Product Brief

4th Gen Intel® Xeon® Scalable Processors

Leading performance with the most built-in accelerators

It's now more critical than ever for technology to deliver business value as organizations look to scale, drive down costs, and deliver new services. Instead of customizing systems for new applications, which can add complexity, enterprises can achieve the performance needed to meet a wide variety of deployments—both today and in the future—with a scalable platform.

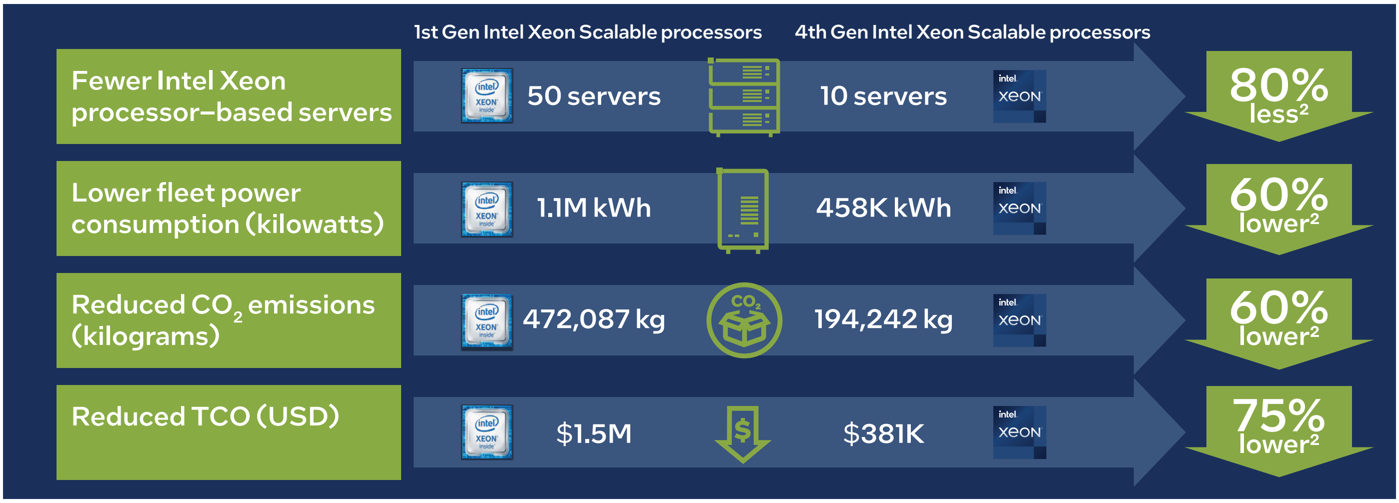
4th Gen Intel Xeon Scalable processors feature Intel® Accelerator Engines designed to accelerate performance across the fastest-growing workloads. These processors have the most built-in accelerators of any CPU on the market to help improve performance efficiency for emerging workloads, especially those powered by AI.[1]

In addition to performance improvements, 4th Gen Intel Xeon Scalable processors have advanced security technologies to help protect data in an ever-changing landscape of threats while unlocking new opportunities for business insights. Together with the largest ecosystem of partners, Intel makes it easier for enterprises to stay competitive, offering the most choice to scale infrastructure and quickly achieve business value.

Reduce your total cost of ownership (TCO)

Intel Accelerator Engines offer an alternative, more efficient way to achieve higher performance and increase virtual and physical CPU utilization. They allow you to improve performance without having to purchase additional specialized hardware. They can also help improve power efficiency by offloading common tasks from the embedded CPU cores on a chip, boosting overall application performance while reducing power usage. The efficiency of 4th Gen Intel Xeon Scalable processors with built-in accelerators delivers massive total cost of ownership (TCO) value, offering up to 75 percent lower TCO than the prior generation of Intel Xeon Scalable processors.[2]

Whether you are refreshing older hardware used for conventional compute scenarios or building a foundation for scalable and efficient operations in emerging processing-intensive workloads like AI, HPC, and analytics, Intel is the right partner to help optimize the power efficiency of your data centers and reduce hardware waste.



	1st Gen Intel Xeon Scalable processors	4th Gen Intel Xeon Scalable processors	
Fewer Intel Xeon processor–based servers	50 servers	10 servers	80% less[2]
Lower fleet power consumption (kilowatts)	1.1M kWh	458K kWh	60% lower[2]
Reduced CO_2 emissions (kilograms)	472,087 kg	194,242 kg	60% lower[2]
Reduced TCO (USD)	$1.5M	$381K	75% lower[2]

Figure 1. Moving to the 4th Gen Intel Xeon Gold 5420+ processor from the 1st Gen Intel Xeon Silver 4110 processor yields significant savings[2]

Intel Accelerator Engines redefine performance and power efficiency

Redefine what you expect from a processor. Built-in acceleration is an alternative, more efficient way to achieve higher performance than growing the CPU core count. With all-new accelerated matrix multiply operations, 4th Gen Intel Xeon Scalable processors have exceptional AI training and inference performance. Other integrated accelerators speed up data movement, encryption, and compression for faster networking and storage, boost query throughput for more responsive analytics, and offload scheduling and queue management to dynamically balance loads across multiple cores.

The latest Intel Accelerator Engines and software optimizations help improve power efficiency across AI, data analytics, networking, and storage. The result is that you can achieve a 3x average performance-per-watt efficiency improvement for targeted workloads using built-in accelerators with 4th Gen Intel Xeon Scalable proccessors, compared to the 3rd Gen Intel Xeon Scalable processors.[3]

- **Intel® Advanced Matrix Extensions (Intel® AMX)** accelerates deep learning (DL) inference and training workloads, such as natural language processing (NLP), recommendation systems, and image recognition.
- **Intel® Data Streaming Accelerator (Intel® DSA)** drives high performance for storage, networking, and data-intensive workloads by improving streaming data movement and transformation operations.
- **Intel® In-Memory Analytics Accelerator (Intel® IAA)** improves analytics performance while offloading tasks from CPU cores to accelerate database query throughput and other workloads.
- **Intel® Dynamic Load Balancer (Intel® DLB)** provides efficient hardware-based load balancing by dynamically distributing network data across multiple CPU cores as the system load varies.
- **Intel® Advanced Vector Extensions (Intel® AVX) for vRAN** increases virtual radio access network (vRAN) density up to 2x, compared to the previous generation, with the same power envelope.[3]
- **Intel® Advanced Vector Extensions 512 (Intel® AVX-512)** supports up to two fused-multiply add (FMA) units and includes optimizations to accelerate performance for demanding computational tasks.
- **Intel® QuickAssist Technology (Intel® QAT)** accelerates encryption, decryption, and data compression, offloading these tasks from the processor core to help reduce system resource consumption.
- **Intel® Crypto Acceleration** reduces the penalty of implementing pervasive data encryption and increases the performance of encryption-sensitive workloads, such as for Secure Sockets Layer (SSL) web servers, 5G infrastructure, and VPNs/firewalls.

AI

With accelerated vector instructions and matrix multiply operations, 4th Gen Intel Xeon Scalable processors provide exceptional AI inference and training performance. Intel AMX can provide a substantial performance increase for DL workloads, such as recommendation systems, NLP, image recognition, media processing and delivery, and media analytics.

HPC

4th Gen Intel Xeon Scalable processors are ready to improve performance for the highly threaded code common in HPC workloads found in manufacturing simulations, molecular dynamics, earth systems modeling, and AI inferencing and training. Built-in accelerators provide high levels of precision while speeding up processing of AI datatypes. And support for DDR5 memory, PCIe Gen5, Intel® Ultra Path Interconnect (Intel® UPI) 2.0, and Compute Express Link (CXL) also enhances overall data throughput.

Data analytics

Built-in accelerators enhance performance for in-memory databases, big data, data warehousing, business intelligence (BI), enterprise resource planning (ERP), and operational analytics. Intel DSA improves the streaming data movement and transformation operations common in data processing–intensive applications, while Intel IAA offloads tasks from CPU cores to accelerate throughput for database operations.

Network and storage

Intel DLB balances operations between cores and provides network-packet prioritization. Intel DSA offloads data-copy and common data-transformation operations to free up core cycles. These built-in accelerators enhance cloud computing by enabling efficient network data placement and enterprise storage data movement, and through improved memory-management operations in cloud computing.

Encryption

Intel QAT, now built into 4th Gen Intel Xeon Scalable processors, accelerates cryptography and compression. Intel QAT can significantly boost CPU efficiency and application throughput, while reducing data footprint and power utilization, enabling organizations to strengthen encryption without sacrificing performance.

Security

Intel® Software Guard Extensions (Intel® SGX) is the most researched, updated, and deployed confidential computing technology in data centers on the market today, with the smallest trust boundary of any confidential computing technology in the data center.

4th Gen Intel Xeon Scalable processors compared to 3rd Gen Intel Xeon Scalable processors:

Up to **1.53x**

average performance gain over the previous generation[4]

Up to **10x**

higher PyTorch performance for both real-time inference and training with built-in Intel AMX (BF16) versus the previous generation (FP32)[5]

Up to **3x**

higher RocksDB performance using integrated Intel IAA versus the previous generation[6]

Up to **1.6x**

higher input/output operations per second (IOPS) and up to 37% latency reduction for large packet sequential reads using integrated Intel DSA versus the previous generation[7]

Up to **2x**

the capacity at the same power envelope for vRAN workloads versus previous-generation processors[8]

Up to **95%**

fewer cores and 2x higher level-1 compression throughput using integrated Intel QAT versus previous-generation processors[9]

Technology overview

4th Gen Intel Xeon Scalable processors feature a new architecture with higher per-core performance than the previous generation. They also feature up to 60 cores per socket and one, two, four, or eight sockets per system. To balance those core-count increases, the platform provides accompanying advances in the memory and input/output (I/O) subsystems. DDR5 memory provides up to 1.5x the bandwidth and speed of DDR4, for 4,800 megatransfers per second (MT/s).[10] The platform also features 80 lanes of PCIe Gen5 per socket, for dramatically improved I/O compared to earlier platforms.[11] It provides CXL 1.1 to support high fabric bandwidth and attached accelerator efficiency. 4th Gen Intel Xeon Scalable processors support technologies that let you scale and adapt as workload requirements change. They also enable you to:

- Further boost networking, storage, and compute performance, while improving CPU utilization, by offloading heavy tasks to an Intel® Infrastructure Processing Unit (Intel® IPU).
- Increase multi-socket bandwidth with Intel UPI 2.0 (up to 16 gigatransfers per second [GT/s]).
- Configure your CPU to meet specific workload needs with Intel® Speed Select Technology (Intel® SST).
- Increase shared last-level cache (LLC) (up to 100 MB LLC shared across all cores).
- Strengthen your security posture with hardware-enhanced security.
- Eliminate the need for a separate RAID card with Intel® Virtual RAID on CPU (Intel® VROC).

Capabilities in 4th Gen Intel Xeon Scalable processors

PCI Express Gen5 (PCIe 5.0)

Unlock new I/O speeds with opportunities to enable the highest possible throughput between the CPU and connected devices. 4th Gen Intel Xeon Scalable processors have up to 80 lanes of PCIe 5.0—ideal for fast networking, high-bandwidth accelerators, and high-performance storage devices. PCIe 5.0 doubles the I/O bandwidth from PCIe 4.0,[11] maintains backward compatibility, and provides foundational slots for CXL.

DDR5

Improve compute performance by overcoming data bottlenecks with higher memory bandwidth. DDR5 offers up to 1.5x bandwidth improvement over DDR4,[12] enabling opportunities to improve performance, capacity, power efficiency, and cost. 4th Gen Intel Xeon Scalable processors offer up to 4,800 MT/s (1 DPC) or 4,400 MT/s (2 DPC) with DDR5.

CXL

Reduce compute latency in the data center and help lower total cost of ownership (TCO) with CXL 1.1 for next-generation workloads. CXL is an alternate protocol that runs across the standard PCIe physical layer and can support both standard PCIe devices and CXL devices on the same link. CXL provides a critical capability to create a unified, coherent memory space between CPUs and accelerators, and it will revolutionize how data center server architectures will be built for years to come.

Scale with the most choice and flexibility—Intel Xeon Scalable processors

From hardware to systems to software, Intel provides a trusted foundation of technology designed to help organizations meet an ever-expanding set of business goals while keeping data more secure. Whether it's delivering greater compute density to reduce power footprint, accelerating AI workflows, or supporting a transition to cloud-native architecture, Intel Xeon Scalable processors help solve the most important business challenges while offering the greatest cloud choice and application portability.

Unlock hardware performance with oneAPI software tools

Intel is committed to an open software strategy, designed with developers in mind so that the investments made in Intel® technologies continue to add value in future generations. Intel® oneAPI toolkits are a comprehensive set of advanced compilers, libraries, and analysis, debug, and porting tools for Intel architecture. These toolkits enable cutting-edge features for hardware performance, and they help reduce software development and maintenance costs. Intel oneAPI toolkits enable developers to optimize features, including built-in accelerators, on 4th Gen Intel Xeon Scalable processors.

Overview of 4th Gen Intel Xeon Scalable processors

Intel Xeon Platinum 8400 processors are the foundation for security-enabled, agile, and hybrid cloud data centers. They are designed for advanced data analytics, AI, high-density infrastructure, and multicloud workloads. These processors deliver high levels of performance, platform capabilities, and industry-leading workload acceleration. They offer enhanced hardware-based security and exceptional multi-socket processing performance—with up to 8 sockets on select Intel Xeon Platinum 8400 processors. With trusted, hardware-enhanced data-service delivery and new I/O and connectivity technologies, these processors deliver improvements in I/O, memory, storage, and network technologies to harness actionable insights from the increasingly data-fueled world, including:

- Up to 60 cores per Intel Xeon Scalable processor
- 8 memory channels per processor at up to 4,800 MT/s (1 DPC)
- AI acceleration with Intel AMX for a giant leap in DL inference and training performance

With up to four-socket scalability,[13] **Intel Xeon Gold 6400 and Intel Xeon Gold 5400 processors** are optimized for demanding mainstream data center, multicloud compute, and network and storage workloads. With support for higher memory speeds and enhanced memory capacity, these processors deliver improved performance, enhanced memory capabilities, hardware-enhanced security, and workload acceleration.

Intel Xeon Silver 4400 processors deliver essential performance, improved memory speed, and power efficiency. They offer the hardware-enhanced performance required for entry-level data center compute, network, and storage.

Learn more

For more about how 4th Gen Intel Xeon Scalable processors can advance your business, visit intel.com/xeonscalable and intel.com/4thgenxeon.

intel. xeon PLATINUM

Up to 8-socket scalability

Four Intel UPI ports at 16 GT/s

80 lanes of PCIe 5.0 with CXL

DDR5 at up to 4,800 MT/s (1 DIMM per channel) or 4,400 MT/s (2 DIMMs per channel)

Intel AVX-512 (two 512-bit FMAs)

Intel® Hyper-Threading Technology (Intel® HT Technology) and Intel® Turbo Boost Technology

Intel AMX

Intel SST

Advanced reliability, availability, and serviceability (RAS) capabilities

Intel SGX up to 128 GB max enclave size (up to 512 GB max enclave size on select SKUs)

Workload acceleration with Intel QAT, Intel DLB, Intel DSA, and Intel IAA

intel. xeon GOLD

Up to 4-socket scalability

Three Intel UPI ports at 16 GT/s

80 lanes of PCIe 5.0 with CXL

DDR5 at up to 4,800 MT/s (1 DIMM per channel) or 4,400 MT/s (2 DIMMs per channel)

Intel AVX-512 (two 512-bit FMAs)

Intel HT Technology and Intel Turbo Boost Technology

Intel® Deep Learning Boost (Intel® DL Boost) and Intel AMX

Intel SST

Advanced RAS capabilities

Intel SGX up to 128 GB max enclave size

Workload acceleration with Intel QAT, Intel DLB, Intel DSA, and Intel IAA

intel. xeon SILVER

Up to 2-socket scalability

Two Intel UPI ports at 16 GT/s

80 lanes of PCIe 5.0 with CXL

DDR5 at up to 4,800 MT/s (1 DIMM per channel) or 4,400 MT/s (2 DIMMs per channel)

Intel AVX-512 (two 512-bit FMAs)

Intel HT Technology and Intel Turbo Boost Technology

Intel DL Boost and Intel AMX

Intel SGX up to 64 GB max enclave size

Workload acceleration with Intel QAT, Intel DLB, Intel DSA, and Intel IAA

[1] Intel. Intel® Accelerator Engines web page. intel.com/content/www/us/en/products/docs/accelerator-engines/overview.html.

[2] Comparing benefits transitioning from Intel Xeon Silver 4110 processors to Intel Xeon Gold 5420+ processors. Performance varies by use, configuration, and other factors. Calculations as of March 28, 2023, based on the Intel node TCO and power calculator using default cost, power, and TCO assumptions over a five-year TCO horizon, and comparing replacing 50 older servers equipped with Intel Xeon Silver 4110 processors with new servers equipped with Intel Xeon Gold 5420+ processors. Results may vary. Performance measurements based on published SPECrate2017_int_base on spec.org as of March 28, 2023, and October 2020: spec.org/cpu2017/results/res2020q4/cpu2017-20201015-24218.html.

[3] See [E1] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

[4] See [G1] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

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

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

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

[8] See [N9] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

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

[10] See [G2] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

[11] 4th Gen Intel Xeon Scalable processor: 80 lanes of PCIe 5.0 with flex bus/CXL per CPU vs. 3rd Gen Intel Xeon Scalable processor: 64 lanes of PCIe 4.0 per CPU.

[12] 4th Gen Intel Xeon Scalable processor: 8 channels DDR5, up to 4,800 MT/s (1 DPC) vs. 3rd Gen Intel Xeon Scalable processor: 8 channels DDR4, 3,200 MT/s (2 DPC).

[13] Up to four-socket scalability available only on select Intel Xeon Gold 6400 processors.

Availability of accelerators varies depending on SKU. Visit the Intel Product Specifications page for additional product details.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at www.intc.com.

Printed in USA 0923/MG/PRW/PDF Please Recycle 352722-002US