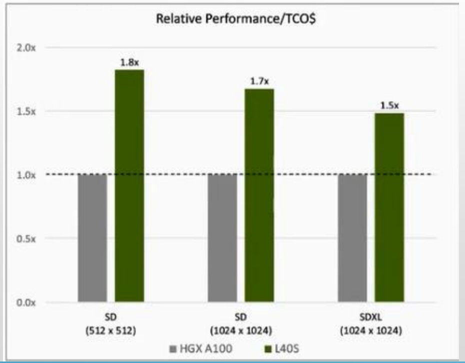
Confidential

Performance (based on Tensor core)

	GPU PCIe	Arch.	NVLINK Bridge	PCIe Form factor	TDP	Memory	Decode encoder	Tensor Core (TFLOPS / TOPS)							Cuda Core (TFLOPS)	
								FP64	TF32	FP16	BF16	FP8	INT8	INT4	FP64	FP32
	H100 NVL 80GB	Hopper	2x w/ 600GB/s enabled	• PCIe 5.0 x16 • dual-slot air cooling • FHFL	2x 350W - 400W	188GB HBM3, 7.8 TB/s	14 NVDEC 14 JEPG	134	1979	3958	3958	7916	7916	-	68	134
																based on sparsity matrix
	H100 80GB	Hopper	600GB/s	• PCIe 5.0 x16 • Dual-slot air cooling • Single-slot liquid cooling • FHFL	350W	80GB HBM2e, 2 TB/s	7 NVDEC 7 JEPG	51	756	1513	1513	3026	3026	-	26	51
																based on sparsity matrix
Available in Sep. 	**L40S 48GB**	Ada	-	• PCIe Gen4 x16 • Dual-slot air cooling • FHFL	350W	48GB GDDR6, 864 GB/s	3x NVENC(+AV1) 3x NVDEC 4x NVJEPG	-	366	733	733	1466	1466	1466	-	91.6
																based on sparsity matrix
	**A100 80GB**	Ampere	600GB/s	• PCIe 4.0 x16 • Dual-slot air cooling • Single-slot liquid cooling • FHFL	300W	80GB HBM2e, 2 TB/s	5x NVDEC 1 NVJEPG	19.5	312	624	624	-	1248	2496	9.7	19.5
																based on sparsity matrix

"?" indicates Not revealed by Nvidia; "-" indicates Not Support





According to Nvidia Test Result:
L40S delivers better performance than A100 in
LLM Inference/Training, Generative AI & **TCO$**
(Not support FP64 application)

SUPERMICRO

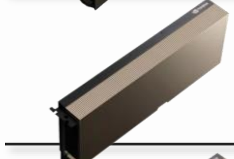
Confidential

Workstation GPU PCIe Solutions

Compute Performance – AI/ML & Generative AI Training, Data Analytics, HPC

Performance (based on Tensor core)

Available in Sep.

	GPU PCIe	Arch.	NVLINK Bridge	PCIe Form factor	TDP	Memory	Decode encoder	Tensor Core (TFLOPS / TOPS)							Cuda Core (TFLOPS)	
								FP64	TF32	FP16	BF16	FP8	INT8	INT4	FP64	FP32
	H100 NVL 80GB	Hopper	2x w/ 600GB/s enabled	• PCIe 5.0 x16 • dual-slot air cooling • FHFL	2x 350W - 400W	188GB HBM3, 7.8 TB/s	14 NVDEC 14 JEPG	134	1979	3958	3958	7916	7916	-	68	134
													based on sparsity matrix			
	H100 80GB	Hopper	600GB/s	• PCIe 5.0 x16 • Dual-slot air cooling • Single-slot liquid cooling • FHFL	350W	80GB HBM2e, 2 TB/s	7 NVDEC 7 JEPG	51	756	1513	1513	3026	3026	-	26	51
													based on sparsity matrix			
	L40S 48GB	Ada	-	• PCIe Gen4 x16 • Dual-slot air cooling • FHFL	350W	48GB GDDR6, 864 GB/s	3x NVENC(+AV1) 3x NVDEC 4x NVJEPG	-	366	733	733	1466	1466	1466	-	91.6
													based on sparsity matrix			
	A100 80GB	Ampere	600GB/s	• PCIe 4.0 x16 • Dual-slot air cooling • Single-slot liquid cooling • FHFL	300W	80GB HBM2e, 2 TB/s	5x NVDEC 1 NVJEPG	19.5	312	624	624	-	1248	2496	9.7	19.5
													based on sparsity matrix			
	L40 48G	Ada	-	• PCIe Gen4 x16 • Dual-slot air cooling • FHFL	300W	48GB GDDR6, 864 GB/s	3x NVENC(+AV1) 3x NVDEC 4x NVJEPG	-	181	362	362	724	724	1488	-	90.5
													based on sparsity matrix			

Comparison between L40 & L40S, much improvement on TF32/FP16/BF16/FP8/INT8 data format

a; “-” indicates Not Support

Workstation GPU PCIe Solutions

Collection of GPU Accelerating PCIe Card P/N

Type	GPU PCIe	Description	Part Number
NVIDIA	H100 NVL	NVIDIA H100 NVL 80GB PCIe 5.0	GPU-NVH100NVL
	H100	NVIDIA H100 80GB PCIe 5.0	GPU-NVH100-80
	A100	NVIDIA A100 80GB HBM2 PCIe 4.0 (w/o CEC)	GPU-NVA100-80-NC
	RTX 6000 Ada	NVIDIA RTX6000 Ada 48GB GDDR6 PCIe 4.0	GPU-NVQRTX6000-ADA
	RTX 5000 Ada	NVIDIA RTX6000 Ada 32GB GDDR6 PCIe 4.0	GPU-NVQRTX5000-ADA
	RTX 4500 Ada	NVIDIA RTX4500 Ada 24GB GDDR6 PCIe 4.0	**Applying**
	RTX 4000 Ada	NVIDIA RTX4000 Ada 20GB GDDR6 PCIe 4.0	GPU-SMP-RTX4000ADA-PS
	RTX 4000 Ada SFF	NVIDIA RTX4000 Ada 20GB GDDR6 PCIe 4.0	GPU-NVQRTX4000-ADA-SFF
	RTX A6000	NVIDIA RTXA6000 48GB GDDR6 PCIe 4.0	GPU-NVQRTX-A6000
	RTX A5500	NVIDIA RTX A5500 24GB GDDR6 PCIe 4.0	GPU-NVQRTX-A5500
	RTX A2000	NVIDIA RTX A2000 6GB GDDR6 PCIe 4.0	GPU-NVQRTX-A2000
	L40S	NVIDIA Ada L40S 48GB GDDR6 PCIe 4.0	GPU-NVL40S
	L40	NVIDIA Ada L40 48GB GDDR6 PCIe 4.0	GPU-NVL40
	L4	NVIDIA Ada L4 24GB GDDR6 PCIe 4.0	GPU-NVL4
	RTX T1000	NVIDIA Quadro T1000 4GB GDDR6 PCIe 3.0	GPU-NVQT1000
		NVIDIA Quadro T1000 8GB GDDR6 PCIe 3.0	GPU-NVQT1000-8
	RTX T400	NVIDIA Quadro T400 4GB GDDR6 PCIe 3.0	GPU-NVQT400-4
AMD	AMD MI210	AMD Instinct MI210 64GB HBM2e PCIe 4.0	GPU-AMDMI210-PCIE-0008H

OS Compatibility

Confidential

Developers can run a GNU/Linux environment on Windows

	Type	X13DAI-T	X13DEG-QT
Windows	Windows 10 Enterprise	v	v
	Windows 10 Pro Workstation	v	v
	Windows 10 IoT Enterprise	v	v
	Windows 11 Enterprise	v	v
	Windows 11 Pro Workstation	v	v
	Windows 11 IoT Enterprise	v	v
	Windows 11 with WSL2	**v**	**TBC**
	Windows Server 2019	v	v
	Windows Server 2022	v	v
Linux	RHEL 8.7/9.1	v	
	RHEL 8.6/9.0		v
	CentOS 8.5		v
	Oracle 8.7	v	
	Oracle 8.6		v
	Rocky 8.7/9.1	v	
	Rocky 8.6		v
	SLES 15 SP4	v	v
	Ubuntu Server 22.04	v	v
vmware ESXi	VMWare ESXi 8.0	v	
	VMWare ESXi 7.0u3d		v

OS Compatibility

Windows 11 with WSL2

Confidential

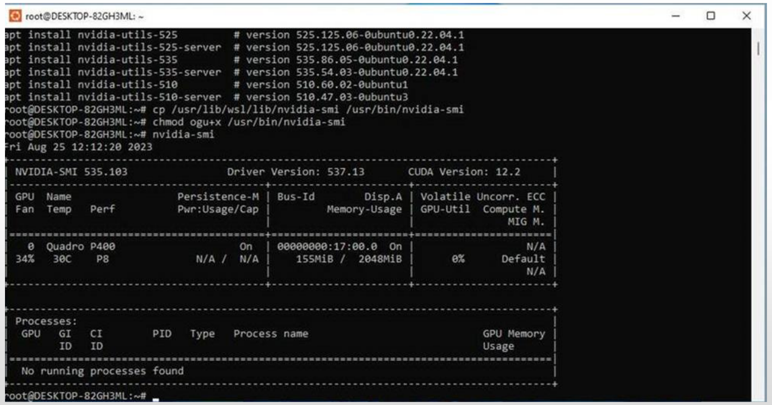
What is WSL?

Windows Subsystem for Linux (WSL) is a Windows 11 feature that enables you to run **native Linux command-line** tools directly on Windows, without requiring the complexity of a dual-boot environment such as installing VM on Windows to get the slowly execution.

Containerized environment that is tightly integrated with the Microsoft Windows operating system. This allows it to run Linux applications alongside traditional Windows desktop and modern store apps.

Benefits

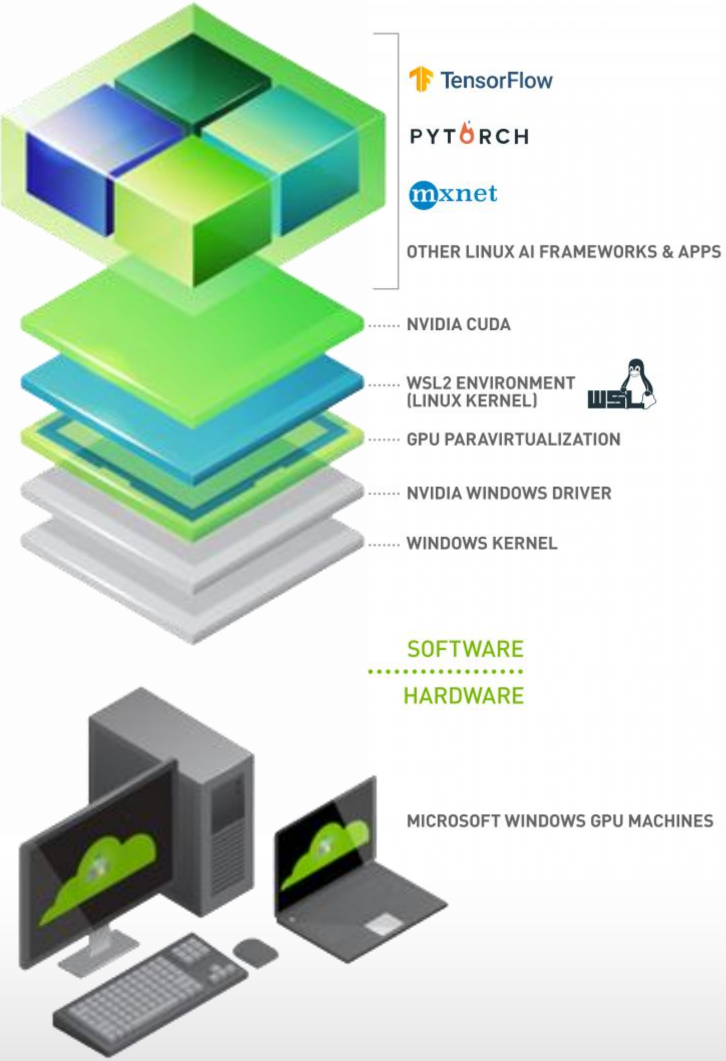
Without switching effort between Windows and Linux while working on CUDA development because some CUDA packages are only compatible with Linux platforms.



```
apt install nvidia-utils-525         # version 525.125.06-0ubuntu0.22.04.1
apt install nvidia-utils-525-server  # version 525.125.06-0ubuntu0.22.04.1
apt install nvidia-utils-535         # version 535.86.05-0ubuntu0.22.04.1
apt install nvidia-utils-535-server  # version 535.54.03-0ubuntu0.22.04.1
apt install nvidia-utils-510         # version 510.60.02-0ubuntu1
apt install nvidia-utils-510-server  # version 510.47.03-0ubuntu3
root@DESKTOP-82GH3ML:~# cp /usr/lib/wsl/lib/nvidia-smi /usr/bin/nvidia-smi
root@DESKTOP-82GH3ML:~# chmod ogu+x /usr/bin/nvidia-smi
root@DESKTOP-82GH3ML:~# nvidia-smi
Fri Aug 25 12:12:20 2023
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.103                 Driver Version: 537.13       CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  Quadro P400                     On | 00000000:17:00.0  On |                  N/A |
| 34%   30C    P8              N/A /  N/A |    155MiB /  2048MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|  No running processes found                                                           |
+---------------------------------------------------------------------------------------+
root@DESKTOP-82GH3ML:~#
```

X13DAI-T with Windows 11 + Ubuntu WSL2 installed, Nvidia GPU is recognized with "**nvidia-smi**" command

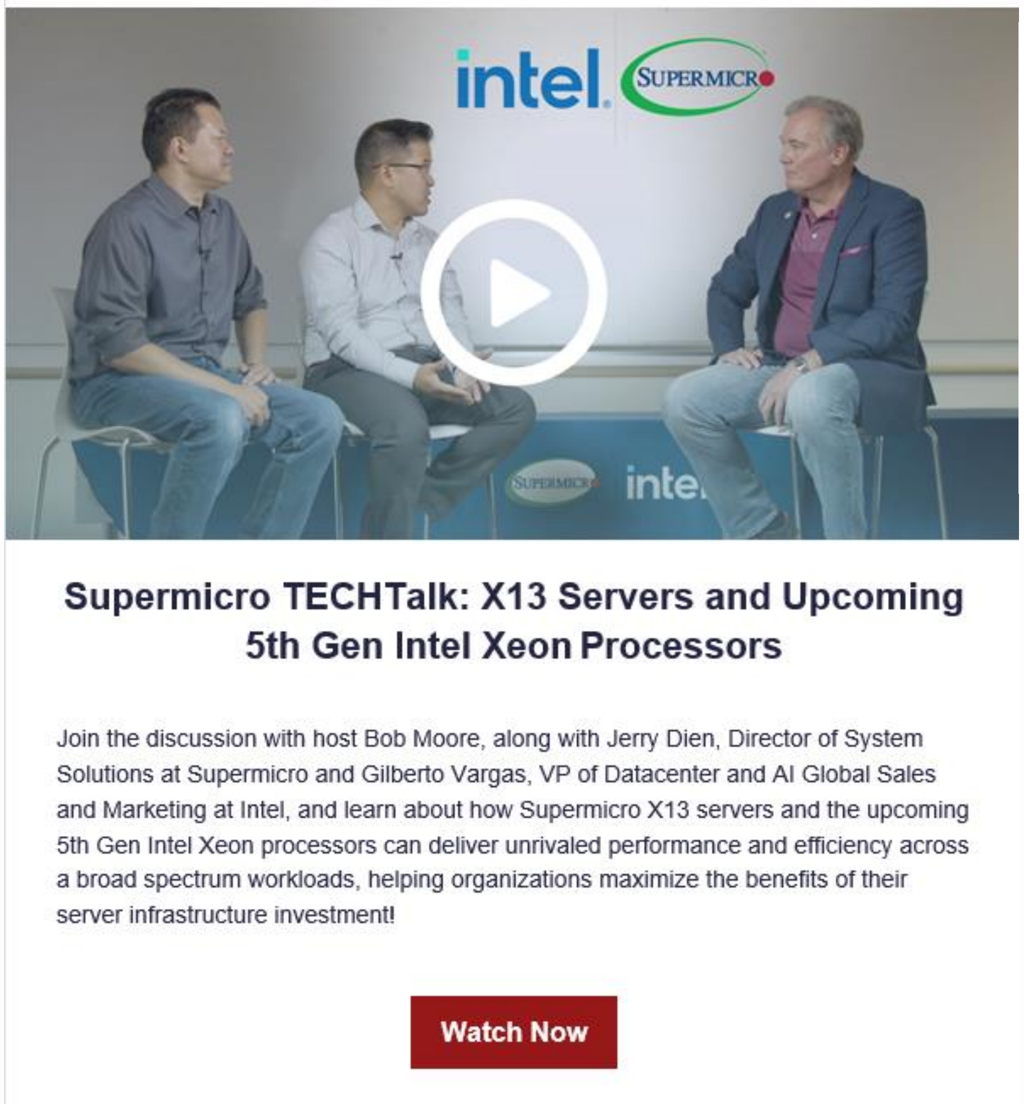


CUDA on Windows WSL2: https://developer.nvidia.com/blog/announcing-cuda-on-windows-subsystem-for-linux-2/

Intel Gen 5 Xeon SP - EMR

TECH Talk

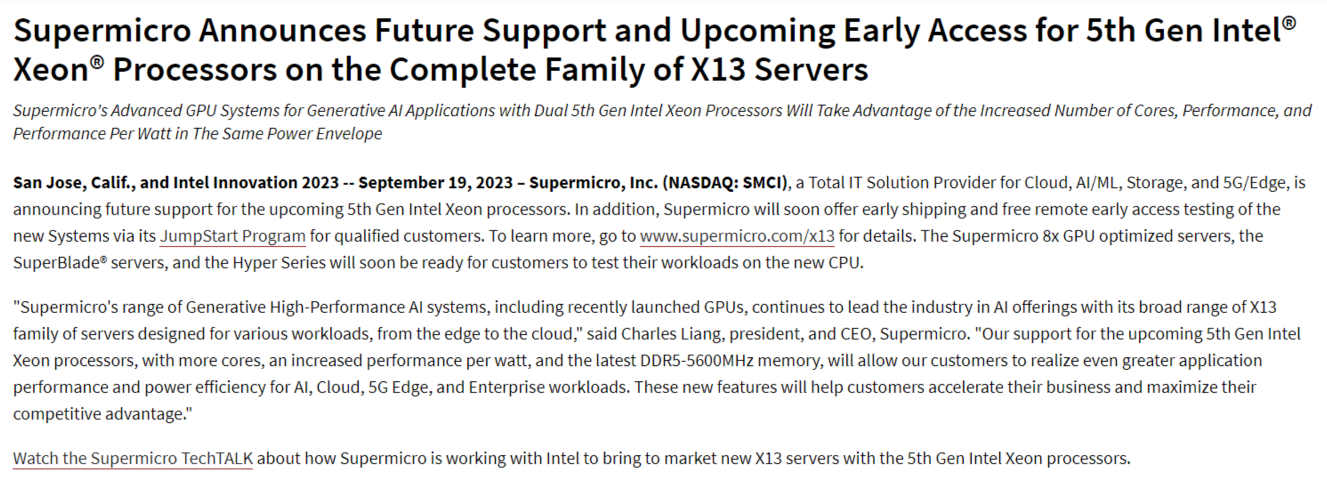
Confidential



Supermicro TECHTalk: X13 Servers and Upcoming 5th Gen Intel Xeon Processors

Join the discussion with host Bob Moore, along with Jerry Dien, Director of System Solutions at Supermicro and Gilberto Vargas, VP of Datacenter and AI Global Sales and Marketing at Intel, and learn about how Supermicro X13 servers and the upcoming 5th Gen Intel Xeon processors can deliver unrivaled performance and efficiency across a broad spectrum workloads, helping organizations maximize the benefits of their server infrastructure investment!

Watch Now



Supermicro Announces Future Support and Upcoming Early Access for 5th Gen Intel® Xeon® Processors on the Complete Family of X13 Servers

Supermicro's Advanced GPU Systems for Generative AI Applications with Dual 5th Gen Intel Xeon Processors Will Take Advantage of the Increased Number of Cores, Performance, and Performance Per Watt in The Same Power Envelope

San Jose, Calif., and Intel Innovation 2023 -- September 19, 2023 – Supermicro, Inc. (NASDAQ: SMCI), a Total IT Solution Provider for Cloud, AI/ML, Storage, and 5G/Edge, is announcing future support for the upcoming 5th Gen Intel Xeon processors. In addition, Supermicro will soon offer early shipping and free remote early access testing of the new Systems via its JumpStart Program for qualified customers. To learn more, go to www.supermicro.com/x13 for details. The Supermicro 8x GPU optimized servers, the SuperBlade® servers, and the Hyper Series will soon be ready for customers to test their workloads on the new CPU.

"Supermicro's range of Generative High-Performance AI systems, including recently launched GPUs, continues to lead the industry in AI offerings with its broad range of X13 family of servers designed for various workloads, from the edge to the cloud," said Charles Liang, president, and CEO, Supermicro. "Our support for the upcoming 5th Gen Intel Xeon processors, with more cores, an increased performance per watt, and the latest DDR5-5600MHz memory, will allow our customers to realize even greater application performance and power efficiency for AI, Cloud, 5G Edge, and Enterprise workloads. These new features will help customers accelerate their business and maximize their competitive advantage."

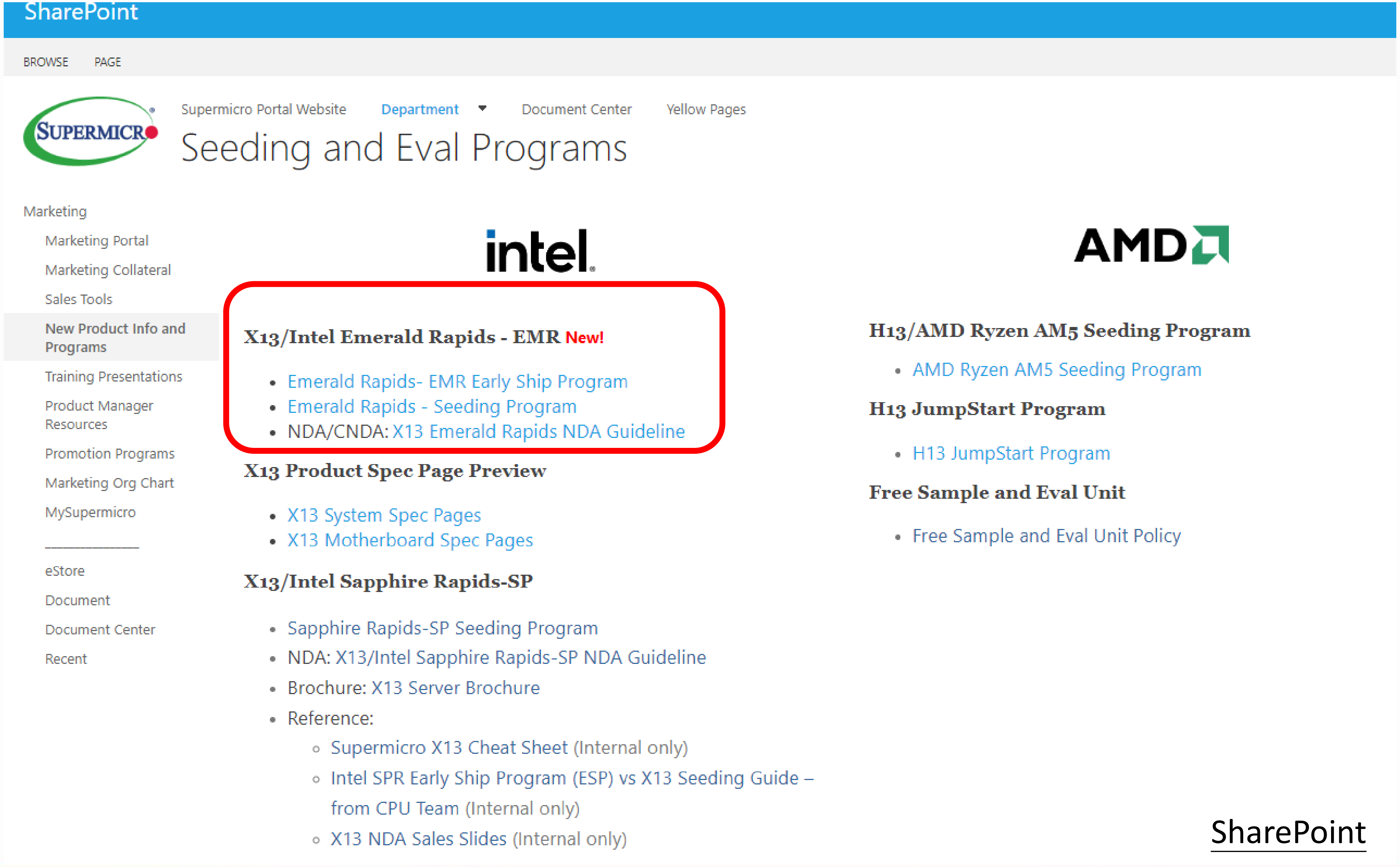
Watch the Supermicro TechTALK about how Supermicro is working with Intel to bring to market new X13 servers with the 5th Gen Intel Xeon processors.

Supermicro's Expansive X13 Server Portfolio
Coming Soon with the 5th Gen Intel® Xeon® Processors

Supermicro TECHTalk – X13 Servers and Upcoming 5th Gen Intel® Xeon® Processors - YouTube

Early Shipment / Seeding Programme

Confidential



Seeding and Eval Programs

X13/Intel Emerald Rapids - EMR New!

- Emerald Rapids- EMR Early Ship Program
- Emerald Rapids - Seeding Program
- NDA/CNDA: X13 Emerald Rapids NDA Guideline

X13 Product Spec Page Preview

- X13 System Spec Pages
- X13 Motherboard Spec Pages

X13/Intel Sapphire Rapids-SP

- Sapphire Rapids-SP Seeding Program
- NDA: X13/Intel Sapphire Rapids-SP NDA Guideline
- Brochure: X13 Server Brochure
- Reference:
  - Supermicro X13 Cheat Sheet (Internal only)
  - Intel SPR Early Ship Program (ESP) vs X13 Seeding Guide – from CPU Team (Internal only)
  - X13 NDA Sales Slides (Internal only)

H13/AMD Ryzen AM5 Seeding Program

- AMD Ryzen AM5 Seeding Program

H13 JumpStart Program

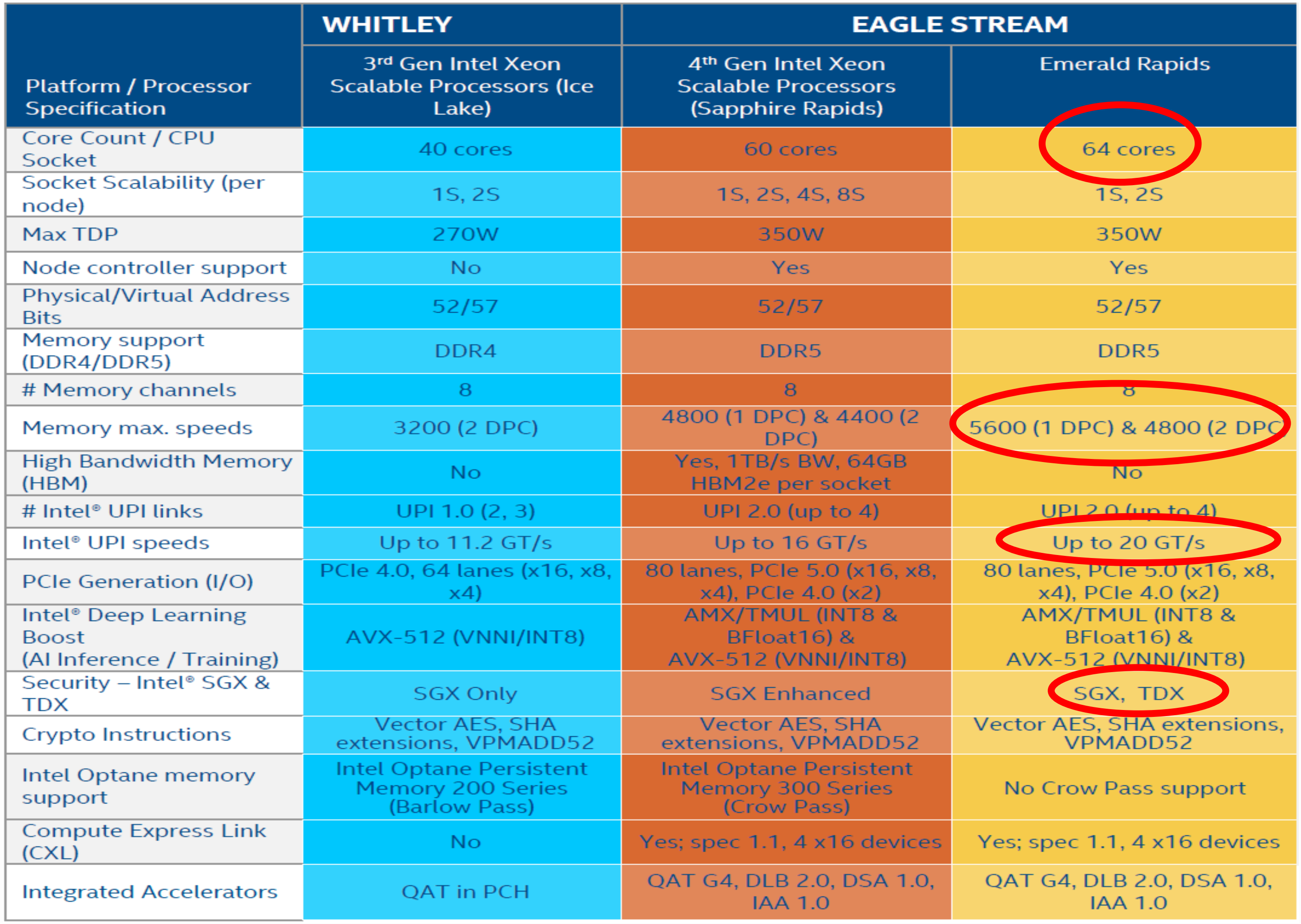
- H13 JumpStart Program

Free Sample and Eval Unit

- Free Sample and Eval Unit Policy

SharePoint

Intel 5th Gen Xeon SP – Emerald Rapids



Platform / Processor Specification	WHITLEY	EAGLE STREAM	
	3rd Gen Intel Xeon Scalable Processors (Ice Lake)	4th Gen Intel Xeon Scalable Processors (Sapphire Rapids)	Emerald Rapids
Core Count / CPU Socket	40 cores	60 cores	64 cores
Socket Scalability (per node)	1S, 2S	1S, 2S, 4S, 8S	1S, 2S
Max TDP	270W	350W	350W
Node controller support	No	Yes	Yes
Physical/Virtual Address Bits	52/57	52/57	52/57
Memory support (DDR4/DDR5)	DDR4	DDR5	DDR5
# Memory channels	8	8	8
Memory max. speeds	3200 (2 DPC)	4800 (1 DPC) & 4400 (2 DPC)	5600 (1 DPC) & 4800 (2 DPC
High Bandwidth Memory (HBM)	No	Yes, 1TB/s BW, 64GB HBM2e per socket	No
# Intel® UPI links	UPI 1.0 (2, 3)	UPI 2.0 (up to 4)	UPI 2.0 (up to 4)
Intel® UPI speeds	Up to 11.2 GT/s	Up to 16 GT/s	Up to 20 GT/s
PCIe Generation (I/O)	PCIe 4.0, 64 lanes (x16, x8, x4)	80 lanes, PCIe 5.0 (x16, x8, x4), PCIe 4.0 (x2)	80 lanes, PCIe 5.0 (x16, x8, x4), PCIe 4.0 (x2)
Intel® Deep Learning Boost (AI Inference / Training)	AVX-512 (VNNI/INT8)	AMX/TMUL (INT8 & BFloat16) & AVX-512 (VNNI/INT8)	AMX/TMUL (INT8 & BFloat16) & AVX-512 (VNNI/INT8)
Security – Intel® SGX & TDX	SGX Only	SGX Enhanced	SGX, TDX
Crypto Instructions	Vector AES, SHA extensions, VPMADD52	Vector AES, SHA extensions, VPMADD52	Vector AES, SHA extensions, VPMADD52
Intel Optane memory support	Intel Optane Persistent Memory 200 Series (Barlow Pass)	Intel Optane Persistent Memory 300 Series (Crow Pass)	No Crow Pass support
Compute Express Link (CXL)	No	Yes; spec 1.1, 4 x16 devices	Yes; spec 1.1, 4 x16 devices
Integrated Accelerators	QAT in PCH	QAT G4, DLB 2.0, DSA 1.0, IAA 1.0	QAT G4, DLB 2.0, DSA 1.0, IAA 1.0

- **BIOS update to be compatible with current X13 MBD**
- **Intel On Demand support**
- **Enhanced Security for VMs workloads with Trust Domain Extension (TDX)**

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