X13 Universal GPU

Multi-Architecture Flexibility with Future-Proof Open-Standards-Based Design

Ultimate modularity and customization options for AI and HPC environments

- **Dual 4th Gen Intel® Xeon® Scalable processors**
- **Support for the latest industry standards including PCIe 5.0, DDR5 and Compute Express Link (CXL) 1.1**
- **Innovative modular architecture designed for flexibility and futureproofing with a range of form factors from 4U to 8U**
- **Supports next-generation GPUs including NVIDIA H100 and Intel Data Center GPU Max Series**
- **Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling**
- **PCIe 5.0 x16 networking slots and up to 16 U.2 NVMe drive bays**

Open-Standards-Based Platform

Supermicro X13 Universal GPU systems feature an open, modular, standards-based architecture designed for maximum flexibility. Support for multiple industry-standard GPUs allows organizations to take advantage of different GPU configurations based on workload while only deploying a single server architecture, reducing infrastructure complexity and simplifying future upgrades. Designed for serviceability with hot-swappable, tool-less components in a modular construction, the chassis design is optimized for thermal capacity.

Designed for Demanding HPC and AI Workloads

The Supermicro Universal GPU platform has been designed from the ground up to support a combination of CPU and GPU configurations, allowing customization for specific HPC and AI workloads within the data center using a single platform.

- High-performance computing including energy, molecular dynamics, physics, computational chemistry and climate sciences
- Deep Learning for image and video detection/recognition, life sciences & drug discovery, autonomous driving and robotics

Supports Industry-Standard GPU Form Factors

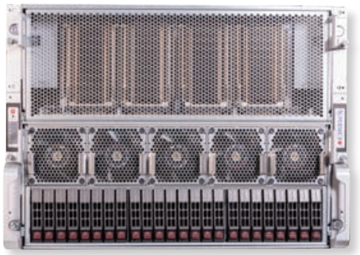
The Supermicro Universal GPU platform is designed to work with a wide range of GPUs based on an open standards design. By adhering to an agreed-upon set of hardware design standards, such as Universal Baseboard (UBB) and OCP Accelerator Modules (OAM), IT administrators can choose the GPU architecture best suited for their HPC or AI workloads. Additionally, support for GPU interconnects including the latest NVIDIA NVLink® v4.0 with a total bandwidth of 900 GB/s facilitates ultra-fast GPU-to-GPU communication, reducing bottlenecks caused by traditional GPU interlinks.

High Performance Networking and Storage

Extraordinary data throughput from next-generation GPUs and CPUs is nothing without networking and storage to match. All Universal GPU systems support high-speed networking up to 400Gb/s directly to each GPU via PCIe 5.0 x16 slots with additional OCP 3.0 compliant AIOM options for DPUs and other accelerator cards. For storage, systems support up to 16 front-accessible hot-swap U.2 NVMe drive bays. There is also an additional front I/O variation available for selected models for enhanced maintenance and management.

AI Accelerated with 4th Gen Intel Xeon Scalable Processors

The latest 4th Gen Intel Xeon Scalable processors include built-in accelerator engines optimized for AI and HPC workloads. The purpose-built Intel Advanced Matrix Extensions (Intel AMX) accelerator improves the performance of deep learning workloads to deliver robust AI capabilities for AI training and inference.



Universal GPU	SYS-821GE-TNHR/FTNHR	SYS-521GU-TNXR	SYS-421GU-TNXR
Processor Support	Dual Socket E (LGA-4677) 4th Gen Intel® Xeon® Scalable processors[†]	Dual Socket E (LGA-4677) 4th Gen Intel® Xeon® Scalable processors[†]	Dual Socket E (LGA-4677) 4th Gen Intel® Xeon® Scalable processors[†]
Outstanding Features	Highest GPU communication using NVIDIA® NVLINK™ + NVIDIA® NVSwitch™ High density 8U system with NVIDIA® HGX™ H100 8-GPU 8 NVMe for GPU direct storage 8 NIC for GPU direct RDMA (1:1 GPU Ratio) 2 M.2 NVMe for boot drive only	Highest GPU communication using NVIDIA® NVLINK™ High density 5U Universal GPU system with NVIDIA® HGX™ H100 4-GPU 8 NICs for high bandwidth networking	Highest GPU communication using NVIDIA® NVLINK™ High density 4U Universal GPU system with NVIDIA® HGX™ H100 4-GPU 8 NICs for high bandwidth networking
Memory Slots & Capacity	32 DIMM slots Up to 8TB: 32x 256 GB DRAM 4800MHz ECC DDR5	32 DIMM slots Up to 8TB: 32x 256 GB DRAM 4800MHz ECC DDR5	32 DIMM slots Up to 8TB: 32x 256 GB DRAM 4800MHz ECC DDR5
GPU Support	HGX H100 8-GPU SXM5 Multi-GPU Board NVIDIA® NVLink™ with NVSwitch™	HGX H100 4-GPU SXM5 Multi-GPU Board NVIDIA® NVLink™	HGX H100 4-GPU SXM5 Multi-GPU Board NVIDIA® NVLink™
I/O Ports	2x 10GbE BaseT with Intel® X550-AT2 (optional) 2x 25GbE SFP28 with Broadcom® BCM57414 (optional) 2x 10GbE BaseT with Intel® X710-AT2 (optional) 1 VGA Port	2x 10GbE RJ45 port(s) with Intel® Ethernet Controller X550-AT2 1 VGA Port	2x 10GbE RJ45 port(s) with Intel® Ethernet Controller X550-AT2 1 VGA Port
Motherboard	X13DEG-OA	X13DGU	X13DGU
Form Factor	8U Rackmount Enclosure: 437 x 355.6 x 843.28mm (17.2" x 14" x 33.2") Package: 698 x 750 x 1300mm (27.5" x 29.5" x 51.2")	5U Rackmount Enclosure: 449 x 222.5 x 833mm (17.67" x 8.75" x 32.79") Package: 700 x 370 x 1260mm (27.55" x 14.57" x 49.6")	4U Rackmount Enclosure: 449 x 175.6 x 833mm (17.67" x 6.91" x 32.79") Package: 700 x 370 x 1260mm (27.55" x 14.57" x 49.6")
Expansion Slots	8 PCIe 5.0 x16 LP slot(s) 2 PCIe 5.0 x16 FHFL slot(s)	10 PCIe 5.0 X16 LP Slots	8 PCIe 5.0 X16 LP Slots

Drive Bays	20x 2.5" hot-swap NVMe/SATA drive bays;
	12x 2.5" NVMe dedicated;
		10x 2.5" hot-swap NVMe/SATA drive bays;
		10x 2.5" NVMe hybrid;
			6x 2.5" hot-swap NVMe/SATA drive bays;
			10x 2.5" NVMe hybrid;

Cooling	10 heavy duty fan(s)	5 heavy duty fan(s)	5 heavy duty fan(s)
Power	6x 3000W (3+3) Redundant Power Supplies, Titanium Level Optional: 8x 3000W (4+4) Redundant Power Supplies, Titanium Level	4x 3000W Redundant Power Supplies, Titanium Level	4x 3000W Redundant Power Supplies, Titanium Level

[†] Supports up to 350W TDP CPUs (Aircooled). CPUs with high TDP supported under specific conditions. Contact Technical Support for details.