intel® XEON®

Improve performance with Intel® Xeon® Scalable processors featuring built-in accelerators

E-Guide

Table of contents

What are built-in accelerators and why should you use them?

What if, instead of building customized workload solutions every time you need to establish new capabilities, you could rely on the technology already built into your CPU? With Intel® Xeon® Scalable processors, you can. These trusted CPUs you already depend on to run all your workloads are supporting growing data center needs with Intel® Accelerator Engines, which deliver enhanced performance across various emerging workloads.

In this guide, we'll explore how these purpose-built, integrated accelerators support today's most demanding workloads spanning AI, security, HPC, networking, analytics and storage.

The real-world benefits of Intel Accelerator Engines

Whether you're using Intel Xeon Scalable processors for your workloads on-prem, in the cloud or at the edge, Intel Accelerator Engines can help your business reach new heights. These built-in accelerators are designed to provide a range of benefits including **increased application performance, reduced costs and improved power efficiency:**

Performance	Intel Accelerator Engines are designed to help deliver higher performance on targeted workloads.
Cost savings	Intel Accelerator Engines allow you to improve performance without having to purchase additional specialized hardware.
Power savings	Intel Accelerator Engines can help improve power efficiency by offloading common tasks from the embedded CPU cores on chip, boosting overall application performance while minimizing power usage to meet your sustainability goals.

Which Intel Accelerator Engines are right for your business?

With the most built-in accelerators of any CPU on the market, 4th Gen Intel® Xeon® Scalable processors are designed to improve performance and help protect data across today's fastest-growing workloads. To help you decide which Intel Accelerator Engines can best support your business, let's look at our top offerings across six key categories: AI, security, HPC, networking, analytics and storage.

Intel® AI Engines
Intel® Advanced Matrix Extensions (Intel® AMX)
Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

Intel® Security Engines
Intel® Software Guard Extensions (Intel® SGX)
Intel® Trust Domain Extensions (Intel® TDX)
Intel® Crypto Acceleration

Intel® HPC Engines
Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
Intel® Advanced Matrix Extensions (Intel® AMX)
Intel® Data Streaming Accelerator (Intel® DSA)

Intel® Network Engines
Intel® QuickAssist Technology (Intel® QAT)
Intel® vRAN Boost
Intel® Dynamic Load Balancer (Intel® DLB)

Intel® Analytics Engines
Intel® In-Memory Analytics Accelerator (Intel® IAA)
Intel® Data Streaming Accelerator (Intel® DSA)
Intel® QuickAssist Technology (Intel® QAT)

Intel® Storage Engines
Intel® Data Streaming Accelerator (Intel® DSA)
Intel® QuickAssist Technology (Intel® QAT)
Intel® Volume Management Device (Intel® VMD)

Intel AI Engines

Intel Advanced Matrix Extensions (Intel AMX)

Intel AMX is Intel's next-generation advancement for deep-learning training and inference on 4th Gen Intel Xeon Scalable processors. Ideal for workloads like natural language processing, recommendation systems and image recognition, this new feature extends the built-in AI acceleration capabilities of previous Intel Xeon Scalable processors while also offering significant performance gains.

Up to

10x higher PyTorch performance

for both real-time inference and training workloads compared to the previous generation (FP32).[1]

Intel Advanced Vector Extensions 512 (Intel AVX-512)

A continuing feature of Intel Xeon Scalable processors, Intel AVX-512 is a general-purpose performance-enhancing accelerator with a wide range of uses. When it comes to AI, Intel AVX-512 can accelerate machine learning workloads for training and inferencing. 4th Gen Intel Xeon Scalable processors with Intel AVX-512 are also designed to speed up data preprocessing.

Intel Security Engines

Intel Software Guard Extensions (Intel SGX)

Intel SGX is the most researched, updated and deployed confidential computing technology in data centers on the market today. This continuing feature of Intel Xeon Scalable processors provides the foundation for confidential computing solutions across edge and multi-cloud infrastructures.

Intel SGX offers a hardware-based security solution that is designed to prohibit access to protected data in use via unique application-isolation technology. By helping protect selected code and data from inspection or modification, Intel SGX allows developers to run sensitive data operations inside enclaves to help increase application security and protect data confidentiality.

Intel Trust Domain Extensions (Intel TDX)

Intel TDX is a new capability available through select cloud providers in 2023 that offers increased confidentiality at the virtual machine (VM) level, enhancing privacy and control over your data. Within an Intel TDX confidential VM, the guest OS and VM applications are isolated from access by the cloud host, hypervisor and other VMs on the platform.

Intel Crypto Acceleration

Intel Crypto Acceleration uses single instruction, multiple data (SIMD) techniques to process more encryption operations in every clock cycle. This can help increase the total throughput of applications that require strong data encryption, with minimal impact on performance and user experience.

Intel HPC Engines

Intel Advanced Vector Extensions 512 (Intel AVX-512)

With ultrawide 512-bit vector operations capabilities, Intel AVX-512 is especially suited to handle the most demanding computational tasks commonly encountered in HPC applications. It's used by organizations across educational, financial, enterprise, engineering and medical industries. By enabling users to run complex workloads on existing hardware, Intel AVX-512 accelerates performance for tasks like financial analytics, 3D modeling and scientific simulations.

Intel Advanced Matrix Extensions (Intel AMX)

In addition to accelerating AI workloads, Intel AMX is also designed to deliver performance gains across popular HPC workloads. This new built-in accelerator transforms large matrix math calculations into a single operation and uses a two-dimensional register file to store large chunks of data.

Intel Data Streaming Accelerator (Intel DSA)

Intel DSA is a new feature designed to optimize and speed up streaming data movement and transformation operations common in networking, data-processing-intensive applications and high-performance storage. Intel DSA accelerates HPC workloads by offloading the most common data movement tasks that cause CPU overhead in data-center-scale deployments.

Up to

1.56x higher average HPC performance

on 4th Gen Intel Xeon Scalable platform vs. prior gen[2]

Intel Network Engines

Intel QuickAssist Technology (Intel QAT)

Intel QAT boosts performance to meet the demands of today's networking workloads, helping systems serve more clients. It can deliver significant workload acceleration for cryptography, including symmetric and asymmetric encryption and decryption.

Intel QAT using RSA4K can increase client density on an open-source NGINX web server compared to software running on CPU cores without acceleration.

Intel vRAN Boost

Intel vRAN Boost is a new feature that eliminates the need for an external accelerator card by integrating vRAN acceleration directly into the Intel 4th Gen Xeon Scalable processor. By reducing vRAN component requirements, it reduces overall vRAN solution complexity and provides power savings for customers.

Up to

2x the capacity

at the same power envelope for vRAN workloads vs. prior-generation processors[3]

Intel Dynamic Load Balancer (Intel DLB)

Intel DLB is a new feature that enables the efficient distribution of network processing across multiple CPU cores. It also restores the order of networking data packets processed simultaneously on CPU cores.

With Intel DLB, customers can gain higher performance on packet forwarding compared to software queue management on cores without acceleration. Additionally, applications can achieve higher packet processing performance than the previous generation.
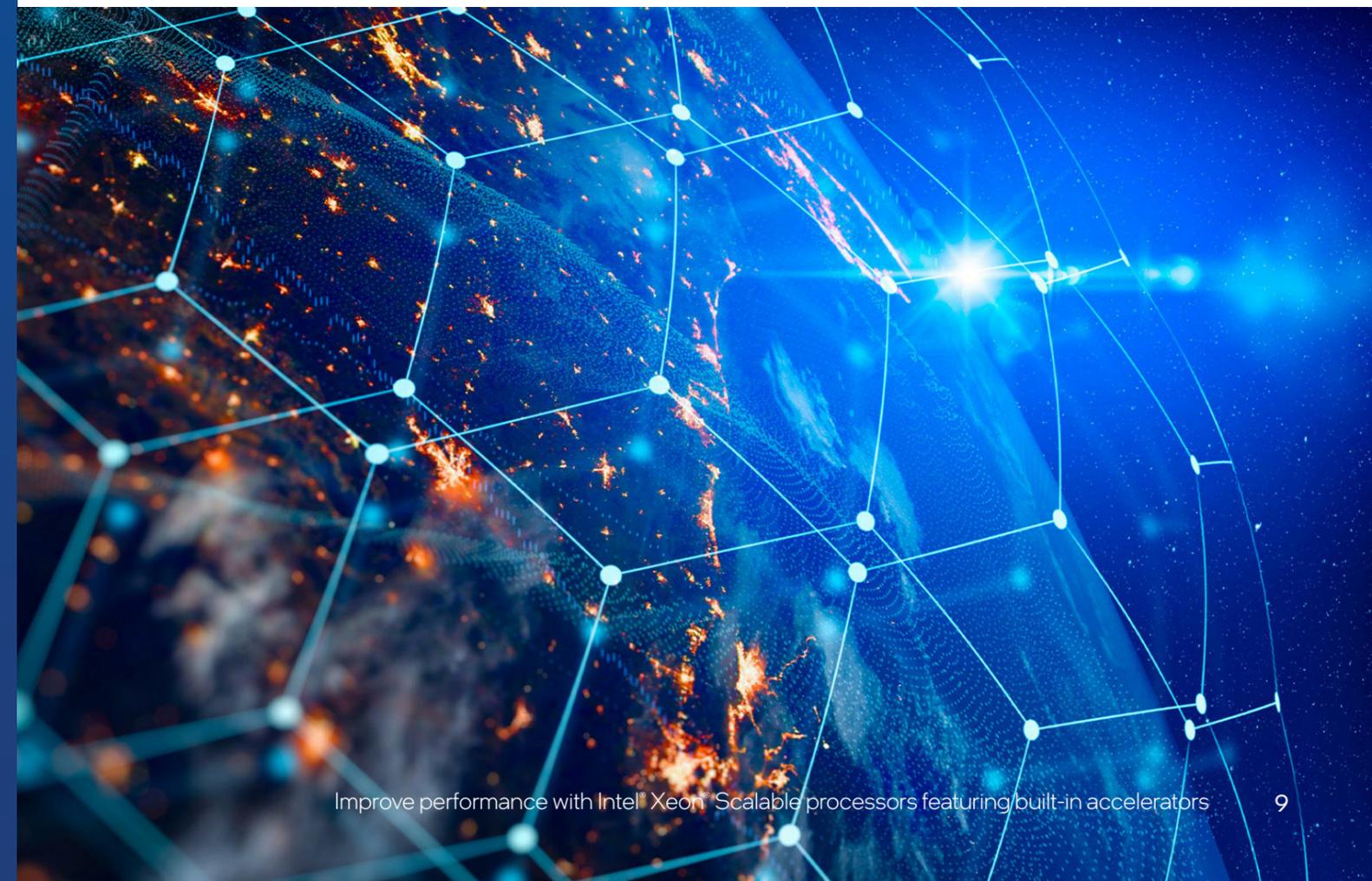
Intel Analytics Engines

Intel In-Memory Analytics Accelerator (Intel IAA)

Intel IAA is designed to accelerate database and analytics performance. By increasing query throughput and decreasing the memory footprint for in-memory databases and advanced analytics workloads, this new feature can provide faster data movement and improve CPU core utilization by reducing dependency on CPU cores.

Intel IAA is ideal for in-memory databases, open-source databases and data stores like RocksDB, Redis, Cassandra and MySQL. Customers using 4th Gen Intel Xeon Platinum 8490H with Intel IAA will gain up to 3x higher RocksDB performance compared to the previous generation.[4]

Intel Data Streaming Accelerator (Intel DSA)

Intel DSA is a new feature designed to optimize and speed up streaming data movement and transformation operations common in data-intensive applications, driving better business outcomes. By offloading tasks like data movement, data copying and error checking, Intel DSA enables the CPU to focus on business-critical database functions or other analytics workloads. This reduces query latencies and increases throughput, delivering faster data processing.

Intel QuickAssist Technology (Intel QAT)

Intel QAT can be used to accelerate database backups. When QAT was enabled as a new feature in SQL Server 2022, it delivered more efficient and higher performance applications.

With Intel QAT, SQL Server customers achieve up to 2.3x faster backup operations and a 6% reduction in backup storage capacity.[5]

Intel Storage Engines

Intel Data Streaming Accelerator (Intel DSA)

Intel DSA works on the CPU — between DRAM, caches and processor cores — and extends across I/Os to attached memory and storage, plus networked resources. As Intel's next-generation direct memory access (DMA) engine, it speeds transfers between volatile and persistent memory and supports virtualized memory and I/Os.

Customers using 4th Gen Intel Xeon Platinum 8490H with Intel DSA will gain up to 1.6x higher IOPS and up to 37% latency reduction for large packet sequential read compared to the previous generation.[6]

Intel QuickAssist Technology (Intel QAT)

Intel QAT increases performance of storage workloads and applications by accelerating cryptography and data compression/decompression. Customers using 4th Gen Intel Xeon Platinum 8490H with Intel QAT will gain up to 95% fewer cores and 2x higher level 1 compression throughput compared to the previous generation.[7]

Intel Volume Management Device (Intel VMD)

Intel VMD is a legacy feature that enables direct control and management of NVMe SSDs from the PCIe bus without the need for additional hardware adapters. It allows for a smoother and lower-cost transition to fast NVMe storage while limiting the downtime of critical infrastructure. With benefits like bootable RAID, robust surprise hot-plug and blink status LED, Intel VMD increases serviceability and enables you to deploy next-generation storage with confidence.

Learn more about what Intel Accelerator Engines on Intel Xeon Scalable processors can accomplish by visiting the [Intel Products page](#).

Conclusion

There are numerous workloads for which Intel Accelerator Engines deliver greater business value to customers. Whether you're looking to increase performance, support sustainability initiatives or help ensure the protection of your most sensitive data, Intel Xeon processors with Intel Accelerator Engines offer a wide range of solutions without the need for additional hardware.

With the latest Intel Xeon Scalable processors, Intel redefines what customers can expect from a CPU — and delivers leading performance with the most built-in accelerators.

Learn more about Intel Xeon Scalable processors by visiting
https://www.intel.com/xeonscalable.

[1]See [A16] and [A17] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
[2]See [H1] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary
[3]See [N10] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
[4]See [D1] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
[5]SOFTWARE BASELINE: Test by Intel as of 10/26/2021. 1-node, 2x Intel® Xeon® Platinum 8358 Processor on Wilson City (ref. platform), 32 cores
 with 1024 GB (16 slots/ 64GB/ 3200[3200]) total DDR4 memory, ucode 0xD000323, HT on, Turbo on, Windows Server 2019,10.0.17763.1999,
 [Hyper-V VM], 1x Intel X550, SQL Server 2022 CTP1.1 private test build, 525GB Database, Database Backup with Compression.
[6]See [N18] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
[7]See [N16] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

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.

For workloads and configurations, visit 4th Gen Xeon Scalable processors at www.intel.com/processorclaims. Results may vary.

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

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

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

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