PRODUCT BRIEF

SUPERMICRO UNIVERSAL GPU SYSTEM DELIVERS MULTI-ARCHITECTURE FLEXIBILITY AND FUTURE PROOF OPEN-STANDARDS DESIGN FOR AI, TRAINING, AND HPC ENVIRONMENTS

Simplifies Customer Deployments, Supports All Major CPU, GPU, and Fabric Architectures in the Highest Performance Configurations

Universal GPU – 4U | Universal GPU – 5U | Universal GPU w/NVIDIA HGX A100 4-GPU | Universal GPU w/AMD Instinct™ MI250 OAM Accelerator

Supermicro Universal GPU Options

TABLE OF CONTENTS

Executive Summary

Modern data centers require the flexibility to respond to changing workload requirements, even in specialized environments such as High-Performance Computing (HPC) and Artificial Intelligence (AI). While specialized configurations of servers can be created that contain specific CPUs or GPUs, a reduction in costs can be achieved when a standard architecture is designed to accommodate a range of CPUs, GPUs, communication paths, and networking options. Based on open standards, the Supermicro Universal GPU is a revolutionary product design that accommodates various components and delivers a state-of-the-art GPU system for demanding applications. The balance between CPUs, GPUs, storage, and networking in a single enclosure is the key to maximizing performance. The Supermicro Universal GPU system has been

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

designed to maximize the performance of applications that require CPUs, GPUs, networking, and storage, all working seamlessly together and avoiding common bottlenecks that many server designs do not address.

The Supermicro Universal GPU is a modular system that has been designed for maximum flexibility, utilizing several standards-based technologies. With the innovations in HPC and AI hardware accelerators, a new system that could easily accommodate the accelerator innovations was created to easily accept the most recent and future HPC and AI hardware. Furthermore, with its innovative design, the Supermicro Universal GPU will accommodate current and subsequent generations of CPUs and GPUs in an air-cooled environment, reducing capital expenditures.

Future-Proofing

With the increasing power requirements of next-generation CPUs and GPUs, combined with the need for more high-performance and low latency networking and storage devices, having a system design that will be able to accommodate future products will reduce costs. This will also allow for quicker adoption of new technologies when available, eliminating data center redesign and forklift upgrades to the existing infrastructure. By designing a chassis that can accommodate the highest performing (and highest wattage) components today and working with future generations of components easily, costs will be reduced, leading to faster deliveries and a faster-to-market. Additionally, multiple generations of CPUs and GPUs will be able to use the Supermicro Universal GPU chassis.

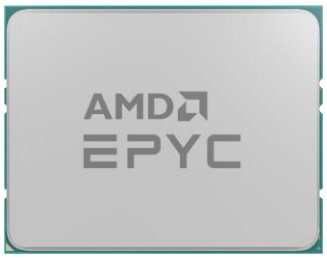
Workloads Require Flexibility

The Supermicro Universal GPU is designed for a number of demanding compute environments in HPC and AI within a data center. As a key component of a modern data center infrastructure, whether located in an on-premise data center or in a public cloud environment, the Supermicro Universal GPU is designed to excel at workloads that require the latest CPU and GPU technology. The combination of the fastest and most powerful CPUs from Intel and AMD, together with the latest advances in GPU technology, is in high demand as HPC and AI workflows are integrated into enterprises of all sizes. Different combinations of CPUs and GPUs will deliver the best results for a given workload, as applications can be tuned to specific hardware capabilities. However, various workloads may require different combinations in an enterprise data center. Thus a single product that can be managed easily and contains various components will be ideal for many computing environments. In addition, future-proofing a system like the Supermicro Universal GPU to accommodate yet to be announced components will allow for faster and more efficient deliveries as new technology becomes available.

CPU Choices

Today's most powerful servers are designed to house either the 3rd Gen Intel® Xeon® Scalable processors or the 3rd Gen AMD EPYC™ processors. Different organizations require various CPUs yet may also mix and match within a data center, typically to match a workload with a CPU architecture. The Universal GPU server gives users the choice of a powerful server with the latest CPUs configured in a dual-socket system. With up to 128 cores, applications will have the compute performance to execute

applications quickly. By designing a system that can easily accommodate different CPUs with minimal changes to the server, the right mix of servers can be installed in a data center for specific workloads, both current and future.



GPU Choices

Multiple GPUs families can be installed in the Supermicro Universal GPU. While AMD and NVIDIA's leading accelerator products can be installed today, the system has also been designed to accommodate future accelerators. For example, this new Supermicro server will support Intel's data center GPU, codenamed Ponte Vecchio accelerator, in the future.

- The NVIDIA HGX A100 4-GPU accelerators can be paired with either the 3rd Gen Intel Xeon 83xx Series or the 3rd Gen AMD 7003/7003X processors. The NVIDIA HGX A100 4-GPU accelerators are designed for data centers supporting AI, data analytics, and high-performance computing applications. The NVIDIA A100 performs approximately 20X [1] faster than the previous generation of NVIDIA GPUs. The A100 can be efficiently scaled up to be partitioned into seven isolated GPU instances with Multi-Instance (MIG). This provides a unified platform for elastic data centers. Several different math precisions are supported, which enables acceleration for any workload.

- The AMD MI250 is based on the 2nd Gen AMD CDNA architecture and is constructed from multi-die modules and the 3rd Gen AMD Infinity Architecture. AMD CDNA™ 2 architecture is the industry's first multi-chip GPU module. This GPU is designed for the most complex scientific computing and machine learning applications. This architecture powers the AMD Instinct MI250 OAM Accelerator product that targets solutions ranging from single compact systems to enterprise and research large-scale implementations. The AMD Infinity Architecture empowers system builders and cloud architects alike to unleash the very latest in server performance without sacrificing power, manageability, or the ability to help secure their organization's most important assets, its data.

GPU SDKs

Software Development Kits are critical for getting the maximum performance from a GPU. The following SDKs are supported on the Supermicro Universal GPU servers:

- The NVIDIA HPC SDK C, C++, and Fortran compilers support GPU acceleration of HPC modeling and simulation applications with standard C++ and Fortran, OpenACC® directives, and CUDA®. GPU-accelerated math libraries maximize performance on common HPC algorithms, and optimized communications libraries enable standards-based multi-GPU and scalable systems programming. Performance profiling and debugging tools simplify porting and optimization of HPC applications, and containerization tools enable easy deployment on-premises or in the cloud. The HPC SDK provides the tools needed to build NVIDIA GPU-accelerated HPC applications.

- AMD ROCm™ is an open software platform allowing researchers to tap the power of AMD Instinct™ accelerators to drive scientific discoveries. The ROCm platform is built on the foundation of open portability, supporting environments across multiple accelerator vendors and architectures. With ROCm 5.0, AMD extends its platform powering top HPC and AI applications with AMD Instinct MI200 series accelerators, increasing accessibility of ROCm for developers and delivering outstanding performance across critical workloads. In addition, various precision-based math is available, which enables acceleration for a wide range of workloads.

Below is a comparison of the initial Supermicro Universal GPU Systems

System	Height	CPU	Max Memory	GPU	GPUs/System	Fabric	Power Supplies
A+ Server 4124GQ-TNMI	4U or 5U	AMD 7003/7003X Series	8TB DDR4-3200GHz	AMD Instinct MI250 OAM	4	xGMI or PCI-E	4x 3000W (2+2) Titanium Level efficiency power supplies
A+ Server 4124GQ-TNMI	4U or 5U	AMD 7003/7003X Series	8TB DDR4-3200GHz	NVIDIA HGX A100 4-GPU	4	NVLink or PCI-E	
SYS-420GU-TNXR	4U or 5U	Intel 83xx Processors	8TB DDR4-3200MHz or 12TB with Intel® Optane® Persistent Memory	NVIDIA HGX A100 4-GPU	4	NVLink or PCI-E	

Table 1 - Supermicro Universal GPU Product Summary

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), OCP Accelerator Modules (OAM), and PCI-E and platform-specific interfaces, IT administrators can choose the GPU architecture best suited for their HPC or AI workloads. This will meet the demanding requirements of many enterprises and will simplify the installation, testing, production, and upgrades of GPU solutions. In addition, IT administrators will easily choose the right combination of CPUs and GPUs to create the most optimal system for their users.

- Universal Baseboard (UBB) [2] - The Universal Baseboard is designed to be modular and flexible in supporting current and future OAM modules and providing maximum design flexibility for many conceivable system designs. The UBB supports 8 OAM modules, but the board has been engineered to support a wide range of interconnecting fabrics and

topologies, power domains, TDPs, cooling solutions, and scale-out options. While the board is optimized for a few standard configurations and released OAM modules, great care was taken to accommodate future trends and customer needs.

- OCP Accelerator Modules (OAM) - The OCP Accelerator Module (OAM) specification [3] defines the form factor and interconnects for an open-hardware compute accelerator module. Facebook (now Meta) worked [4,5] with partners within the Open Compute Project (OCP) community and developed mezzanine-based OCP Accelerator Modules (OAM) as a standard form factor for different types of hardware accelerator solutions can follow. Facebook (now Meta) was an original contributor and participant in the OAM specification.

Image 1 – Example of an OAM Module

GPU Fabrics and Universal GPU System Layouts

The system architecture for the Supermicro Universal GPU servers allows for a very fast inclusion of various CPUs and GPUs in a single, consistent manner. This innovative design gives IT administrators a uniform interface to high-powered GPU servers installed in a data center. The key to the Supermicro Universal GPU server is that different GPU fabrics can be used, depending on the application requirements. In addition, the ability to handle the following two high-speed fabrics or PCI-E based GPUs will also be accommodated in this system, which is a first for the industry.

- NVIDIA® NVLink® is a high-speed, direct GPU-to-GPU interconnect. NVIDIA NVLink technology addresses interconnect issues by providing higher bandwidth, more links, and improved scalability for multi-GPU system configurations. For example, a single NVIDIA A100 Tensor Core GPU supports up to 12 third-generation NVLink connections for a total bandwidth of 600 gigabytes per second (GB/sec)—almost 10X the bandwidth of PCIe Gen 4.

- The AMD Infinity Architecture is the GPU to GPU fabric for the AMD MI250 accelerator. The AMD Infinity Architecture, introduced with the 2nd Gen AMD EPYC™ Processors, empowers system builders and cloud architects alike to unleash the very latest in server performance. Calculations as of SEP 18th, 2021. AMD Instinct™ MI250 built on AMD CDNA™ 2 technology accelerators support AMD Infinity Fabric™ technology providing up to 100 GB/s peak total aggregate theoretical transport data GPU peer-to-peer (P2P) bandwidth per AMD Infinity Fabric link, and include up to eight links providing up to 800GB/s peak aggregate theoretical GPU (P2P) transport rate bandwidth performance per GPU OAM card for 800 GB/s.

Below are the block diagrams for the Supermicro Universal GPU server with an AMD CPU and AMD MI250 accelerators and the similar server with an Intel CPU and OAM modules. Note how easy it will be to incorporate future CPUs or GPUs with this system architecture.

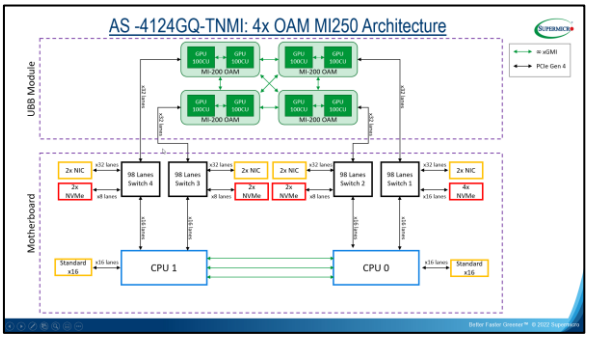


Image 2 - AS-4124GQ-TNMI Block Diagram

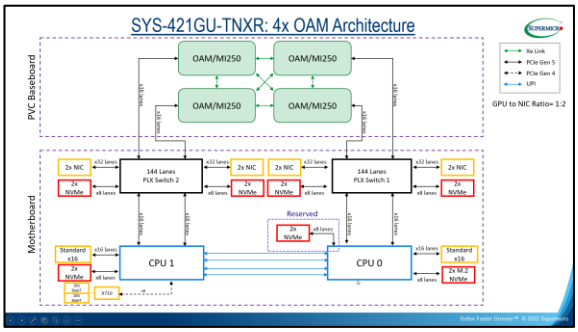


Image 3 - SYS-421GU-TXNR Block Diagram

The Supermicro Universal GPU servers will also accommodate GPUs that are not interconnected, and rely on the PCI-E communications path to communicate with other GPUs when needed. These PCI-E based versions of the Supermicro Universal GPU will accommodate up to 10 GPUs, which communicate through the PCI-E bus.

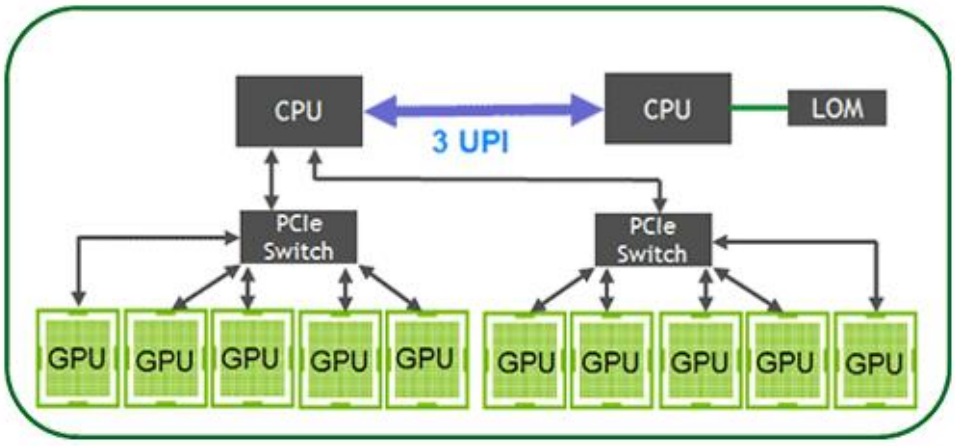


Image 4 - Universal GPU (intel CPUs) with PCI-E Switch and GPUs

Optimal System - Storage, Networking, Power, and Cooling mix for top performance configurations

The Supermicro Universal GPU incorporates the optimal networking, storage, power, and cooling mix to support the highest performing and efficient system configurations. This includes:

- Up to 10 High-Speed NVMe U.2 storage drives connected to the PCI-E Switch or 10 High Speed 2.5" SATA Drives.
- Power Supplies up to 4x 3000W (2+2) with up to Titanium Level efficiency
- Expansion up to 8 PCI-E x16 LP via a PLX Switch
- Standard I/O ports

- Additional 2x PCI-E 4.0 x16 LP slots in the 5U chassis.
- Support for 2 OCP 3.0 AIOM Modules Direct From CPU in the 5U chassis.

The features were selected to optimize the GPU performance and provide upside for future requirements. For example, the four AMD Instinct MI-250 OAM modules are comprised of two GPU chips (GCDs or Graphics Compute Dies) each, so the 8 PCI-E expansion slots can be used to provide dedicated networking for each GPU chip. In addition, with four 3000W power supplies, the systems can support today's 500W GPU cards as well as future GPU cards that will require up to or more than 700W.

Supermicro Universal GPU 4U/5U Options

The 4U or 5U Universal GPU server will be available for accelerators that use the UBB standard and PCI-E 4.0 and soon PCI-E 5.0. In addition, 32 DIMM slots of memory and a wide range of storage and networking options are available, which can also be connected using the PCI-E standard. The Supermicro Universal GPU server can accommodate GPUs using baseboards in the SXM or OAM form factors that utilize very high-speed GPU to GPU interconnects such as NVIDIA NVLink or the AMD xGMI Infinity fabric, or directly connect GPUs via a PCI-E slot. All major current CPU and GPU platforms will be supported, giving customers choices that match their exact workloads.

The Universal GPU is available in two chassis, one that is 4U in height the other is 5U in height. The 5U model can accommodate CPUs above today's highest TDP values and GPUs up to 700 TDP, extending the Universal GPU to handle the cooling demands as next-generation CPUs and GPUs emerge. The 5U version, besides accommodating the highest CPU and GPU TDPs, also incorporates a 1U expansion module, which includes support for 2 OCP 3.0 AIOM modules directly accessible from the CPU, which can increase networking performance.

Image 5 - Supermicro Universal GPU - 4U and 5U Chassis Options

Summary

The Supermicro Universal GPU server is designed with flexibility for organizations needing various CPU and GPU combinations. Different CPUs, GPUs, and the type of fabric within the same architecture can be specified. These systems are designed with enough power and cooling capacity to accommodate the next generation of CPUs or GPUs. In addition, significant amounts of memory can be addressed for large, complex applications.

References

1 - https://www.nvidia.com/en-us/data-center/a100/

2 - https://www.opencompute.org/documents/universal-baseboard-design-specification-v0p42-pdf

3 - https://www.opencompute.org/documents/ocp-accelerator-module-design-specification-v1p0-3-pdf

4 - https://engineering.fb.com/2019/03/14/data-center-engineering/accelerator-modules/

5 - https://www.servethehome.com/facebook-ocp-accelerator-module-oam-launched/

Additional Information

Supermicro Universal GPUs - https://www.supermicro.com/en/products/universal-gpu

Supermicro All GPUs – https://www.supermicro.com/gpu