WHITE PAPER

SUPERMICRO H13 SERVERS ENABLE HIGH-PERFORMANCE DATA CENTERS

New Supermicro H13 Servers Utilizing the Latest AMD EPYC™ 9004 Series Processors (codenamed "Bergamo" and "Genoa-X") Give More Capabilities for Increased Performance for Innovative Enterprises

Contents

Introduction to High-Performance Data Centers

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 new 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, and path width and performance contribute to specific servers for workloads. In addition, the server which 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

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

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 new H13 Supermicro product line, based on 4th Gen AMD EPYC™ CPUs, supports a broad spectrum of workloads and excels at helping a business achieve its goals, which are highlighted here:

- 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 all the way 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. To provide value to end users, a solution needs to contain a choice of CPUs, GPUs, and the proper software stack. Various aspects, such as the number of cores, communication latency between cores, GHz, and which generation of CPU architectures, can influence the benchmark performance of real-world AI applications.

In this white paper, we take an in-depth look at Supermicro's latest H13 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:

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

- 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 H13 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 also 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 H13 servers can house up to 10 GPUs in a 4U 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 H13 servers, with the 4th 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 H13 systems with increased performance and reduced costs. In addition, existing workloads will execute faster, using less power than previous generations of Supermicro servers.

How 4th 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,

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

require performance improvements to take advantage of the new 4th Gen AMD EPYC processors' new performance levels. HPC and AI require both increased core numbers, increased GHz, and parallelization and networking outside the system.

- More cores: Maximum of 128 cores compared to 64 cores in the 3rd Gen AMD EPYC processors, with the new 4th Gen AMD EPYC 9004 Series (codename "Bergamo")
- Increased Level 3 Cache: 4th Gen AMD EPYC Processors with AMD 3D V-Cache™ technology (Codename "Genoa-X") increase the performance of technical applications such as FEA, CFD, and EDA, as more cache is closer to the processors. This is due to an L3 cache of up to 1,152MB, a 50% increase per CPU than the previous generation of AMD EPYC processors with AMD 3D V-Cache technology. This results in 4th Gen AMD EPYC processors with 3D V-Cache technology having 3x the amount of L3 cache compared to the highest-end regular 4th Gen AMD EPYC processors. All Supermicro Aplus servers that use the AMD EPYC 9004 Series processors can also use the new CPU with AMD 3D V-Cache™ technology.
- Faster communication: PCIe 5.0 is 2X faster than the previous generation of CPUs with PCIe 4.0.
- Addressable memory: 4th Gen AMD EPYC processors can address up to 6TB of DRAM per socket.
- Memory performance: The new AMD processors can utilize DDR5-4800Mhz memory, 33% faster than previous generations.
- Faster communication between CPUs: More and faster xGMI interconnects are available with the 4th Gen AMD EPYC processors.
- AI Acceleration – 4th Gen AMD EPYC processors now 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, inflight or instore.

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

4TH GENERATION AMD EPYC PROCESSOR DESCRIPTION

Next Generation Server Architecture – AMD EPYC™ 9004 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 9004 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 EPYC 9004 CPUs, IT professionals can use fewer servers to get the job done compared with EPYC 7003 CPUs.

Latest Technology: The 9004 Series EPYC CPUs amplify the AMD history of x86 architecture innovations and record-breaking performance[1] with next-generation 5nm technology as well as introducing support for high-performant DDR5 DIMMs and fast PCIe® Gen 5 I/O. EPYC 9004 CPUs support 12 memory channels with 2 DIMMs/channel capability[2], delivering the resources needed for memory-hungry AI, ML, HPC, and large in-memory computations. These 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 4th Gen EPYC CPUs' high PCIe5 lane counts.

There are several advantages to using the 4th Gen AMD EPYC processors for different workloads with different models for various workloads. The various models can be categorized for:

- AMD EPYC 9004 CPUs are a family of CPUs that address a wide range of data center requirements. The AMD EPYC 9004 CPUs specifications are listed below and include the latest "Bergamo" and AMD EPYC™ 9004 Series Processors with AMD 3D V-Cache™ technology (codename "Genoa-X")processors.

AMD EPYC™ 9004 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
9754 "Bergamo"	128	256	2.25	3.10	360	256	12	4800	460.8	128	2P/1P
9734 "Bergamo"	112	224	2.20	3.00	340	256	12	4800	460.8	128	2P/1P
9684X "Genoa-X"	96	192	2.55	3.70	400	1150	12	4800	460.8	128	2P/1P
9654P	96	192	2.40	3.70	360	384	12	4800	460.8	128	1P
9634	84	168	2.25	3.70	290	384	12	4800	460.8	128	2P/1P
9554P	64	128	3.1	3.75	360	256	12	4800	460.8	128	1P
9534	64	128	2.45	3.70	280	256	12	4800	460.8	128	2P/1P
9474F	48	96	3.60	4.10	360	256	12	4800	460.8	128	2P/1P
9454P	48	96	2.75	3.80	290	256	12	4800	460.8	128	1P
9384X "Genoa-X"	32	64	3.10	3.90	320	768	12	4800	460.8	128	2P/1P
9374F	32	64	3.85	4.30	320	256	12	4800	460.8	128	2P/1P
9354P	32	64	3.25	3.80	280	256	12	4800	460.8	128	1P
9334	32	64	2.70	3.90	210	128	12	4800	460.8	128	2P/1P
9274F	24	48	4.05	4.30	320	256	12	4800	460.8	128	2P/1P
9254	24	48	2.90	4.15	200	128	12	4800	460.8	128	2P/1P
9224	24	48	2.50	3.70	200	64	12	4800	460.8	128	2P/1P
9184X "Genoa-X"	16	32	3.55	4.20	320	768	12	4800	460.8	128	2P/1P
9174F	16	32	4.10	4.40	320	256	12	4800	460.8	128	2P/1P
9124	16	32	3.00	3.70	200	64	12	4800	460.8	128	2P/1P

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

Supermicro's Wide Range of Product Lines

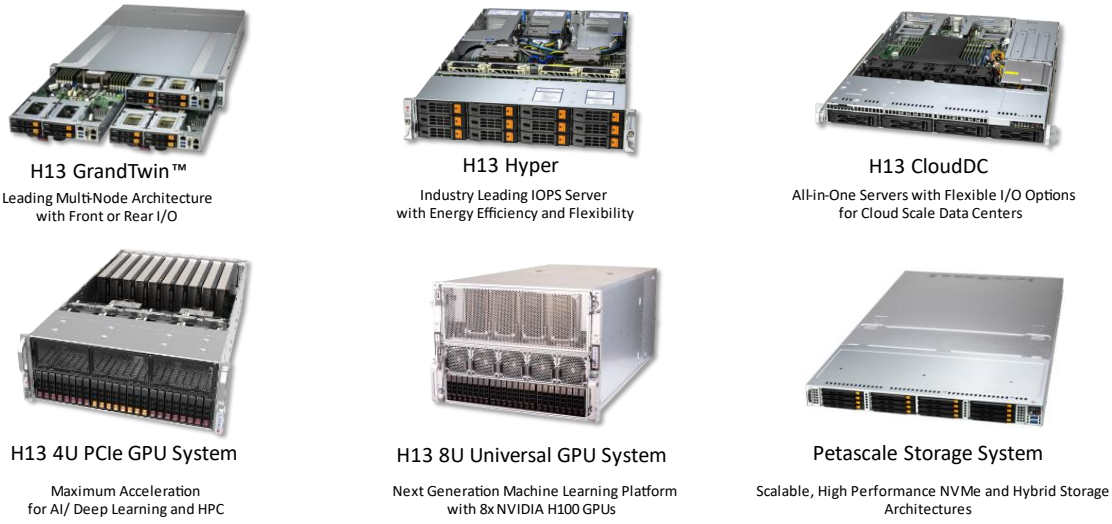

Figure 1 - Supermicro Product Lineup of H13 Systems

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 4th Gen AMD processors. The Supermicro product line can be segmented into the following areas. This white paper will look more closely at the following product lines:

GrandTwin™ – The Supermicro GrandTwin is an innovative system that puts multiple independent servers within the same enclosure. This lowers operating expenses by allowing the use of shared resources, such as the 2U enclosure, heavy-duty fans, backplane, and N+1 power supplies.

Common workloads include:

Diskless HPC • All-Flash HCI • Hybrid Cloud • All-Flash NVMe Storage • High-Performance File Systems • Software-Defined Storage

Figure 2 - Supermicro GrandTwin(TM)

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

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 at tremendous rates. 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 4th Gen AMD EPYC can accommodate, up to 6TB per socket.

a. **GPU with HGX** – With Supermicro's advanced architecture and thermal design, including liquid cooling and custom heatsinks, our 8U GPU system featuring NVIDIA's latest HGX H100 8-GPU baseboard, can deliver up to 6x AI training performance and 7x inference workload capacity and highest density in a flexible 4U system. The H13 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.

   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 4th 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.

Figure 3 - 5U GPU Server with OAM GPUs

Figure 4 - 8U 8 GPU Server

b. **GPU System with PCI-E GPUs** – The GPU systems that attach the GPU accelerators via the PCI-E 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 4th Gen AMD EPYC processors. Various platforms can accommodate from one to 10 GPUs.

   Common workloads include:

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

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 PCI-E will benefit significantly from the PCI-E 5.0 communications bus, the increased number of cores, and up to 6TB of memory that this system can accommodate.

Figure 5 - GPU System w/PCI-E GPUs

Hyper Family - The Hyper servers are designed for maximum rackmount flexibility with rear and front I/O for today's data center requirements as a single or 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 PCI-E slots (up to eight) for extreme flexibility, are toolless for fast and easy servicing, and come with various storage devices (NVMe/SAS/SATA). The Hyper systems can also support up to 2 AIOM/OCP 3.0 NICs.

Figure 6 – Hyper Server

Common workloads for the Hyper family include:

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

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

Hyper systems will benefit from the increased core count at similar pricing with the 4th Gen AMD EPYC processors. In addition, the faster PCI-E 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, and the systems can accommodate up to two double-width GPUs. The CloudDC family comes with dual AIOM OCP 3.0 support, which gives the product family tremendous expandability and flexibility. The CloudDC family also supports up to 6 PCI-E 5.0 slots. The PCI-E slots are equally split between the CPUs, which results in additional flexibility. 12 NVMe storage devices are supported for maximum I/O performance and capacity.

Common workloads for the CloudDC family include:

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

The 4th Gen AMD EPYC processor features that would benefit these applications include the increased core count and PCI-E 5.0.

Figure 7 - CloudDC Server

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

MEMORY CAPACITY

The 4th Generation AMD EPYC processors increase memory capacity that can be addressed directly per socket. This is due to the increased number of memory channel. A 2-socket system can address six Terabytes (TB) of memory 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	4th Gen AMD EPYC processors	% Increase
Memory DIMMs (max/socket)	16	24	50%
Max Memory (DRAM/socket)	4TB	6TB	50%

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	4th Gen AMD EPYC processors	% Increase
Memory Performance	3200 MHz	4800 MHz	50%
Number of Channels	8	12	50%
Total Memory Bandwidth	= 8 * 3200 MHz = 25,600 MT/s	= 12 * 4800 MHz = 57,600 MT/s	125%

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

FASTER CONNECTIONS TO PERIPHERALS

The 4th 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.

	PCI-E 4.0 (3rd Gen)	PCI-E 5.0 (4th 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 from a single unified console to IT administration by integrating data center tasks into a single intelligent management solution.

Applications Benefits Summary

With the new 4th Gen AMD EPYC processors, applications will benefit from several innovations.

- More cores – for applications that scale with the number of available cores, performance will increase.
- More extensive memory access – with more memory that can be 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 PCI-E 5.0, applications can communicate with PCI-E devices at twice the speed as before, resulting in overall application performance increases.
- Interconnect between sockets – for applications requiring socket-to-socket communication, the faster xGMI channels will reduce execution time.

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

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 PCI-E 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.

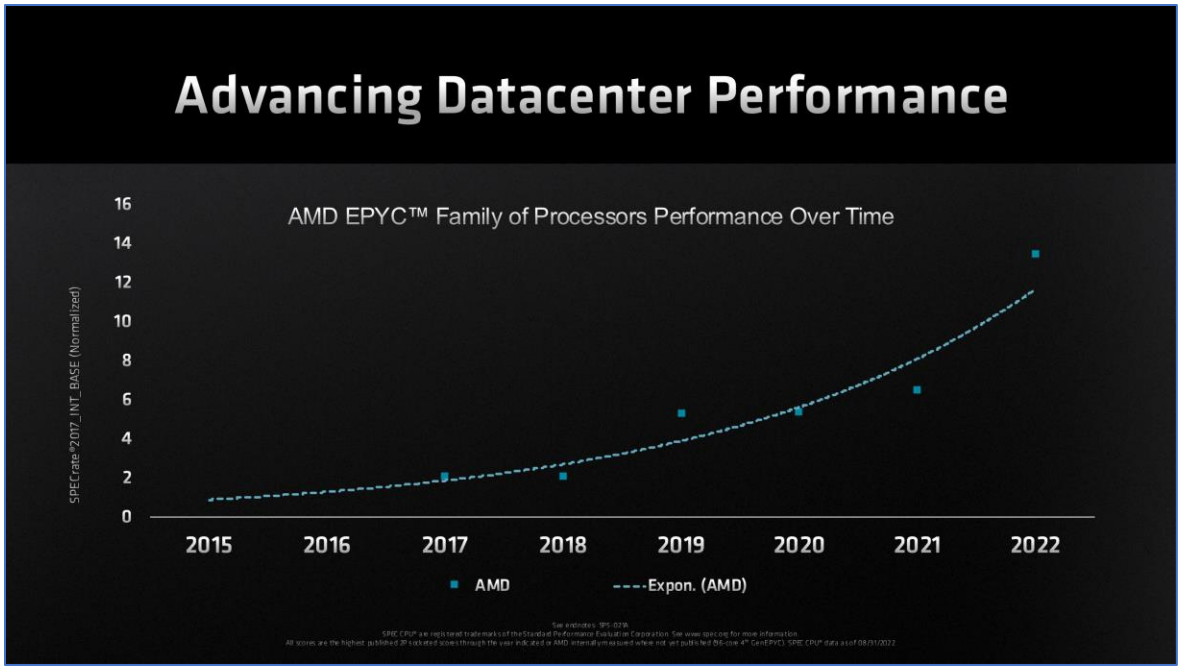

Image Courtesy of AMD

Summary

The new H13 product line from Supermicro enables all organizations to take full advantage of AMD's latest CPUs. Ranging from a single processor to the latest in blade technology and from 16 cores per socket to an amazing 128 cores per socket, Supermicro has a server designed for your workload. With the increase in the amount of memory that can be addressed and

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

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 H13 product lineup is designed for workloads that range from the Edge to the data center.

Footnotes

1. Based on AMD Estimates

Resources

www.supermicro.com/aplus

© 2023 Copyright Super Micro Computer, Inc. All rights reserved