WHITE PAPER

SUPERMICRO H14 SERVERS: UPGRADE YOUR DATA CENTER TO DELIVER MAXIMUM PERFORMANCE

Industry Leading Performance and Power Efficiency with Supermicro Rack-Scale Solutions Featuring AMD EPYC™ 9005 Series Processors

TABLE OF CONTENTS

Introduction to High-Performance Data Centers

Data centers are changing with the emergence of AI as part of an enterprise computing strategy. At the same time, the entire organization's needs require a range of server technologies. These servers and their associated performance need to be carefully chosen to meet the SLAs for a wide range of uses while, at the same time, reducing power usage. With the new AMD EPYC 9005 series processors, the performance of many workloads and the work per watt will also increase.

The modern data center must be both highly performant and energy efficient. Massive amounts of data are generated at the edge and then analyzed in the data center. New CPU technologies are constantly being developed to analyze data, determine the best course of action, and speed up the time to understand the world and make better decisions.

With the digital transformation continuing, a wide range of data acquisition, storage, and computing systems continue to evolve with each generation of new CPUs. The latest generations of CPUs continue to innovate within their core computational units and the technology to communicate with memory, storage devices, networking, and accelerators.

Servers and, by default, the CPUs within the servers form a continuum of computing and I/O power. The combination of cores, clock rates, memory access, path width, and performance contribute to specific servers for workloads. In addition, the server that houses the CPUs may take different form factors and be used when the environment where the server is placed has airflow or power restrictions. A key for a server manufacturer to be able to address a wide range of applications is to use a building block approach to designing new systems. In this way, a range of systems can be simultaneously released in many form factors, each tailored to the operating environment.

The Supermicro H14 products have been designed with the following components:

1. Broad Selection — The Supermicro H14 product line offers a wide range of choices optimized for specific workloads.

2. Compute Power—The Supermicro H14 products with the AMD EPYC™ 9005 series processors offer top-level performance for many metrics and, when combined with high core counts, are ideal for a range of workloads.

3. Max Core Counts—Supermicro H14 servers have been designed to house AMD's most powerful and energy-intensive CPUs for high-end computing environments.

4. Thermal Design—By optimizing the airflow within a system, high-performing CPUs can be used without concern for overheating. Liquid cooling increases the compute density and lowers the data center PUE.

Maximum Performance, Efficiency, and Density

Supermicro H14 servers with AMD EPYC™ 9005 processors deliver maximum performance (192 "Zen 5c" core, 384 threads) for dense, high core count and frequency. Better efficiency and high density compute.

Broadest Portfolio and Flexible options with liquid cooling at rack-scale

Supermicro's Building Block architectures allow custom configuration of expansion, acceleration, networking, and storage supporting AMD's EPYC 9004 and 9005 series processors to deliver the broadest portfolio of Enterprise, Cloud, and AI solutions powered by AMD EPYC CPUs.

Industry-Leading Time-to-Deployment

Supermicro's large-scale production and integration capacity is up to 5,000+ racks per month globally (including 1,350+ liquid-cooled racks), from design to validation and delivery in weeks rather than months. The new H14 Supermicro product line, based on 5th Gen AMD EPYC™ CPUs, supports a broad spectrum of workloads and excels at helping a business achieve its goals, which can be summarized:

- Best business outcomes across industries and workloads
- Highest performance x86 server processor
- Leadership x86 energy efficiency
- Assurance of confidential computing
- A significant ecosystem of solutions

While the performance of CPUs continues to grow and can quickly meet many enterprise computing requirements, certain domains, such as HPC and AI, require technologies that work in parallel and software stacks that can take advantage of thousands of computing elements to work together. These applications require the maximum number of CPU cores working together and specialized accelerators that have been designed for a smaller class of applications. Fast internal networking between the components and state-of-the-art communication between systems allows innovative organizations to explore new algorithms while minimizing power usage and, thus, costs.

Supermicro designs and manufactures a wide range of servers and storage systems deployed from the Edge to hyperscale data centers. Different form factors with varying amounts of CPUs, memory capacity, storage types, capacity, and environmental considerations are engineered and delivered by Supermicro. The key to offering many different systems is advanced engineering and teaming up with leading-edge CPU manufacturers, such as AMD.

As CPUs run faster, with more cores, more heat is generated. Supermicro designs systems that efficiently remove this heat, lowering cooling costs and allowing CPUs to run up to their maximum thermal design power (TDP). With a design philosophy that enables customers to upgrade individual components, whether CPUs, RAM, storage, or I/O devices, users can choose to replace only what needs to be updated, reducing E-Waste while using the latest and most efficient components.

AI workloads require optimized systems incorporating the proper hardware and tuning software to deliver maximum performance at a given price point. A solution must contain a choice of CPUs, GPUs, and the proper software stack to provide value to end users. Various aspects, such as the number of cores, communication latency between cores, GHz, and the generation of CPU architectures, can influence the benchmark performance of real-world AI applications.

In this white paper, we look in-depth at Supermicro's latest H14 portfolio of servers and how these systems help organizations thrive in today's digital landscape.

Wide Range of Products for Varying Workloads

Supermicro's customers span many industries, with some common objectives:

- Ability to meet Service Level Agreements (SLAs) – Whether servicing employees or end-user customers, the CPU and I/O systems' responses are expected to fall within a specific time range.
- Provide new services to customers – As customers demand new services, which may run partially on edge devices as "apps," organizations must set up the back-end infrastructure to handle and respond to more data and processing than ever before.
- Reduce costs with more powerful systems – Some workloads do not increase at the same rate as new processors' computational and I/O power. Therefore, new CPUs allow them to reduce costs by assigning more work to lesser systems for these organizations.

- Enable new insights – Using the latest CPU designs, scientists, engineers, and data analytics professionals can gain new insights and simulate physical systems more accurately.

Various workloads are all addressed by the Supermicro H14 servers and storage systems. These include:

- **High-Performance Computing (HPC)** – HPC systems are used by more than university and national lab researchers. Enterprises integrate HPC systems into everyday workflows to bring products to market faster or discover new vaccines and drugs. HPC systems require fast cores, large amounts of memory, and fast networking between systems.
- **Cloud** – Designing and implementing a cloud solution requires a wide range of optimized products for different workloads, not just for environments where the price performance of the compute aspect is most important. Storage and networking are critical for a productive and cost-effective cloud data center.
- **Artificial Intelligence (AI)** – Systems with fast CPUs and associated GPU sub-systems are required for the growing AI use cases. Supermicro H14 servers can house up to 10 GPUs in a 5U rack height and excel at AI applications, enabling faster training and inference applications. Supermicro designs servers specifically to accommodate a high number of GPUs for maximum AI application performance. In addition, the Supermicro GPU servers incorporate the latest GPUs from several vendors in various form factors.
- **Big-Data Analysis** – As the volume of data generated everywhere explodes, the systems must access, analyze, and present structured and unstructured data to the user. These tasks require the ability to hold an increasing amount of data in memory, fast computation, and quick data communication to GPUs if needed.
- **Virtualization** – With many enterprises utilizing virtualization technologies to get higher utilization from existing servers, the new Supermicro H14 servers, with the 5th Gen AMD EPYC processors, allow for higher-powered virtualization machines, as more cores are available and faster CPUs.
- **Enterprise** – Typical enterprise workloads will benefit from the new Supermicro H14 systems with increased performance and reduced costs. In addition, existing workloads will execute faster, using less power than previous generations of Supermicro servers.

How 5th Generation AMD EPYC™ Processors Enhance Workloads and CPU Highlights

While the performance of computing systems continues to increase over time with AMD's innovations, different workloads require this new performance, while other workloads benefit from the lower cost per unit of work. For example, while the performance of CPUs increases, typical Enterprise workloads (HR, ERP, Inventory Control, etc.) mainly do not require the performance gains from generation to generation but rather benefit from assigning more work to a given CPU. New Enterprise workloads, such as analytics, video conferencing, and application delivery, require performance improvements to take advantage of the new 5th Gen AMD EPYC processors' new performance

levels. HPC and AI require both increased core numbers, increased GHz, and parallelization and networking outside the system.

- Cores: Maximum of 192 cores with the "Zen 5c" core, compared to 64 cores in the 3rd Gen AMD EPYC processors. There is a maximum of 128 cores (Zen 5).
- Faster communication: PCIe 5.0 is 2X faster than the previous generations of CPUs with PCIe 4.0.
- Addressable memory: 5th Gen AMD EPYC™ processors can address up to 9TB of DRAM per socket, significantly more than the 2nd or 3rd generation of AMD EPYC servers.
- Memory performance: The new AMD processors utilize DDR5-6000 MT/s memory, almost twice as fast as the H12 Generation of processors.
- Faster communication between CPUs: Compared to the 2nd Gen AMD EPYC processors, 5th Gen AMD EPYC processors have more and faster Infinity Fabric interconnects.
- AI Acceleration – 5th Gen AMD EPYC processors include support for the AVX512 instructions.
- Security by design is a set of state-of-the-art security features that help keep data secure, whether in use, flight, or rest.

5TH GENERATION AMD EPYC PROCESSOR DESCRIPTION

Next Generation Server Architecture – AMD EPYC™ 9005 Series CPUs are raising the bar for workload performance and helping IT professionals everywhere excel.

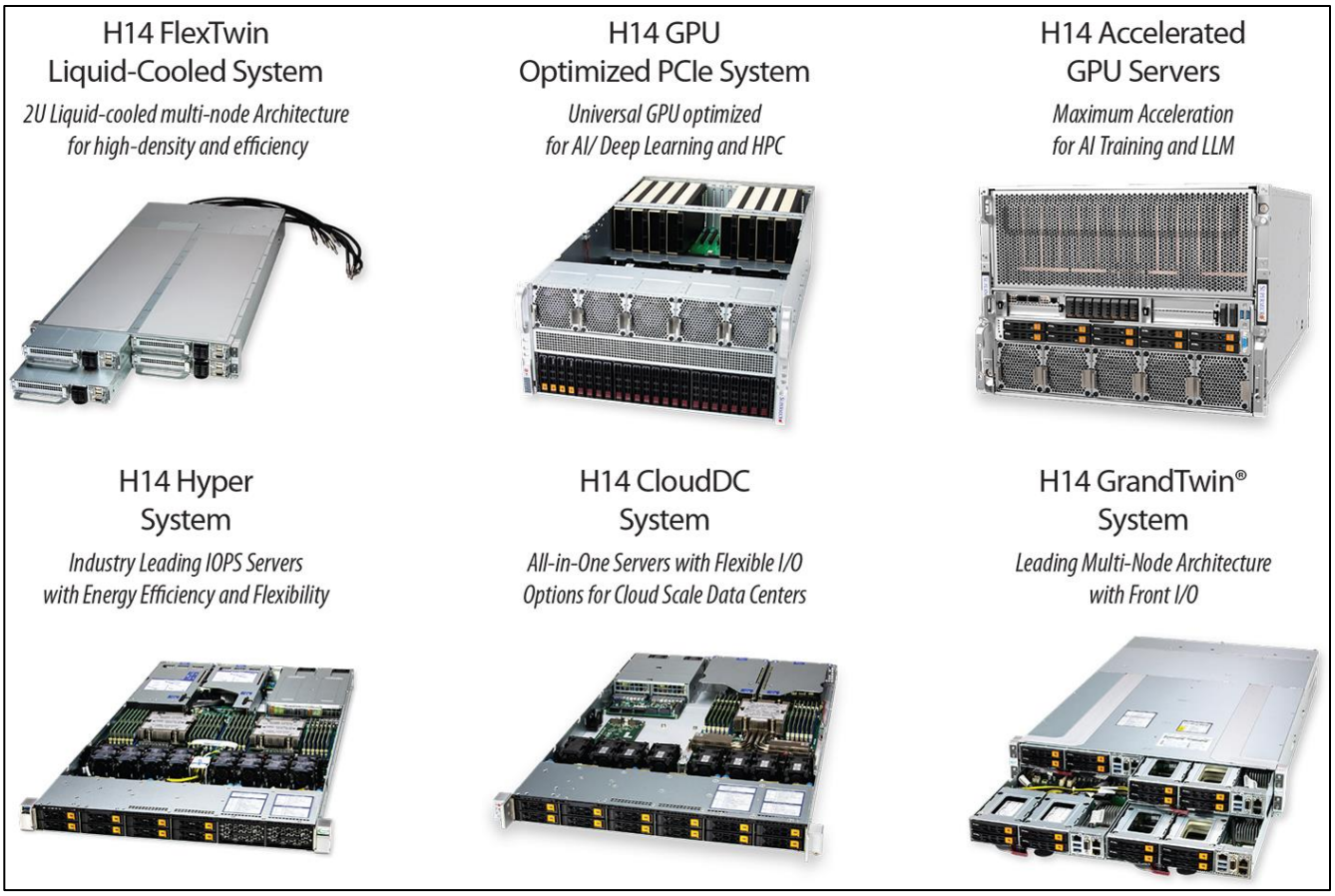
Performance + Efficiency are the new metrics for success in IT. Servers powered by EPYC 9005 CPUs can deliver faster time to results, helping provide more and better insights for decision making driving better business outcomes. AMD EPYC CPUs – performant, efficient, on time.

Efficient: With AMD EPYC 9005 CPUs, IT professionals can use fewer servers to get the job done compared with EPYC 7003 CPUs.

Latest Technology: The AMD EPYC 9005 Series CPUs amplify the AMD history of x86 architecture innovations and record-breaking performance with next-generation 5nm technology as well as introducing support for high-performant DDR5 DIMMs and fast PCIe Gen 5 I/O. EPYC 9005 CPUs support 12 memory channels with 2 DIMMs/channel capability, delivering the resources needed for memory-hungry AI, ML, HPC, and large in-memory computations. These AMD EPYC CPUs also uniquely provide 128 PCIe5 lanes in a 1-socket server and up to an astounding 160 PCIe5 lanes in 2-socket servers. This enables the high-performant demands of today's AI and ML applications and the increasing use of accelerators, GPUs, FPGAs, and high-capacity LAN cards natively with 5th Gen EPYC CPUs' high PCIe5 lane counts.

There are several advantages to using the 5th Gen AMD EPYC™ processors for different workloads with different models for various workloads. The various models can be categorized as follows:

Supermicro's Wide Range of Product Lines Featuring the 5th Gen AMD EPYC™ CPUs



The Supermicro AMD product family contains many servers designed for customer workloads. All of the systems take advantage of the new features and capabilities of the 5th Gen AMD EPYC processors. The Supermicro product line can be segmented into the following areas. This white paper will look more closely at the following product lines:

Figure 1 - Supermicro FlexTwin (TM)

FlexTwin - Supermicro H14 FlexTwin is a new platform designed for maximum performance density and serviceability in a multi-node architecture, featuring support for the latest CPU, memory, storage, and cooling technologies. Purpose-built to support demanding HPC workloads, including financial services, manufacturing, scientific research, and complex modeling, H14 FlexTwin can be customized to suit specific HPC applications and customer requirements thanks to Supermicro's modular Building Block architecture.

Common workloads include: HPC Data Center • Financial Services • Manufacturing • Climate & Weather Modeling • Oil & Gas • Scientific Research

GrandTwin® – The Supermicro GrandTwin is an innovative system that puts multiple independent servers within the same enclosure. This feature lowers operating expenses by allowing the use of shared resources, such as the 2U enclosure, heavy-duty fans, backplane, and N+1 power supplies. Supermicro's GrandTwin® family of servers is purpose-built for single-processor performance, with front-serviceable hot-swap nodes allowing easier installation and servicing in space-constrained environments. Powered by AMD EPYC processors, the GrandTwin architecture delivers high performance in a modular design that can be optimized for a wide range of applications. Supermicro's Resource Saving Architecture delivers improved power efficiency and lower materials costs thanks to shared components, including power and cooling.

Figure 2 - Supermicro GrandTwin(TM)

Common workloads include: Diskless HPC • All-Flash HCI • Hybrid Cloud • All-Flash NVMe Storage • High-Performance File Systems • Software-Defined Storage.

GPU Family of Servers – The Supermicro GPU family of servers excels at HPC and AI applications. Systems have been designed to house multiple GPUs in a single server so applications can process data rapidly. While many Supermicro server lines can accommodate one or two GPUs, the GPU family extends the quantity of GPUs in a single server up to 10 in a 4U form factor. In addition, the GPU family of servers can house multiple GPUs and is designed so that GPUs can efficiently communicate with each other, allowing GPU systems to bypass internal communication paths for faster results. The GPU systems can also address the maximum memory that the 5th Gen AMD EPYC can accommodate, up to 9TB per socket.

a. **GPU with OAM/HGX** – With Supermicro's advanced architecture and thermal design, including liquid cooling and custom heatsinks, the GPU systems feature either the AMD Instinct™ MI325X or NVIDIA B200 GPUs, can deliver up to 6x AI training performance and 7x inference workload capacity and highest density in a flexible 8U or 10U system. The H14 GPU systems feature the latest technology stacks, with up to 400G networking, NVIDIA NVLink and NVSwitch, 1:1 GPUDirect RDMA, GPUDirect Storage, and NVMe-oF on InfiniBand.

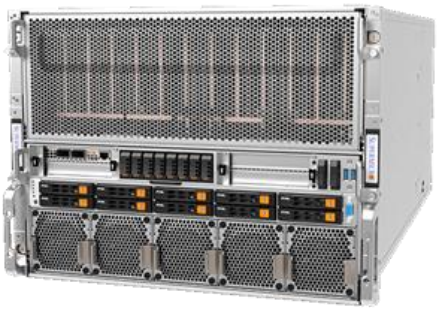


Figure 3 – 8U GPU Server with Dual AMD EPYC 9005 CPUs and 8x AMD Instinct MI325X GPUs

Figure 4 - 10U GPU Server with Dual AMD EPYC 9005 CPUs and 8x NVIDIA HGX B200 GPUs

Common workloads include:

AI/ML • Deep Learning Training and Inference • High-performance Computing (HPC) • Building Block for Scalable AI Infrastructure

For the Supermicro GPU systems, all of the new features of the 5th Gen AMD EPYC processors will help high-end applications perform better and return faster with the latest GPU systems from Supermicro. More and faster cores, higher bandwidth to the GPUs and other devices, and the ability to address vast amounts of memory are exactly what large HPC and AI applications demand.

b. **GPU System with PCIe GPUs** – The GPU systems that attach the GPU accelerators via the PCIe bus are ideal for environments requiring multiple GPUs that perform their work with direct commands from the CPU. HPC and AI/ML environments will benefit significantly from the 5th Gen AMD EPYC processors. Various platforms can accommodate from one to 10 GPUsCommon workloads include:

AI/ML • Deep Learning Training and Inference • High-performance Computing (HPC) • Rendering Platform for High-end Professional Graphics • Best-in-Class VDI Infrastructure Platform

GPU Systems with PCIe will benefit significantly from the PCIe 5.0 communications bus, the increased number of cores, and up to 9TB of memory that this system can accommodate.

Figure 5 - GPU System w/PCIe GPUs

Hyper Family - The Hyper servers are designed for maximum rackmount flexibility with front I/O for today's data center requirements as a dual-socket server. These systems can handle the maximum wattage of CPUs and the maximum number of DIMMs to accelerate a wide range of workloads. In addition, the Hyper systems sport many PCIe slots (up to eight) for extreme flexibility, are toolless for fast and easy servicing, and come with various

Figure 6 – Hyper 2U Dual Socket Server

storage devices (NVMe/SAS/SATA). The Hyper systems can also support 1 AIOM/OCP 3.0 NIC.

Common workloads for the Hyper family include:

• Enterprise Server • Cloud Computing • Big Data Analytics • Hyperconverged Storage • AI Inference and Machine Learning • Network Function Virtualization

Hyper systems will benefit from the increased core count at pricing similar to that of 5th Gen AMD EPYC processors. In addition, the faster PCIe 5.0 communications bus will give more rapid access to storage devices.

CloudDC Family – The CloudDC family is explicitly designed for cloud data centers where space is premium. The CloudDC product line is toolless, meaning servicing these servers (hot-swapping) is quick and easy. The I/O options vary; the systems can accommodate up to two singles-width GPUs. The CloudDC family can support a single AIOM/OCP 3.0 card, which gives the product family tremendous expandability and flexibility. The CloudDC family also supports up to 2 PCIe 5.0 slots. 12 NVMe storage devices are supported for maximum I/O performance and capacity. The Supermicro H14 CloudDC system is designed with DC-MHS compliance.

Figure 7 - CloudDC 1U Single Socket Server

Common workloads for the CloudDC family include:

Cloud Computing • Web Servers • Hyper-converged Storage • Virtualization • File Servers • Head-node Computing • 5G Telco • AI Inferencing

5th Gen AMD EPYC Compared to Previous CPUs

AMD EPYC™ 9005 SERIES PROCESSORS											
MODEL	CORES	THREADS	BASE FREQ. (GHZ)	UP TO MAX BOOST FREQ. (GHZ)[a]	DEFAULT TDP (W)	L3 CACHE (MB)	DDR5 CHANNELS	UP TO MAX DDR5 FREQ. (1DPC)	PER-SOCKET THEORETICAL MEMORY BANDWIDTH (GB/S)	PCIE® GEN 5 LANES	2P/1P
9755	128	256	2.70	4.10	500	512	12	6000	576	128	2P/1P
9655	96	192	2.60	4.50	400	384	12	6000	576	128	2P/1P
9655P	96	192	2.60	4.50	400	384	12	6000	576	128	1P
9575F	64	128	3.30	5.00	400	256	12	6000	576	128	2P/1P
9555	64	128	3.20	4.40	360	256	12	6000	576	128	2P/1P
9555P	64	128	3.20	4.40	360	256	12	6000	576	128	1P
9355	32	64	3.55	4.40	280	256	12	6000	576	128	2P/1P
9355P	32	64	3.55	4.40	280	256	12	6000	576	128	1P
9135	16	32	3.65	4.30	200	64	12	6000	576	128	2P/1P
9965	192	384	2.25	3.70	500	384	12	6000	576	128	2P/1P

CORE COUNT INCREASE OVER GENERATIONS

The 5th Gen AMD EPYC has more cores than previous generations of AMD EPYC processors. More cores enable faster processing for applications that have been designed to take advantage of multiple cores, or more applications can be run on the same CPU simultaneously.

	3rd Gen AMD EPYC processors	4th Gen AMD EPYC processors	5th Gen AMD EPYC processors	% Increase (3rd Gen to 5th Gen – max)
Number of Cores	Up to 64	Up to 128	Up to 160 (Zen 5) and 192 (Zen 5c)	3X

MEMORY CAPACITY

The 5th Generation AMD EPYC processors increase memory capacity that can be addressed directly per socket. This is due to the increased number of memory channels. A 2-socket system can address 18 Terabytes (TB) of memory (9 TB per socket).

Increased memory allows for more extensive applications to be run in less time. Data analytics, HPC, and more VMs can easily take advantage of this increased memory capacity to deliver results to users faster. By keeping more data in memory than on storage devices, performance is improved, and more extensive and complex simulations or analytics can be executed to gain more in-depth insight.

	3rd Gen AMD EPYC processors (UP)	5th Gen AMD EPYC processors (UP)	% Increase
Memory DIMMs (max/socket)	16	24	50%
Max Memory (DRAM/socket)	4TB	9TB	125%

MEMORY ACCESS PERFORMANCE

The speed at which the CPU can access memory greatly affects the overall execution time of a task. The 3rd Gen has improved memory access bandwidth of up to 3200 Megatransfers per second (MT/s). The faster the MT/s rate, the faster that the CPUs can retrieve data and act on it. The previous generation of AMD processors limit was 3200 MT/s, and eight channels could deliver 8 x 3200 MT/s = 25,600 MT/s. The 4th Gen AMD EPYC uses 12 channels for memory access, thus, the maximum performance per socket = 12 * 4800 MT/s = 57,600 MT/s, a 125% improvement.

	3rd Gen AMD EPYC processors	5th Gen AMD EPYC processors	% Increase
Memory Performance	3200 MHz	6000 MHz	87.5%
Number of Channels	8	12	50%
Total Memory Bandwidth	= 8 * 3200 MHz = 25,600 MT/s	= 12 * 6000 MHz = 72,000 MT/s	181%

FASTER CONNECTIONS TO PERIPHERALS

The 5th Gen AMD EPYC processors supports the PCIe Gen 5 standard, which has a peak performance of twice that of the previous PCIe Gen 4 standard. PCIe Gen 5 delivers 32 GT/second per lane. The performance of a system for communicating with PCIe devices is computed as follows:

PCIe Performance (GT/s/lane) x Number of lanes / 8 (since 1 GigaTransfer = .125 GB)

Thus, for PCIe Gen 5 a system with 16 lanes, the communication can achieve 32 GT/s x 16 lanes / 8 = approximately 64GB/second. The aggregate performance is 2X what PCIe Gen 4 delivers. A faster PCIe bus is critical when using GPUs or FPGAs.

	PCIe 4.0 (3rd Gen)	PCIe 5.0 (5th Gen)	% Increase
Per Lane Performance	16 GigaTransfers/Second	32 GigaTransfers/Second	100%

Supermicro Intelligent Management

SuperCloud Composer is a composable cloud management platform that provides a unified dashboard to administer software-defined data centers. Supermicro's cloud infrastructure management software brings speed, agility, and simplicity to IT administration by integrating data center tasks into a single intelligent management solution. Our robust composer engine can orchestrate cloud workloads through a streamlined industry-standard Redfish API. SuperCloud Composer monitors and manages the broad portfolio of multi-generation Supermicro servers and third-party systems through its data center lifecycle management feature set, which ranges from a single unified console to IT administration by integrating data center tasks into a single intelligent management solution.

Applications Benefits Summary

Applications will benefit from several innovations with the new 5th Gen AMD EPYC™ processors.

- More cores – performance will increase for applications that scale with the number of available cores.
- More extensive memory access – with more memory accessed on the main memory bus, applications will perform better without waiting for data to be retrieved from storage devices.
- Faster memory access – with higher memory bandwidth, applications will execute faster, requiring less time to wait for critical data.
- Faster communication – with PCIe 5.0, applications can communicate with PCI-E devices at twice the speed as before, increasing overall application performance.
- Interconnect between sockets – for applications requiring socket-to-socket communication, the faster xGMI channels will reduce execution time.

How Does Supermicro Do It?

Supermicro incorporates a Building Block® approach, allowing us to design individual components with the latest technology and then engineer these different components into various systems. Using this design process, Supermicro can create many variations, including additional CPUs, the number of memory slots, the number of PCIe lanes, and the number and type of storage devices. Application-optimized systems can quickly develop depending on the form factor, cooling, and memory requirements. Innovative design allows for efficient cooling and the sharing of other mechanical components. Supermicro's servers can accommodate high-end CPUs in various form factors.

Performance / Power over time - Why this is important to data centers

Over time, with AMD's advancement of CPU technology, more computing power is available at a given price and a given amount of energy. AMD has increased the amount of work performed per unit of electricity by a factor of 5 over the past 12 years. This means more work can be performed at a constant power draw, enabling organizations to offer more services and applications to their employees or the public. Below is a chart of the AMD EPYC performance over time using the SPECrate®2017_INT_BASE (Normalized) benchmark and successive generations of AMD EPYC processors.

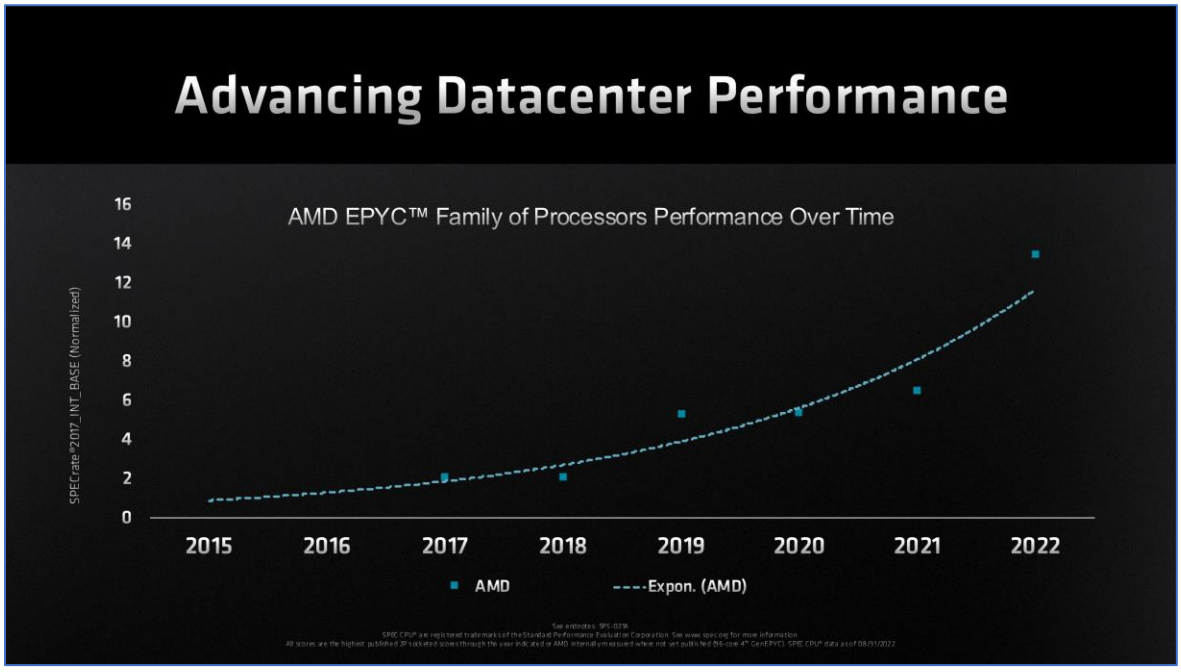


Image Courtesy of AMD

Summary

The new H14 product line from Supermicro enables all organizations to take full advantage of AMD's latest CPUs. Supermicro has a server designed for your workload, ranging from a single processor to the latest in blade technology, from 8 cores to an amazing 192 cores per socket. With the increase in the amount of memory that can be addressed and the performance of the memory sub-system, applications can access more data faster. The increase in core count numbers and clock rates results in a faster time-to-solution and more performance per watt. The Supermicro H14 product lineup is designed for workloads that range from the edge to the data center.

For More Information

www.supermicro.com/aplus