PRODUCT BRIEF

ACCELERATE EVERYTHING – FROM 5G TO DATA CENTER WORKLOADS

Dramatically Increase Performance Using Supermicro Servers with 4th Gen Intel Xeon Scalable Processors

TABLE OF CONTENTS

Executive Summary

The 4th Gen Intel Xeon Scalable processors contain several advanced technologies that speed up various workloads. However, specific workloads can benefit from dedicated acceleration engines on the CPU itself, dramatically improving performance, reducing power consumption, and reducing the number of cores needed for a particular workload. Compared to previous generations of Intel Xeon Scalable processors (2nd Gen and 3rd Gen), there are more cores per CPU, faster cores (both base clock and turbo boost), and the ability to communicate faster to devices external to the CPU. These new technologies include DDR5-4800MHz memory and PCIe Gen 5, which communicate to a wide range of peripherals. In addition, one of the most technological advances in the 4th Gen Intel Xeon Scalable processors is the built-in Intel Accelerator Engines that can increase the workload's performance well beyond just the increase in cores and clock rates. This Supermicro Product Brief looks more closely at the available accelerators and how specific workloads will benefit.

Modern Workloads That Can Be Accelerated

While each industry has specific applications, some common technologies are essential across industries that can be significantly accelerated with the new Intel Accelerator Engines. The table below summarizes acceleration opportunities across workloads such as AI, HPC, 5G, Storage/Networking, and Enterprise/Analytics.

	AI	HPC	5G	Storage/Networking	Enterprise/Analytics
Intel® Xeon CPU Max (HBM)	X	X			
Intel® AMX	X	X			
Intel® QAT				X	
Intel® DSA				X	
Intel® DLB			X	X	
Intel® IAA					X
Intel® vRAN Boost			X		
Ideal Supermicro Systems	SuperBlade® Hyper Twin Family	SuperBlade® Hyper	GrandTwin™ SuperEdge Hyper-E	Petascale Twin Family CloudDC	MP Servers Petascale SuperBlade

Application: AI

Artificial intelligence is becoming pervasive, with new applications being introduced frequently. With new software optimized for the latest hardware innovation, organizations can make better decisions, realize more optimal outcomes, and eliminate repeatable tasks.

An efficient AI solution deployment will incorporate several technologies, including processors, which can be accelerated using the accelerators on the 4th Gen Intel Xeon Scalable processors. The following options give applications a performance boost for AI environments without additional hardware simply by activating these accelerators:

- Intel Advanced Matrix Extensions (Intel® AMX)
- Intel QuickAssist Technology (Intel ® QAT)

The ideal Supermicro X13 servers for AI include:

- SuperBlade - Supermicro's 8U SuperBlade is optimized for performance, density, and advanced networking. With up to 20 nodes in an 8U chassis and available air and liquid cooling options, each node occupies just 0.4U of rack space, delivering unprecedented compute density. The modular resource-saving architecture utilizes shared power and cooling to reduce power consumption and allow modular refresh of subsystems, reducing e-waste and TCO.

- Hyper - The Hyper series are enterprise-focused servers built with versatility and performance, optimized for supporting the highest TDPs and offering a flexible range of computing, networking, storage, and I/O expansion capabilities.
- Twin Family Servers - The industry's unparalleled multi-node design strikes a balance between density and performance, enabling cloud services to optimize their use of space and resources. The Twin family is specially crafted to deliver high-density computing power, complemented by lightning-fast high performance storage, making the most out of available space and resources.

8U SuperBlade®

Hyper 2U 2Socket

BigTwin® 2U 4Node

Application: HPC

High-Performance Computing (HPC) is a technology that involves many servers working together to solve complex mathematical models that simulate a process. Sometimes referred to as a supercomputer, an HPC environment today is comprised of hundreds to thousands of systems working together to solve numerical equations that unleash new knowledge or result in a better understanding of physical processes.

HPC applications will run faster with each new CPU generation that Intel releases. The increased core count per CPU, the increased clock rate per core (aggregate – Core-GHz), and microarchitecture improvements have led to tremendous increases in performance. Using the new Accelerators incorporated in the 4th Gen Intel Xeon Scalable processors, applications can get an additional performance increase, above and beyond what the base CPU can offer. Since the overall performance of an HPC system needs to work closely with the storage and networking sub-systems, there are several Intel Accelerator Engines that will help to increase the performance of an HPC system, whether running a low number of large (thousands of cores) jobs, or many, smaller (tens of cores) jobs.

- Intel Xeon CPU Max Series (High Bandwidth Memory)
- Intel Advanced Matrix Extensions (Intel® AMX) for compute acceleration
- Intel QuickAssist Technology (Intel® QAT) for storage and networking tasks

Supermicro X13 servers that are ideal for HPC workloads include:

- Hyper - The Hyper series are enterprise-focused servers built with versatility and performance, optimized for supporting the highest TDPs and offering a flexible range of computing, networking, storage, and I/O expansion capabilities.

- SuperBlade - Supermicro's 8U SuperBlade is optimized for performance, density, and advanced networking. With up to 20 nodes in an 8U chassis and available air and liquid cooling options, each node occupies just 0.4U of rack space, delivering unprecedented compute density. The modular resource-saving architecture utilizes shared power and cooling to reduce power consumption and allow modular refresh of subsystems, reducing e-waste and TCO.

8U SuperBlade®

Hyper 2U 2Socket

Application: 5G

With more data generated at the network's edge, 5G has become an essential technology for more than just cell phones. All types of devices will use 5G for communication with other edge devices and sending data over the network to distributed data centers. The rise of 5G and the server and CPU innovation that brings significant computing power to the edge enables a world where everything is connected.

5G environments may have specific environmental constraints for the servers, which may dictate the power or size envelope of the server. Using just the CPU-based functions can reduce power consumption and allow small physical server form factors. The 4th Gen Intel Xeon Scalable processors incorporate accelerators that speed up specific 5G workloads and reduce the CPU work by offloading certain functions. Performance improvements specifically for 5G include:

- vRAN Boost
- Intel QuickAssist Technology (Intel® QAT)
- Intel Data Streaming Accelerator (Intel® DSA)

Supermicro servers for 5G environments:

- SuperEdge – Supermicro's SuperEdge is designed to handle increasing compute and I/O density requirements of modern edge applications. With three customizable single-processor nodes, SuperEdge delivers high-class performance in a 2U, short-depth form factor. Each node is hot-swappable and offers front access I/O, making the system ideal for remote IoT, edge, or telco deployments.

- Hyper-E – Supermicro's Hyper-E brings the performance and flexibility of Supermicro's flagship Hyper series to the edge with short-depth form factors designed for edge data centers and telco deployments. Telco-optimized configurations are NEBS Level 3 certified and feature optional DC power supplies on selected models.
- GrandTwin™ – Supermicro's GrandTwin™ family of servers is a new multi-node architecture purpose-built for single-processor performance. Front-serviceable hot-swap nodes allow easier installation and servicing in space-constrained environments. The GrandTwin architecture delivers high performance in a modular design optimized for a wide range of applications, with Supermicro's Resource Saving Architecture delivering improved power efficiency and lower materials costs.

GrandTwin™

SuperEdge

Hyper-E

Application: Storage/Networking

Cloud-Native applications are being built around several individual services that must communicate rapidly to pass information and data back and forth. The heavy networking traffic of a modern data center requires low latencies and high bandwidth, allowing CPUs to continue running the application. In addition, with security so critical, data needs to be encrypted, decrypted, compressed, and decompressed with minimal impact on the application running on the server. Some of the same acceleration that boosts 5G networking performance gains also apply to storage and networking. Both storage work and networking workloads benefit from the same Intel Accelerator Engines:

- Intel QuickAssist Technology (Intel® QAT) for networking tasks that offload the CPU.
- Intel Dynamic Load Balancer (Intel® DLB) optimizes and distributes networking tasks across cores based on traffic and loads.

Supermicro servers for Storage and Networking environments:

- Petascale – Supermicro X13 All-Flash systems offer industry-leading storage density and performance with EDSFF drives allowing unprecedented storage capacity in a single 1U chassis. The advanced high-density server design paired with the unmatched efficiency of EDSFF flash media provides exceptional IOP-per Watt performance. This combination of performance and TCO value will accelerate the transition from legacy HDD for many large scale, capacity hungry applications used across a range of data-intensive industries.
- BigTwin® – The industry's unparalleled multi-node design strikes a balance between density and performance, enabling cloud services to optimize their use of space and resources. The Supermicro BigTwin is designed to deliver

high-density computing power, complemented by lightning-fast high performance storage and networking, making the most out of available space and resources.

- CloudDC (for networking) – Ultimate flexibility on I/O and storage with up to 6 PCIe 5.0 slots and dual AIOM slots (PCIe 5.0; OCP 3.0 compliant) for maximum data throughput. Supermicro X13 CloudDC systems are designed for convenient serviceability with toolless brackets, hot-swap drive trays, and redundant power supplies that ensure rapid deployment and more efficient maintenance in data centers.

Petascale

BigTwin®

CloudDC

Application: Enterprise/Analytics

The modern enterprise must work with a wide range of users (customers and employees) and relies on databases and analytics to run the company and make forward-looking decisions. Databases were typically a scale-up model, but in recent years have become distributed and can use as many nodes as needed to meet service level agreements (SLAs). Thus, running an efficient database or performing data analytics requires several CPU enhancements. With the deployment of a distributed system, additional accelerators can be utilized to meet or exceed SLAs and give users more insight into corporate data. The Intel Accelerator Engines, which reduce latency and respond quickly to users, include:

- Intel® In-Memory Analytics Accelerator (Intel® IAA)
- Intel QuickAssist Technology (Intel® QAT) for networking tasks that offload the CPU.

Supermicro servers that are designed for Enterprise and Analytics workloads:

- 4Way and 8Way MP Servers – X13 multi-processor systems bring new levels of compute performance and flexibility with support for 4th Gen Intel® Xeon® Scalable processors to support mission-critical enterprise workloads. A large memory footprint is ideal for large databases and in-memory compute applications to enable even the most memory-intensive applications. In addition, dynamic storage options support directly attached full-hybrid all NVMe for lower latency with higher throughput and IOPS.
- Petascale Storage – Supermicro X13 All-Flash systems offer industry-leading storage density and performance with EDSFF drives allowing unprecedented storage capacity in a single 1U chassis. The advanced high-density server design paired with the unmatched efficiency of EDSFF flash media provides exceptional IOP-per Watt performance. This combination of performance and TCO value will accelerate the transition from legacy HDD for many large scale, capacity hungry applications used across a range of data-intensive industries.

- SuperBlade - Supermicro's 8U SuperBlade is optimized for performance, density, and advanced networking. With up to 20 nodes in an 8U chassis and both air and liquid cooling options available, each node occupies just 0.4U of rack space, delivering unprecedented compute density. The modular resource-saving architecture utilizes shared power and cooling to reduce power consumption and allow modular refresh of subsystems, reducing e-waste and TCO.

8 Socket MP Server

4 Socket MP Server

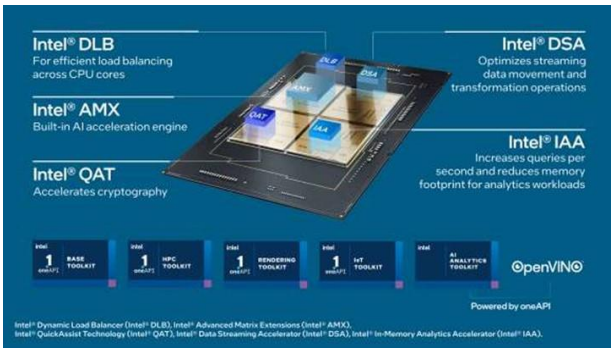
Petascale Server

8U SuperBlade®

Basic Architecture of 4th Gen Intel Xeon Scalable Processors

The 4th Gen Intel Xeon Scalable processors are built with the Intel 7 process technology. With a maximum core count of 60 (50% higher than the previous generation), this latest generation of CPUs from Intel runs a wide variety of workloads faster than ever before. While the raw performance of a server will be higher due to the increased core counts, PCIe Gen 5, and DDR5 memory speeds, there is an additional opportunity for performance gains for specific workloads using the Intel Accelerator Engines that are built into the 4th Gen Intel Xeon Scalable processors.

All 4th Gen Intel Xeon Scalable processors contain the electronic circuitry for different accelerators. By default, based on the specific SKU, a certain number of these accelerators may be available or can be turned on with a key. Learn more about which 4th Gen Intel Xeon Scalable processors contain Intel Accelerators, including Intel Accelerator Engines. Below is a simplified diagram of the 4th Gen Intel Xeon Scalable process, illustrating where the Intel Accelerator Engines reside.



The accelerators that are part of the latest 4th Gen Intel Xeon Scalable processors include:

- Intel® QuickAssist Technology (Intel® QAT)
- Intel® Dynamic Load Balancer (Intel® DLB)
- Intel® Data Streaming Accelerator (Intel® DSA)
- Intel® In-Memory Analytics Accelerator (Intel® IAA)
- Intel® Advanced Matrix Extensions (Intel® AMX)

Intel Accelerator Capacities Per CPU

The 4th Gen Intel Xeon Scalable processors are designed to activate additional performance for specific workloads based on the workload requirements. Below is a simple table that shows the number of Intel Accelerator Engines available on the different CPUs.

Devices	XCC	MCC
DSA	1 or 4	1
IAA	0, 1, or 4	0 or 1
QAT	0, 1, or 4	0, 1, or 2
DLB	0, 1, or 4	0, 1, or 2

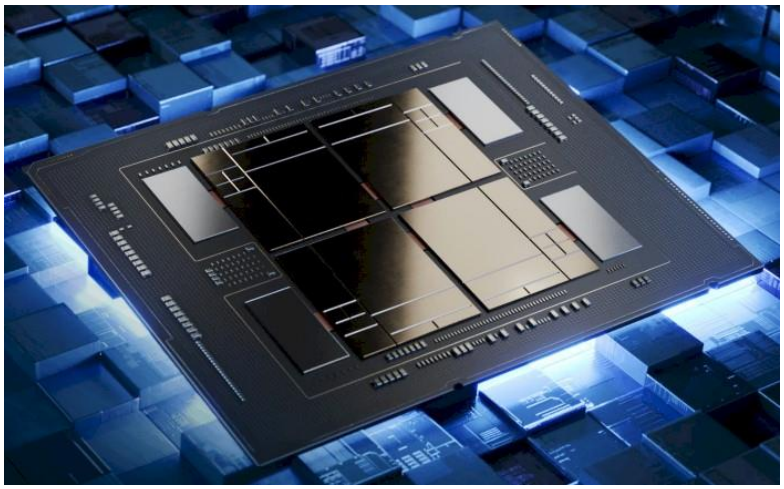
The following are supplemental technologies from Intel but are not considered Intel® Accelerator Engines.

- Intel Xeon CPU Max Series (High Bandwidth Memory)
- Intel vRAN Boost

Intel Xeon CPU Max Series

The Intel Xeon CPU Max Series connects high bandwidth memory (HBM) directly to the CPU, rather than the application having to retrieve data over the common memory bus. This design can significantly increase the performance of memory-intensive applications, such as HPC and Data Analytics. Up to 64GB of HBM can be accommodated on select 4th Gen Intel Xeon Scalable processors (SKUs: 9480, 9470, 9468, 9460, 9462). Memory bandwidth from the HBM memory to the CPU exceeds 1TB/second.

Performance Increase: With the increase in bandwidth using DDR5 memory applications, some applications will see an increase in performance. However, even faster bandwidth to the CPU is possible for certain applications. Compared to the previous generation of Intel Xeon processors (3rd Gen, 8380), a 4th Gen Intel Xeon CPU Max series system showed a performance increase of 1.4X on HPL, 5.3X gain on HPCG, and a 3.0X gain on CosmoFlow applications[1].



Intel Xeon CPU Max Series

The Intel Xeon CPU Max Series includes three usage modes:

• HBM Only – in this mode, the CPU only addresses the HBM memory. There is no DDR present. Applications would then be limited by the amount of HBM memory, which currently cannot exceed 64GB.

• HBM Flat – Applications can use both HBM memory and DDDR5 memory at the same time. The HBM memory will be used first. Some code optimization may have to be done.

• HBM Caching – Since the HBM memory has significantly faster bandwidth to the CPUs than standard memory channels, the HBM memory will be used as a cache from the main memory, which will hold the data used frequently.

Over a range of HPC application tests, the applications ran from 1.2 to 2.44 times faster when using a server containing the Intel Xeon CPU Max series compared to a high-end 4th Gen Intel Xeon Scalable processor (no HBM).

Intel vRAN Boost

Intel® vRAN Boost – Up to twice the performance for vRAN workloads vs. 3rd Gen Intel Xeon Scalable processors. Energy efficiency is twice that of the 3rd Gen Intel Xeon Scalable processors, reducing power consumption and helping to lower the total cost of ownership.

Intel Accelerator Engines

Intel® QuickAssist Technology (Intel® QAT)

Intel® QuickAssist Technology (Intel® QAT) – Helps to reduce system resource consumption—and TCO—by accelerating cryptography and data compression with Intel® QuickAssist Technology (Intel® QAT). By offloading encryption, decryption, and compression, this built-in accelerator helps free up processor cores so that systems can serve more clients or use less power. The most value is when you merge compression and encryption together.

(Image Courtesy of Intel®)

a. Business Value
    i. Accelerated compression/decompression offloading leads to greater CPU efficiency.
    ii. More encrypted connections and secure web connections between devices with less overhead

b. Maximum Number Per CPU: 4

c. Performance Measurement for Networking and Storage operations:
    i. Up to 47% fewer cores to achieve the same connections/s on NGINX with built-in QAT vs. out-of-the-box software [2].
    ii. Up to 95% fewer cores and 2x higher level 1 compression throughput using integrated Intel® QuickAssist Technology (Intel® QAT) vs. the prior generation[3]

d. Scalability
   You can increase your product line performance and scale with the incremental acceleration you need (scaling from 1 to 4 Intel QAT devices on-chip)

e. Efficiency
   Significant Core Utilization Savings translates to Significant Performance/Watt improvements.

Intel® Dynamic Load Balancer (Intel® DLB)

Intel Dynamic Load Balancer (Intel® DLB) – The Intel® DLB enables efficient load balancing across CPU cores by offloading software queue management. As a result, tasks such as packet processing allow for higher performance and better work balancing between cores.

Intel® DLB improves the system performance for handling network data on multi-core Intel® Xeon® Scalable processors. Intel® Dynamic Load Balancer enables the efficient distribution of network processing across multiple CPU cores/threads. It dynamically distributes network data across multiple CPU cores for processing as the system load varies. Intel DLB also restores the order of networking data packets processed simultaneously on CPU cores.

A significant speedup with the Intel DLB accelerator is for microservice environments, where thousands of requests come in per second.

a. Business Value
    i. Increase performance through optimization for environments where there are many requests for work to be done across cores.

    ii. In 5G environments, using Intel DLB will reduce latencies and lead to a better end-user experience.

b. Maximum Number Per CPU: 4

c. Performance Measurement for Networking and Storage operations:
    i. Up to 96% lower latency at the same throughput (RPS) with Intel DLB vs. software for Istio ingress gateway working on six cores/12 threads[4].

d. Scalability:

You can increase your product line performance and scale with the acceleration you need (scaling from 1 to 4 Intel DLB devices on-chip)

Intel® In-Memory Analytics Accelerator (Intel® IAA)

Run database and analytics workloads faster, with potentially greater power efficiency. Intel® IAA increases query throughput and decreases the memory footprint for in-memory database and big data analytics workloads. Intel IAA is ideal for in-memory, open-source, and data stores such as RocksDB and ClickHouse.

a. Business Value
   i. Increase the performance of databases to meet SLAs.
   ii. Decrease the memory footprint for in-memory databases
   iii. Give users faster analytics results

b. Maximum Number Per CPU: 4

c. Performance Measurements for Database Applications
   i. Up to 3x higher RocksDB performance using integrated Intel® In-Memory Analytics (Intel® IAA) vs. the prior generation[5].
   ii. Up to 59% higher ClickHouse DB performance using integrated Intel IAA vs. the prior generation.[2]
   iii. For Microsoft SQL Server, 4th Gen Intel Xeon Scalable processors can deliver:
      i. Up to 53% faster backup enabled by Intel® QuickAssist Technology (Intel® QAT).[3] Intel QAT is a built-in accelerator that helps reduce system resource consumption and TCO by accelerating cryptography and data compression.
      ii. A relative performance gain of up to 22% more NOPM transactions and up to 19% faster query response time compared to the previous generation[6].

d. Scalability:
   You can build your product lines performance scale with the acceleration you need (scaling from 1 to 4 Intel IAA devices on-chip)

Intel® Data Streaming Accelerator (Intel® DSA)

Drive high performance for storage, networking, and data-intensive workloads by improving streaming data movement and transformation operations. Intel® Data Streaming Accelerator (Intel® DSA) is designed to offload the most common data movement tasks that cause overhead in data center-scale deployments. Intel DSA helps speed up data movement across the CPU, memory, and caches, as well as all attached memory, storage, and network devices.

a. Business Value
   i. Faster movement of data throughout the system
   ii. Reduces overall processing time for applications that require large amounts of data movement

b. Maximum Number Per CPU: 4

c. Performance Measurements for applications:
   i. Networking: Intel® Data Streaming Accelerator Up to 95% higher vSwitch throughput for packet sizes above ~800B for 200Gbps bi-directional switching with built-in Intel® Data Streaming Accelerator (Intel® DSA) compared to existing software only implementation[7].
   ii. Storage – Intel® Data Streaming Accelerator Up to 60% higher IOPs and up to 37% latency reduction for large packet sequential read using integrated Intel® Data Streaming Accelerator (Intel® DSA) vs. the prior generation[8].

d. Scalability:
   You can build your product lines performance scale with the acceleration you need (scaling from 1 to 4 Intel DSA devices on-chip)

Intel Advanced Matrix Extensions (Intel® AMX)

Significantly accelerate AI capabilities on the CPU with Intel® Advanced Matrix Extensions (Intel® AMX). Intel AMX is a built-in accelerator that improves the performance of deep learning training and inference on 4th Gen Intel® Xeon® Scalable processors, ideal for workloads like natural language processing, recommendation systems, and image recognition. In addition, HPC applications will benefit when matrix operations must be performed.

a. Business Value
   i. Faster computation of matrix operations that are at the heart of AI applications.
   ii. Frees up CPU cores for other work

b. Maximum Number per CPU: 4

c. Performance Measurements for Deep Learning Applications:
   i. Intel® Advanced Matrix Extensions Up to 10x higher PyTorch real-time inference performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)[9]
   ii. Deep learning training – Intel® Advanced Matrix Extensions Up to 10x higher PyTorch training performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)[10].

d. Scalability:
   Up to four AMX accelerators can be used within an application.

Conclusion

While new CPUs from Intel continue to perform significantly better than previous generations for general workloads, there are significant opportunities to accelerate specific workloads. Intel Accelerator Engines built into the 4th Gen Intel Xeon Scalable processor allow developers and end users to achieve up to 10X the performance of workloads compared to the previous generation of Intel Xeon processors, which far exceeds the increase in cores, clock rates, and architectural improvements. In addition, using the new Intel Accelerator Engines increases performance, efficiency, and offloading processing from the CPU cores.

References

1 - https://edc.intel.com/content/www/us/en/products/performance/benchmarks/overview/

2 - See [N15,16] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/

3 - See [N16] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

4 - See [W6] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

5 - See [D1] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

6 - See [D17] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

7 - See [W8] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

8 - See [N18] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

9 - See [A17] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

10 - See [A16] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.