X12/H12 GPU Systems Roadmap

Alok k Srivastav, Sr Solutions Manager

November 2021

Agenda

- What's New
- Why is AI booming now
- Industry trends of AI/ML & HPC
- GPU/Coprocessors/Accelerator Status/Types
- Supermicro Next Generation X12 and H12 based GPU Servers

X12 GPU System Roadmap

Highest Performance and Flexibility for AL/ML and HPC Applications

4U GPU Systems

X12: 4U-8GPU SYS-420GP-TNAR/+

Integrated Performance, Delta GPU

X12: 4U-10GPU SYS-420GP-TNR

Dual Root Configuration, PCIe GPU

X12: 4U-4GPU SYS-740GP-TNRT

Flexible Solution, PCIe GPU

1U/2U GPU Systems

X12: 2U-4GPU SYS – 220GQ-TNAR+

Scale-able Performance, Redstone GPU

X12: 2U-6GPU SYS-220GP-TNR

Balanced Solution, PCIe GPU

X12: 1U-4GPU SYS-120GQ-TNRT

Highest Density, PCIe GPU

X12: 2U- 2Node 3GPU * SYS-210GP-DNR

Flexible Architecture, PCIe GPU

* Subject to change

HGX A100 8-GPU (Delta) Server: SYS-420GP-TNAR/+

4U NVIDIA SXM A100 + 8-GPU Intel® Xeon® Scalable CPU System

System Front

System Rear

- **Key Features**
  - Supports 8 A100 40GB/80GB SXM4 GPUs
  - Platform with NVIDIA® NVLINK™ + NVIDIA® NVSwitch™
  - Dual 3rd Gen Intel® Xeon® Scalable Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

Specifications

CPU – Dual Socket Dual 3rd Gen Intel® Xeon® Scalable Processors Upto 270W TDP	**Memory –** 32 DIMM Slots 32 DIMMs, up to 8TB Registered ECC DDR4 3200MHz SDRAM
Drives – 6 Hot-Swap Bays 6 NVMe U.2 (4 from PCIe Switch & 2 from CPU) 2 NVMe M.2 (Internal) (Option for up to 10 hot-swap U.2 NVMe 2.5“ available)	**Expansion –** 10 PCI-E Slots 8 PCIe 4.0 x16 LP from PCIe switch 2 PCIe 4.0 x16 LP from CPUs AIOM support
I/O ports 1 BMC LAN port 1 VGA port 2 USB 3.0 ports	**Power Supply –** N+N Redundant 4x 3000W High-efficiency (Titanium level) power supply OR 4x 2200W High-efficiency (Titanium level) power supply (3+1)

Subject to change without notice

NVIDIA HGX A100 4GPU (Redstone) Server: SYS-220GQ-TNAR+

2U NVIDIA HGX A100 4GPU Intel® Xeon® Scalable CPU System

System Front

System Rear

- **Key Features**
  - Supports NVIDIA® HGX A100 40GB/80GB 4GPU
  - Direct connect PCI-E Gen 4 Platform with NVIDIA® NVLink™
  - Dual 3rd Gen Intel® Xeon® Scalable Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

Specifications

CPU – Dual Socket Dual 3rd Gen Intel® Xeon® Scalable Processors Up to 270W TDP	**Memory –** 32 DIMM Slots 32 DIMMs, up to 8TB Registered ECC DDR4 3200MHz SDRAM
Drives – 4 Hot-Swap Bays 4x 2.5" SATA 2x M.2 NVMe/SATA	**Expansion –** 4 PCI-E Slots 4x PCI-E Gen 4 x16 LP
Networking – Dual 10GbE 2x RJ45 10GbE 1x RJ45 1GbE IPMI	**Power Supply –** N+N Redundant 2x 3000W Titanium Level

Data Center PCIe GPU Systems

	A100	A30	A40 / RTX A6000	A10	T4
Design	Highest Performance Compute AI, HPC, Data Analytics	Mainstream Compute AI Inference	Highest Performance Graphics Visual Computing Rendering CloudXR	Mainstream Graphics Online Services 4K Cloud Gaming Video with AI	Inference Edge, Lower Entry Cost
Form Factor	x16 PCIe Gen4 2 Slot FHFL 3 NVLink bridges	x16 PCIe Gen4 2 Slot FHFL 1 NVLink bridges	x16 PCIe Gen4 2 Slot FHFL 1 NVLink bridges	x16 PCIe Gen4 1 Slot FHFL	x16 PCIe Gen3 1 Slot LP
Memory	40 GB HBM2	24 GB HBM2	48 GB GDDR6	24 GB GDDR6	16 GB GDDR6
MIG	Up to 7	Up to 4	N/A	N/A	N/A
Media Acceleration	1 JPEG Decoder 5 Video Decoder	1 JPEG Decoder 4 Video Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode) 4x DP 1.4	1 Video Encoder 2 Video Decoder (+AV1 decode)	1 Video Encoder 2 Video Decoder
RT Core	No	No	Yes	Yes	Yes
FP64	Yes	Yes	No	No	No
TDP	250W	165W	300W	150W	70W
Graphics			Best	Better	Good
DL/Compute	Ultimate	Fastest	Fastest	Faster	Fast

Intel DP 4U 10 GPU System: SYS-420GP-TNR

4U NVIDIA 10- PCIe GPU Intel® Xeon® Scalable CPU System

System Front View

System Rear View

- **Key Features**
  - Supports Upto 10 Double Width PCIe GPUs
  - Dual 3rd Gen Intel® Xeon® Scalable Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning (HPC)
  - Cloud rendering
  - Real-time high quality multi-GPU ray tracing
  - High performance simulation of complex 3D

Specifications

CPU – Dual Socket Dual 3rd Gen Intel® Xeon® Scalable Processors Upto 270W TDP	**Memory –** 32 DIMM Slots 32 DIMMs, up to 8TB Registered ECC DDR4 3200MHz SDRAM
Drives – 24 Hot-Swap Bays 16x HOT SWAP 2.5” SATA/SAS 2x M.2 NVMe 8x HOT SWAP 2.5” NVMe	**Expansion –** 12 PCI-E Slots 12 PCIe 4.0 x16 (FHFL) 10 PCIe GPUs Double Width FHFL AIOM support
I/O ports 1 BMC LAN port 1 VGA port 2 USB 3.0 ports	**Power Supply –** N+N Redundant 4x 2000W High-efficiency (Titanium level) power supply

Subject to change without notice

Intel DP 4U 4 GPU System: SYS-740GP-TNRT

4U NVIDIA 4- PCIe GPU Intel® Xeon® Scalable CPU System

System Rear View

- **Key Features**
  - Supports Upto 4 Double Width GPUs
  - Dual 3rd Gen Intel® Xeon® Scalable Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning (HPC
  - Real-time high quality multi-GPU ray tracing
  - High performance simulation of complex 3D

Specifications

CPU – Dual Socket Dual 3rd Gen Intel® Xeon® Scalable Processors Upto 270W TDP	**Memory –** 16 DIMM Slots 16 DIMMs, up to 4TB Registered ECC DDR4 3200MHz SDRAM
Drives – 8 Hot-swap 3.5" drive bays Up to 8 NVMe drives (4 NVMe drives supported by default) Support 2x M.2 (SATA or NVMe).	**Expansion –** 7 PCI-E Slots 6 PCI-E Gen 4.0 x16 (4 FHFL & 2 LP) 1 PCI-E 4.0 x8 LP
I/O ports Dual 10GbE ports 1 BMC LAN port 1 VGA port • 6 USB 3.0 ports	**Power Supply –** N+N Redundant 2x 2200W High-efficiency (Titanium level) power supply

Subject to change without notice

Intel DP 1U 4 GPU System: SYS-120GQ-TNRT

1U NVIDIA 4- PCIe GPU Intel® Xeon® Scalable CPU System

System Front View

System Rear View

- **Key Features**
  - Supports Upto 4 Double Width GPUs
  - Dual 3rd Gen Intel® Xeon® Scalable Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

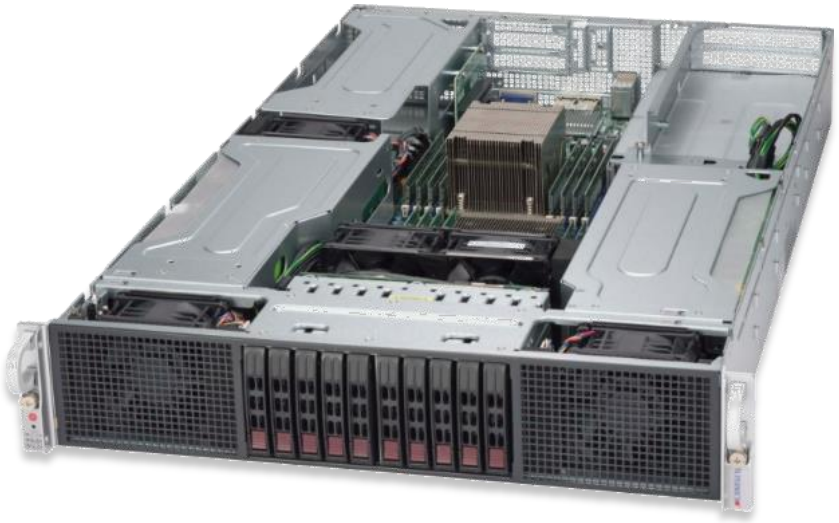
Specifications

CPU – Dual Socket Dual 3rd Gen Intel® Xeon® Scalable Processors Upto 205W TDP	**Memory –** 16 DIMM Slots 16 DIMMs, up to 4TB Registered ECC DDR4 3200MHz SDRAM
Drives – 4 x 2.5" drive bays 1x M.2 NVMe Supported (Internal) 2x 2.5" NVMe Hot-swap drive bays 2x 2.5" SATA Internal Fixed drive bays	**Expansion –** 6 PCI-E Slots 6 PCIe 4.0 x16 (4 FHFL & 2 LP)
I/O ports 1 BMC LAN port 1 VGA port 2 USB 3.0 ports	**Power Supply –** N+N Redundant 2x 2000W High-efficiency (Titanium level) power supply

Subject to change without notice

Intel DP 2U 6 GPU System: SYS-220GP-TNR

2U NVIDIA 6- PCIe GPU Intel® Xeon® Scalable CPU System



System Rear View

System Front View

Subject to change without notice

- **Key Features**
  - Supports Upto 6 Double Width GPUs
  - Dual 3rd Gen Intel® Xeon® Scalable Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning (HPC)
  - Virtual Work station
  - Video Conferencing
  - 4K Cloud Games

Specifications

CPU – Dual Socket Dual 3rd Gen Intel® Xeon® Scalable Processors Upto 270W TDP	**Memory –** 16 DIMM Slots 16 DIMMs, up to 4TB Registered ECC DDR4 3200MHz SDRAM
Drives – 10 x 2.5" drive bays Up to10x 2.5" drive bays Up to 6 NVMe drives 2x M.2	**Expansion –** 8 PCI-E Slots 6 PCIe 4.0 x16 FHFL 2 PCIe 4.0 x8 LP AIOM support
I/O ports 1 BMC LAN port 1 VGA port 2 USB 3.0 ports	**Power Supply –** N+N Redundant 2x 2600W High-efficiency (Titanium level) power supply

Intel UP 2U 2Node GPU System: SYS-210GP-DNR

2U NVIDIA 3- PCIe GPU Per Node Intel® Xeon® Scalable CPU System

System Front View

System Rear View

Key Features

UP 3rd Gen Intel® Xeon® Scalable Processor

Energy-efficient Resource-saving Architecture

Key Applications

Perfect Platform for Video Streaming

High-End Cloud Gaming

Specifications

CPU – Single Socket (Per Node) Single 3rd Gen Intel® Xeon® Scalable Processors Upto 270W TDP	**Memory –** 8 DIMM Slots (Per Node) 8 DIMMs, up to 2TB Registered ECC DDR4 3200MHz SDRAM
Drives – (Per Node) 2 U.2 NMVe Gen 4.0 (Hot-swap drive bay) 2 M.2 NVMe Gen 3.0 (On-board from PCH) 2 SATA DOM for internal OS drive	**Expansion –** 3 PCI-E Slots (Per Node) Up to 3 Double Width FHFL AIOM support
I/O ports (Per Node) BMC LAN port 1 VGA port 2 USB 3.0 ports and 1 COM port	**Power Supply (Enclosure)** 2x 2600W High-efficiency (Titanium level) power supply

H12 GPU System Roadmap

Highest Performance and Flexibility for AL/ML and HPC Applications

SXM GPU Systems

H12: 2U-4GPU AS -2124GQ-NART

Scale-able Performance, Redstone GPU

H12: 4U-8GPU AS -4124GO-NART

Integrated Performance, Delta GPU

PCIe GPU Systems

H12: 2U- 2Node 3GPU AS –2114GT-DNR

Flexible Architecture, PCIe GPU

H12: 4U-8GPU AS -4124GS-TNR

Direct Attach & Low Latency, PCIe GPU

HGX A100 8-GPU (Delta) Server: AS -4124GO-NART(NART+)

4U NVIDIA SXM A100 + 8-GPU AMD EPYC CPU System

System Front

System Rear

- **Key Features**
  - Supports 8 A100 40GB SXM4 GPUs
  - Platform with NVIDIA® NVLINK™ + NVIDIA® NVSwitch™
  - Dual AMD EPYC™ Series Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

Specifications

CPU – Dual Socket Dual AMD EPYC™ 7002/7003 Series (Rome/Milan) Processors Up to 128 Cores, CPU TDP up to 280W	**Memory –** 32 DIMM Slots 32 DIMMs, up to 8TB Registered ECC DDR4 3200MHz SDRAM
Drives – 6 Hot-Swap Bays 6 NVMe U.2 (4 from PCIe Switch & 2 from CPU) 2 NVMe M.2 (Internal) (Option for up to 10 hot-swap U.2 NVMe 2.5“ available)	**Expansion –** 10 PCI-E Slots 8 PCIe 4.0 x16 LP from PCIe switch 1 PCIe 4.0 x16 & 1 PCIe 4.0 x8 LP from CPUs AIOM support
I/O ports 1 VGA port 2 USB 3.0 ports	**Power Supply –** N+N Redundant 4x 3000W High-efficiency (Titanium level) power supply OR 4x 2200W High-efficiency (Titanium level) power supply (3+1)

Subject to change without notice

HGX A100 4-GPU (Redstone) Server: AS -2124GQ-NART/+

2U NVIDIA SXM A100 + 4-GPU AMD EPYC CPU System

System Front

System Rear

- **Key Features**
  - Supports 4 A100 40GB SXM4 GPUs
  - Direct connect PCI-E Gen 4 Platform with NVIDIA® NVLink™
  - Dual AMD EPYC™ Series Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

Specifications

CPU – Dual Socket Dual AMD EPYC™ 7002/7003 Series (Rome/Milan) Processors Up to 128 Cores, CPU TDP up to 280W	**Memory –** 32 DIMM Slots 32 DIMMs, up to 8TB Registered ECC DDR4 3200MHz SDRAM
Drives – 4 Hot-Swap Bays 4x 2.5" SAS/SATA/NVMe Hybrid	**Expansion –** 5 PCI-E Slots 4x PCI-E Gen 4 x16 LP 1x PCI-E Gen 4 x8 LP
Networking – Dual 10GbE 2x RJ45 10GbE 1x RJ45 1GbE IPMI	**Power Supply –** N+N Redundant 2x 2200W Titanium Level 2x 3000W Titanium Level

Subject to change without notice

DP 4U 8 GPU Server: AS -4124GS-TNR

4U PCIe Gen 4 AMD EPYC CPU System

System Front

System Rear

- **Key Features**
  - Supports 8 PCIe GPUs Double Width FHFL
  - Dual AMD EPYC™ Series Processors
- **Key Applications**
  - AI Compute/Model Training/Deep Learning
  - High-performance Computing (HPC)

Specifications

CPU – Dual Socket Dual AMD EPYC™ 7002 /7003Series (Rome/Milan) Processors Up to 128 Cores, CPU TDP up to 280W	**Memory –** 32 DIMM Slots 32 DIMMs, up to 8TB Registered ECC DDR4 3200MHz SDRAM
Drives – 24 Hot-Swap Bays 24 Hot-swap 2.5" drive bays 4x 2.5" SATA drives 4x 2.5" NVMe drives	**Expansion –** 11 PCI-E Slots 8 PCIe 4.0 x16 FHFL 2 PCIe 4.0 x8 (1 FHFL & 2 LP) AIOM support
I/O ports 11 BMC LAN port 1 VGA port 2 USB 3.0 ports	**Power Supply –** N+N Redundant 4x 2000W High-efficiency (Titanium level) power supply

Subject to change without notice

A+ UP 2U 2Node 3 GPU Server

AS –2114GT-DNR

Front View

Rear View

Key Features:

- 3 Direct Attached GPUs
- PCIe Gen 4.0 AMD Radeon Instinct & Nvidia Enterprise GPUs
- Flexible Architecture
- 4x Hot-swap 8cm counter-rotating cooling fans

Specifications

1	**Processor Support** Single EPYC processor (Socket SP3), Up to 64 cores **280W TDP**
2	**Memory Capacity** 8 DIMM, up to 2TB DDR4 3200 MHz Reg. ECC
3	**GPU Support slots per node** **3** Double Width GPUs in PCI-E 4.0 x16 slots or **6** Single Width GPUs in PCI-E 4.0 x16 slots
4	**Expansion slots per node** **6 PCI-E 4.0 x16** FH slots, 1 **AIOM** slot / OCP3.0
5	**Networking & I/O – per node** 1 Flexible networking via **AIOM** 1 RJ45 Dedicated IPMI LAN port 2 USB 3.0 ports 1 VGA port
6	**System Management** IPMI ASPEED AST2600 BMC with dedicated LAN port
7	**Drive Bays per node** 4 Hot-swap PCI-E Gen4 x4 U.2 drives (2 default at front, 2 optional at rear) 2 PCI-E Gen4 x4 M.2 connectors
8	**System Cooling** 4x Heavy duty 8cm PWM fans
9	**Power Supply** 2600W 1+ 1 High-efficiency (Titanium level, 96%) supplies
10	**Dimensions** 17.25" (W) x 3.47" (H) x 29.9" (D)

New Available SXM4 GPU Liquid Cooling System

Redstone System

SYS-220GQ-TNAR+
AS-2124GQ-NART/+

GPU Node:
4x Nvidia SXM4 400W/500W GPU
CPU Node:
2x AMD EPYC 280W CPU /Loop

Delta System

SYS-420GP-TNAR/+
AS-4124GO-NART/+

GPU Node:
8x Nvidia SXM4 400W/500W GPU
CPU Node:
2x AMD EPYC 280W CPU /Loop
2x Intel ICX 270W CPU /Loop

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