WHITE PAPER

THE FUTURE OF COMPUTING WITH SUPERMICRO X12 SERVERS

New Supermicro Servers Designed to Meet Modern Workload Demands Utilizing the 3rd Gen Intel® Xeon® Scalable Processors

Table of Contents

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.

Introduction

Technology demands by enterprises and individuals worldwide increase as the digital transformation of all aspects of life continues. The generation of massive amounts of data continues to grow, which means new solutions are needed that allow employees and the public to quickly make better and faster decisions and participate in the global economy.

As a larger share of all data continues to be generated at the Edge, new technologies, both hardware and software, are needed to acquire, filter, and make autonomous decisions about that data before a select subset of it is sent to a central data center. Computing power appears everywhere, and a range of new systems is required for different workloads. While a single form factor cannot address all computing needs, systems can be designed using common building blocks that respond to low latencies and high bandwidths demands. In addition, a continuum of computing power from trusted vendors allows large-scale infrastructure systems to be built and deployed with a high degree of confidence and reliability.

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 and capacity, and environmental considerations are engineered and delivered by Supermicro. The key to offering so many different systems is advanced engineering and teaming with leading-edge CPU manufacturers, such as Intel.

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 that incorporate 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 numbers of cores, communication latency between cores, GHz, and which generation of CPU architectures can influence benchmark performance of real-world AI applications.

In this white paper, we take an in-depth look at Supermicro's latest X12 portfolio of servers and storage systems 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 do. Therefore, new CPUs allow them to reduce costs by assigning more work to lesser systems for these organizations.
- Enable new insights – By taking advantage of 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 X12 servers and storage systems. These include:

- **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.
- **5G/Telco** – The rapid development and installation of 5G networks drive demand for fast CPUs resilient to the environment. Systems need to be efficiently cooled while performing full analytics. The new X12 lineup provides significantly more computing power per watt with reduced cooling requirements.

- **Artificial Intelligence (AI)** – Systems with fast CPUs and associated GPU sub-systems are required for the growing AI use cases. Supermicro X12 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.
- **High-Performance Computing (HPC)** – HPC systems are used by more than just 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.
- **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.
- **Streaming and Content Delivery** – New services deliver video to end users, both within corporate environments and from data centers to the Edge, in real-time. The X12 systems, with the new fast CPUs and communication channels from storage devices, are suited very well to this task.
- **Virtualization** – With many enterprises utilizing virtualization technologies to get higher utilization from existing servers, the new X12 servers, with the 3rd Gen Intel Xeon Scalable processors, allow for higher-powered virtualization machines, as there are more cores available and faster CPUs.
- **Enterprise** – Typical enterprise workloads will benefit from the new X12 systems with increased performance and reduced costs. In addition, existing workloads will execute faster, using less power than on previous generations of Supermicro servers.

How 3rd Gen Intel Xeon Scalable Processor Enhances Workloads and Highlights

While increasing the performance of computing systems continues over time with Intel'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 3rd Gen Intel Xeon Scalable processors' new performance levels. HPC and AI require both the increased core numbers, increased GHz, and the parallelization and networking outside of the system itself.

- More cores: Maximum of 40 cores compared to 28 cores in the 2nd Gen Intel Xeon Scalable processors

- Faster communication: PCI-E 4.0 is 2X faster and uses more lanes than the maximum PCI-E available in previous Intel CPU generations
- More addressable memory: 3rd Gen Intel Xeon Scalable processors can address up to six TB of DRAM per socket, and even more when using Intel Optane Series 200 PMEM.
- Memory performance: The new Intel processors can utilize DDR4-3200Mhz memory, 9.1 % faster than previous generations.
- Faster communication between CPUs: More and faster Ultra Path Interconnects are available with the 3rd Gen Intel Xeon Scalable processor.
- AI Acceleration – New VPDBUSD instruction for accelerating 8-bit inferencing and Intel Deep Learning Boost.
- Additional security – Security extensions, Crypto acceleration, Total Memory Encryption
- Intel Speed Select technology allows for a limited number of cores to run at a higher GHz for higher performance.

There are several advantages in using the 3rd Gen Intel Xeon Scalable processors for different workloads with different models for various workloads. The various models can be categorized for:

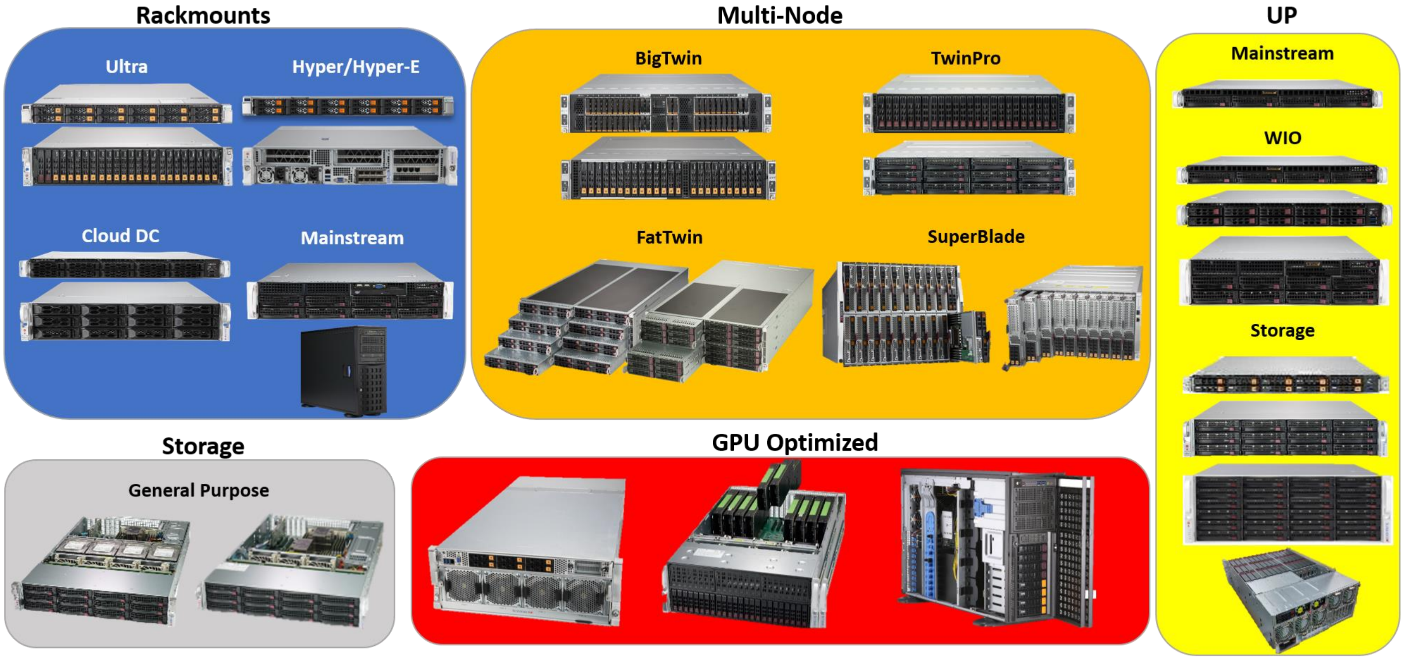
- Mainline SKUs – a range of the number of cores and speed, from eight cores to 32 cores, from 2.2 GHz to 2.8 GHz, and from 105 W to 205 W. These processors are meant for a wide range of applications and deliver higher performance at each price range compared to the previous generation of Intel Xeon Scalable processors.
- The highest Per-Core-Performance SKUs are designed for higher performance at a given number of cores, which means they run at higher frequencies. Processors that run with more cores at higher GHz require more power and produce more heat.
- Network Optimized SKUs – these 3rd Gen Intel Xeon Scalable processors are designed for systems whose primary role is low latency, high throughput, and deterministic performance in networking workloads. Therefore, these SKUs will have an "N" appended to the processor model number.
- NEBS SKUs – Systems that operate at the Edge or in combination with IoT devices must work in a wider range of environments than in a closed and environmentally controlled data center. The "T" processors are designed to withstand harsher environmental conditions yet still deliver the performance of up to 24 cores.
- SGX SKUs – The 3rd Gen Intel Xeon Scalable processors incorporate new features that ensure increased protection of data and applications. The SGX SKUs offer an increased amount of memory that is fully protected from potential attacks.
- Intel Speed Select allows a certain number of cores to run at a higher frequency (getting more work done) than the base frequency. These SKUs from Intel are the "Y", "S", and some "V" and "N" SKUs.

Supermicro Products Are Designed for Demanding Customers

The Supermicro product family contains more than 200 servers and storage systems designed with customer workloads. All take advantage of the new features and capabilities of the new 3rd Gen Intel Xeon Scalable processors. The Supermicro product line can be segmented into the following areas. This white paper will look more closely at the following product lines:

- Twin Family
    - BigTwin
    - FatTwin
    - TwinPro
- Ultra/Ultra-E
- GPU Family
    - GPU with HGX
    - GPU with PCI-E

- Hyper / Hyper-E Family
- Blade Family
- SuperStorage Family
- CloudDC Family



3RD GEN INTEL XEON SCALABLE PROCESSOR DESCRIPTION

The 3rd Gen Intel Xeon Scalable processors are the latest CPUs designed for a range of workloads. From the Edge to on-prem and cloud data centers, this new Intel CPU offers outstanding performance on several benchmarks and includes new features that enable a new level of security and AI acceleration. With more cores at higher frequencies than previous generations of CPUs, the 3rd Gen Intel Xeon Scalable processors enable new and innovative applications to be created and deployed from 5G/Telco applications to large-scale analytics. The increase in I/O performance due to PCI-E 4.0 allows faster communication to accelerator options. The amount of directly addressable memory is also increased, allowing more data to be kept in memory. Along with the Intel Optane 200 Series, it blurs the distinction between memory and storage.

Twin Family – The Supermicro Twin product line comprises innovative systems that put 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. Within the Twin product line are four product families:

a. **BigTwin®** – The BigTwin is a 2U design containing either two or four server nodes, accommodating up to 2 CPUs and up to 6TB of DRAM in each node. BigTwin systems are designed to accommodate a wide variety of storage options, memory topologies with Intel Optane PMem Series 200, and flexible networking options with AIOM, offering a superset of OCP 3.0 features and performance.

   Common workloads include:

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

The 3rd Gen Intel Xeon Scalable processors' best features include more cores, higher clock cycles, two 512-bit FMA units, new AI-optimized instruction sets, increased memory size, SGX support, and PCI-E 4.0 NVMe storage & I/O performance.

BigTwin w/2 Nodes

BigTwin w/4 Nodes

b. **FatTwin®** – The FatTwin product comes in either four or eight nodes, with up to 16 DIMM slots. Each node can support 2 CPUs and has versatile networking and storage options. FatTwin nodes are front accessible to accommodate cold-aisle serviceability environments and increased ease of use. Because of FatTwin's shared component design, energy savings are quickly realized, and each node can be configured with a range of hot-swappable storage devices.

Common workloads include:
Hyperscale / Hyperconverged • Cloud optimized servers • Data Center Enterprise Application • Scale-out of Storage Expansion • Telcom Data Center & ETSI Certified • Virtualization Server

FatTwin® systems will quickly benefit from the new 3rd Gen Intel Xeon Scalable processor features such as increased memory addressability, PCI-E 4.0 communication speeds and lanes, and an increased number of cores. In concrete terms, these mean up to five TB of Intel's Optane Persistent Memory and the ability to take advantage of FatTwin's modularity to support things like Smart NICs and other networking options to take advantage of the increased PCI-Lanes available in the 3rd Gen Intel Xeon Scalable processor architecture. In addition, FatTwin is a battle tested platform deployed worldwide in various use cases from the Edge to HPC. The new FatTwin X12 provides a seamless upgrade for these deployments with little operational risk in the upgrade process.

FatTwin w/4 Nodes

FatTwin w/8 Nodes

c. **TwinPro®** – The TwinPro systems are available in either 1U or 2U enclosure and contain two or four nodes. Each node has multiple storage and shared chassis, fans, backplane, and power supplies. Up to 4 TB of DRAM can be installed per node.

Common workloads include:

HPC • Hyperscale Data Center • Financial Analysis • Render Farms • CDN • Telco • Enterprise Mission-critical Applications • Data Center Cloud Computing • Virtualization • Big Data

TwinPro® systems will excel when utilizing the new 3rd Gen Intel Xeon Scalable processor features, specifically the increased core count and clock rates, faster communication through PCI-E 4.0, larger memory addressability, and support for single-socket or dual-socket configurations to deliver cost-optimized performance with Dual-10G ethernet onboard.

TwinPro w/4 Nodes

Ultra / Ultra-E - The Supermicro Ultra family consists of several systems designed for general enterprise computing and are either 1U or 2U in height. This server product family offers a way to consolidate many workloads into a single product line. The Ultra and Ultra-E servers can address up to 12 TB of memory and support the most powerful processors available. The Ultra family is very versatile. A full range of CPUs can be installed, up to 24 drives (NVMe/SATA/SAS) are supported, and there is ample expansion capacity (up to eight PCI-E slots). Ultra systems can accommodate up to four GPUs, enabling a range of AI and ML applications to be run on a single system. The Ultra-E systems are designed for the Telco industry with a shorter depth and optional redundant power supplies. They are available with NEBS Level 3 certification to operate in more harsh environments.

Common workloads include:

HCI • HPC • CDN • Hybrid Cloud • Container-as-a-Service • Cloud Computing • Big Data Analytics • Back-up and recovery

The 3rd Gen Intel Xeon Scalable processors' new features that will be most beneficial include fast Inter-Socket communication performance, faster PCI-E, and more cores per CPU. The Ultra will benefit most from the mainline and high-performance SKUs, while the Ultra-E will work best with the NEBS SKUs.

2U Ultra

2U Ultra-E

GPU Family – 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 that 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. The GPU family of servers not only can house multiple GPUs but are 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 3rd Gen Intel Xeon Scalable processors up to six TB per socket.

a. **GPU with HGX** – With Supermicro's advanced architecture and thermal design, including liquid cooling and custom heatsinks, our 4U GPU system featuring NVIDIA's latest HGX A100 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 X12 GPU systems feature the latest technology stacks such as 200G 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 3rd Gen Intel Xeon Scalable 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.

GPU with HGX

b. **GPU with PCI-E** – The GPU systems that attach the GPU accelerators via the PCI-E bus are ideal for environments that require multiple GPUs that perform their work with direct commands from the CPU. HPC and AI/ML environments will benefit significantly from the 3rd Gen Intel Xeon Scalable processors. Various platforms can accommodate from one to 10 GPUs.

Common 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 PCI-E will benefit significantly from the PCI-E 4.0 communications bus, the increased number of cores, and the six TB to 12 TB memory per CPU for large applications.

GPU System w/PCI-E

Hyper, Hyper-E Family - The Hyper and Hyper-E servers are designed for maximum rackmount flexibility with rear and front I/O for today's data center requirements. These systems can handle the maximum wattage of CPUs and the maximum number of DIMMs to accelerate a wide range of workloads. The Hyper-E family is designed for 5G and Telco environments, where NEBS Level 3 compliance is a must-have. With this requirement, the "N" and "T" SKUs (for 3rd Gen Intel Xeon Scalable processors) are ideal for the Hyper-E systems. The Hyper-E is also designed as a "short-depth" server, which means it can fit in tighter environments and features front IO accessibility. The Hyper-E is also available with AC and DC power options. The Hyper systems sport many PCIe slots (up to eight) for extreme flexibility, are tool-less 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.

Common workloads for the Hyper and Hyper-E family include:

5G Core and Edge • Telecom Micro Data Center • Enterprise Server • Cloud Computing • Big Data Analytics • Hyperconverged Storage • AI Inference and Machine Learning • Network Function Virtualization

Hyper and Hyper-E systems will benefit from the increased core count at similar pricing with the 3rd Gen Intel Xeon Scalable processors. The faster PCI-E 4.0 communications bus will give more rapid access to storage devices

2U Hyper / Hyper-E

CPU CORE COUNT AND PERFORMANCE

The Intel 3rd Gen Xeon Scalable processor is built with Intel's 10nm process. By reducing the transistor size, more transistors can be placed on a chip, leading to higher overall performance by increasing the number of transistors on a die, and increasing the number of cores per socket. While the previous generation of Intel Xeon Scalable processors had a maximum of 28 cores per socket, the Intel 3rd Gen Scalable processors contain a maximum of 40 cores. For applications that scale well and can take advantage of the increased core count, a performance increase will be automatic, as long as the application does not have an I/O bottleneck. At a minimum, the performance should increase linearly to the 28 core count of the previous generations of Intel CPUs (2nd Gen Intel Xeon Scalable processors). The 3rd Generation Intel Xeon Scalable processors use the new Sunny Cove core, a new microarchitecture that contains new instructions, larger caches, and other enhancements that are available to all applications. Although some applications may need to be recompiled to take advantage of the new instructions, many CPU-dependent applications will see an 18% increase in performance.

	2nd Gen Intel Xeon Scalable processors	3rd Gen Intel Xeon Scalable processors (1S-2S)	% Increase
Max Cores	**28**	**40**	**42%**
GHz at Max Cores	**2.7**	**2.3**	**NA**
Total GHz at Max Cores	**75.6 GHz**	**92.0 GHz**	**22%**

MEMORY CAPACITY

The 3rd Gen Intel Xeon Scalable processors increase memory capacity that can be addressed directly per socket. This is due to the increased number of memory channels and new memory technology (See PMEM below). The Gold and Platinum series of CPUs can address six Terabytes (TB) of memory per socket. Each socket can address 16 DIMMs, configured to contain either DRAM or the Optane 200 Persistent Memory Module (PMEM). However, a system cannot have more DIMMs containing PMEM than DRAM. Thus, for maximum memory addressability, a system with 16 DIMM slots could be set up with 8 DIMM slots with 256GB memory sticks and 8 DIMM slots with 512 Optane sticks. The total capacity per socket would be: (8 * 256 GB) + (8 * 512 GB) = 2 TB + 4 TB = 6 TB. This is 33% more than the previous generation of Intel CPUs, and Intel Optane PMEM was able to address.

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.

	2nd Gen Intel Xeon Scalable processors	3rd Gen Intel Xeon Scalable processors (1S-2S)	% Increase
Memory DIMMs (max/socket)	12	16	33%
Max Memory (DRAM)/socket	3 TB	4 TB	33%
Max Optane PMEM Slots	6	8	33%
Max Total Memory at ½ DRAM, ½ PMEM)/socket	=6 * 256 GB + 6 * 512 GB = 4.5 TB	=8 * 256 GB + 8 * 512 GB = 6 TB	33%

Blade Family – The Supermicro blade family of systems is designed for the most demanding workloads that require a high density of CPUs and the fastest networking available today. The X12 blades are available in one or two-socket varieties and can contain one or two GPU accelerators on each blade. In addition, the X12 blades are available in the following configurations, at a high level:

- 8U SuperBlade, which can accommodate up to 20 hot-pluggable dual-socket nodes, delivers high performance with advanced networking such as 200G HDR InfiniBand.

- 6U SuperBlade, which can accommodate up to 10 hot-pluggable single-socket or dual-socket nodes, provides access to max memory in a high performance, dense infrastructure. For workload acceleration, the single socket blades accommodate one double-width or two single-width GPUs.
- 4U SuperBlade, which can accommodate up to 14 hot-pluggable dual-socket nodes, is optimized for high density and value.

SuperBlade Chassis 8U

SuperBlade Chassis 6U

SuperBlade Chassis 4U

A significant advantage of the Supermicro SuperBlades is that cooling and power are shared between the server blades, which reduces power consumption compared to individual rackmount servers that contain their own and non-sharable fans and power supplies. Supermicro offers liquid cooling for the 8U enclosure, which is required if the blade accommodates two high TDP (>220 W) 3rd Gen Intel Xeon Scalable processors.

Common workloads for SuperBlade systems include:

HPC • Hybrid Cloud • EDA • Virtualization • Health • Financial Services

Many of the 3rd Gen Intel Xeon Scalable processor's new features will benefit all users of the SuperBlades. For example, the increased core count, performance, and amount of directly addressable memory are extremely valuable for workloads running on the SuperBlades. Also, support for PCI-E 4.0 allows for faster communication with GPUs that are installed on the blades.

1 Socket Blade

2 Socket Blade

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 the Intel processors limit was 2933 MT/s, and six channels could deliver 6 x 2933 = 17,600 MT/s. The Intel 3rd Gen uses 8 channels for memory access, thus the maximum performance per socket = 8 * 3200 MT/s = 25,600 MT/s, a 45% improvement.

	2nd Gen Intel Xeon Scalable processors	3rd Gen Intel Xeon Scalable processors	% Increase
Memory Performance	2966 MHz	3200 MHz	9%
Number of Channels	6	8	33%
Total Memory Bandwidth	= 6 * 2933 MHz = 17,600 MT/s	= 8 * 3200 MHz = 25,600 MT/s	45%

FASTER INTERCONNECT BETWEEN SOCKETS (ULTRA PATH INTERCONNECT)

The Intel 3rd Gen has faster communication between sockets with up to three Intel Ultra Path Interconnects (Intel UPI) running at 11.2 GT/s. This feature is essential when applications are using more than one socket, and the sockets must communicate. With faster communications for applications that run across sockets, performance will benefit and show a decrease in time to solution or delivering results.

	2nd Gen Intel Xeon Scalable processors	3rd Gen Intel Xeon Scalable processors (1S - 2S)	% Increase
Number of UPI links (max)	3	3	
Performance	10.4 GT/s	11.2 GT/s	
Total UPI Throughput	= 3 * 10.4 = 31.2	= 3 * 11.2 = 33.6	8%

FASTER CONNECTIONS

The 3rd Gen Intel Xeon Scalable processor supports the PCIe Gen 4 standard, which has a peak performance of twice that of the previous PCIe Gen 3 standard. PCIe Gen 4 delivers 16 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 GT = .125 GB)

Thus, for PCIe Gen 4 a system with 16 lanes, the communication can achieve 16 GT/s x 16 lanes / 8 = 23 GB/second. The aggregate performance is 2X what PCIe Gen 3 delivers. A faster PCIe bus is critical when using GPUs or FPGAs.

	PCI-E Gen 3	PCI-E Gen 4	% Increase
Per Lane Performance	8 GigaTransfers/Second	16 GigaTransfers/Second	100 %
Number of Lanes from CPU	48	64	33 %
Total Performance	8 * 48 = approx. 384 Gb/s	16 * 64 = approx. 1024 Gb/s	167 %

SuperStorage Family – The SuperStorage family from Supermicro is a new generation of top-loading storage servers that allow easy field serviceability and up to 90 drives. Both HDDs and SSDs are supported in both the 2.5" and 3.5" form factors. The SuperStorage system can house both HDDs and SSDs, which may hold different data types in the same system. These systems contain hot-swappable expanders, drives, power supplies, and fans.

Common workloads for the SuperStorage systems include:

Object Storage • Data Intensive HPC/AI • Private & Hybrid Cloud • Backup & Active Archive

The SuperStorage systems will benefit from the increased clock rates and PCI-E 4.0 I/O speeds with the 3rd Gen Intel Xeon Scalable processors.

90 Bay SuperStorage System

CloudDC Family – The CloudDC family is explicitly designed for cloud data centers, where space is premium. The CloudDC product line is toolless, meaning that servicing these servers 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 4.0 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. The tool-less mechanical design makes serviceability easy along with providing hot-swap storage capability.

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 3rd Gen Intel Xeon Scalable processor features that would benefit these applications include the increased core count and PCI-E 4.0.

2U CloudDC

Supermicro Networking Solutions

Supermicro networking solutions are designed to perform at their optimal level when configured with Supermicro systems. Supermicro introduces the Intel E810 Series (code name Columbiaville) and its X12 launch. Intel E810 offers bandwidth from 25 Gbps up to 100 Gbps with RoCE v2, iWARP, Advance Device Queues (ADQ), and Dynamic Device Personalization (DDP). Intel ADQ technology increases data transaction speed with low latency and high throughput for various applications from Artificial Intelligence, Big Data Analysis, and NVMe-oF to content delivery and High-Performance Computing. DDP is designed to enhance 5G telecommunication tunneling protocols. Intel E810 has an intelligent parser integrated into the ASIC that can parse the tunneling header, offload locally, and send the raw data to the CPU for processing. This allows the CPU to focus on the application without processing additional headers and lookups.

With ADQ and DDP, Intel E810 is one of the leading LAN solutions on the market. The optimized performance reduces CPU resource usage and investment cost when measuring its performance against other LAN solutions.

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.

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 the Intel processors limit was 2933MT/s, and six channels could deliver 6 x 2933 = 17,600 MT/s. The Intel 3rd Gen uses 8 channels for memory access, thus the maximum performance per socket = 8 * 3200 MT/s = 25,600 MT/s, a 45% improvement.

	2nd Gen Intel Xeon Scalable processors	3rd Gen Intel Xeon Scalable processors	% Increase
Memory Performance	2966 MHz	3200 MHz	9%
Number of Channels	6	8	33%
Total Memory Bandwidth	= 6 * 2933 MHz = 17,600 MT/s	= 8 * 3200 MHz = 25,600 MT/s	45%

Applications Benefits Summary - With the new 3rd Gen Intel Xeon Scalable 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 4.0, applications can communicate with PCIe devices at twice the speed as before, resulting in overall application performance increases.
- Interconnect between sockets – for applications that require socket-to-socket communication, the faster UPI channels will reduce execution time.
- AI instructions – accelerate AI inferencing applications with Intel Deep Learning Boost.

How Did We Do It?

Supermicro incorporates a Building Block® approach which allows us to design individual components with the latest technology and then engineer these different components together 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. Depending on the form factor, cooling requirements, and memory requirements, application-optimized systems can quickly develop. 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.

Additional Data Centers Requirements for Improved Performance

CPUs alone are not enough to maintain or enhance customer offerings or gain new insights. Each combination of CPUs, memory, storage, and networking needs to be constantly improved so that bottlenecks of one sub-system do not restrict the overall performance. While this paper will focus on the CPU enhancements of the 3rd Gen Intel Xeon Scalable processors, other technologies contribute to new performance levels of servers and clusters. These include:

- **Intel Optane Persistent Memory (PMEM) 200 series** – the next generation of PMEM that resides on the memory bus. Unlike DRAM, PMEM retains the data even when a system is powered off, protecting systems from data loss. Memory capacity can be increased at a lower cost than DRAM, enabling more extensive data sets to be kept in memory and quickly accessible by the CPUs. PMEM also acts in two modes, AppDirect Mode and Memory Mode, allowing applications to use PMEM either as additional memory or persistent memory. Below are some examples of the total memory capacity using DDR4 and PMEM with 3rd Gen Intel Xeon Processors.

Dual CPU	App Direct Mode (AD)						Memory Mode (MM)		
	4+4	6+1	8+1	8+4	8+8	12+2	4+4	8+4	8+8
Max Capacity (DRAM + Pmem)	6TB	4TB	5TB	8TB	12TB	8TB	4TB	4TB	8TB
Hyper (32 DIMM)	x	x	x	x	x	x	x	x	x
Ultra (32 DIMM)	x	x	x	x	x	x	x	x	x
Cloud DC (16 DIMM)	x	x					x		
Mainstream (16 DIMM + 2 PMem)	x	x	x				x		
BigTwin (16 DIMM + 4 PMem)	x	x	x				x		
Fat Twin (16 DIMM)	x	x					x		
SuperBlade B12DPE (32 DIMM)	x	x	x	x	x	x	x	x	x
SuperBlade B12DPT (16 DIMM)	x	x					x		
GPU Optimized (32 DIMM)	x	x	x	x	x	x	x	x	x
GPU Optimized (16 DIMM)	x	x					x		
SuperStorage (16 DIMM)	x	x					x		

Max capacity calculated with DDR4 256GB modules and PMem 512GB modules

Single CPU	App Direct Mode (AD)						Memory Mode (MM)		
	4+4	6+1	8+1	8+4	8+8	12+2	4+4	8+4	8+8
	3TB	2TB	2.5TB	4TB	6TB	4TB	2TB	2TB	4TB
SuperBlade B12SPE (8 DIMM)	x	x					x		
UP (8 DIMM)	x	x					x		

Max capacity calculated with DDR4 256GB modules and PMem 512GB modules

- **Intel Ethernet 800 series network adapters** – Determining the speed at which data is delivered from a storage system (external) to the systems running the applications, a fast and low-cost networking solution is critical for application performance. With up to 200 GbE per PCIe 4.0 slot for bandwidth-intensive workloads, applications that need more than one system can work together faster, producing results quicker.

Performance / Power over time (Intel chart). Why this is important to data centers.

Over time, with Intel's advancement of CPU technology, more computing power is available at a given price and a given amount of energy. Intel has increased the amount of work performed per unit of electricity by a factor of 5 over the past 12 years. This means that more work can be performed at a constant power draw, enabling organizations to offer more services and applications to their employees or the public.

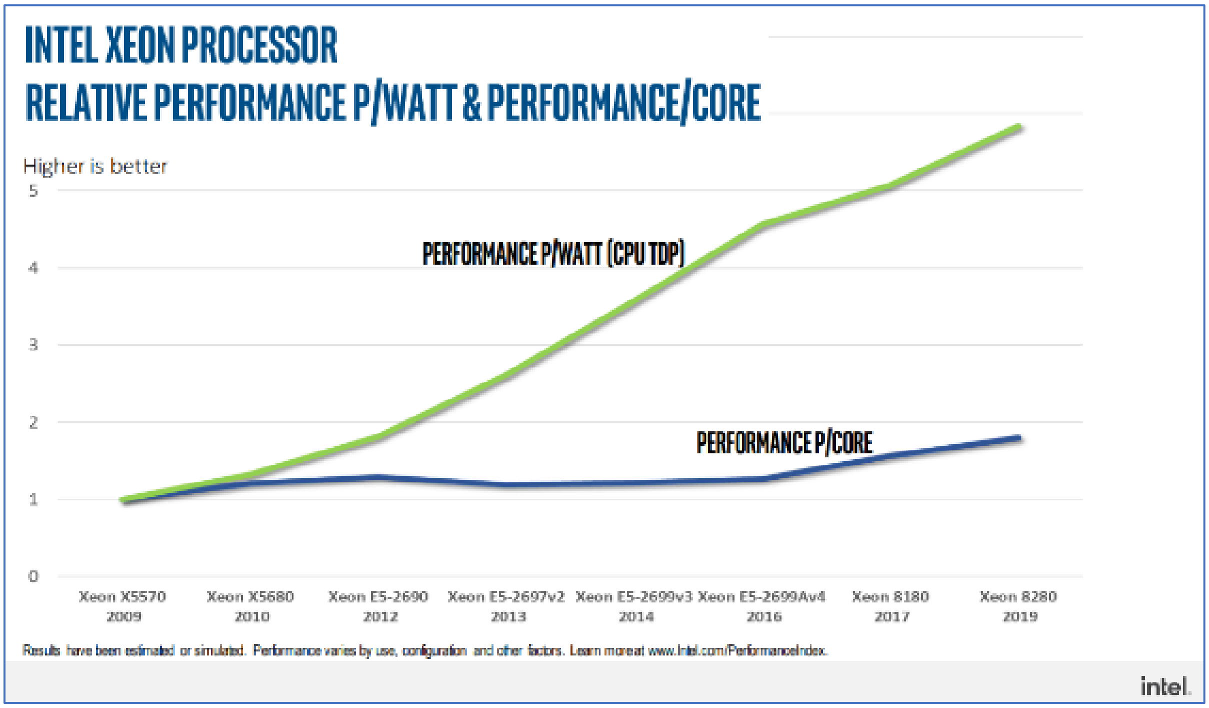


Image Courtesy of Intel

Summary

The new X12 product line from Supermicro enables all organizations to take full advantage of Intel's latest CPUs. Ranging from a single processor to the latest in blade technology and from 8 cores per socket to 40 cores per socket, Supermicro has a server designed for your workload. 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 X12 product lineup is designed for workloads that range from the Edge to the data center.

Resources

www.supermicro.com/x12