Product Brief

HPC Accelerated
Intel® Xeon® Scalable processors

Accelerate the most demanding computational tasks with Intel® HPC engines

Accelerated HPC with 4th Gen Xeon

Up to

1.56x higher

LAMMPS performance on 4th Gen Intel Xeon Scalable processor vs. prior gen[1]

Customer success: Real-world acceleration on Intel Xeon Scalable processors

The University at Buffalo upgrades Industry Compute Cluster; shares HPC and AI with community.

Read the story ›

CERN speeds particle accelerator simulations with Intel Xeon Scalable processors and Intel® Deep Learning Boost.

Read the story ›

High-performance computing (HPC) is essential to scientific discovery, engineering simulations and the modeling of complex systems. Acceleration can provide an efficient and effective alternative to achieving high performance in place of growing the CPU core count. The 4th Generation Intel® Xeon® Scalable processors come equipped with purpose-built accelerators that can elevate HPC workload performance and power efficiency by offloading tasks to these acceleration features.

High-performance computing is entering a new era

Over the years, we've seen a shift in how HPC owners address the need for increased compute speed and access while controlling cost. Industries are increasingly turning to HPC as a tool to help arrive at business insights faster to make critical business decisions — all while lowering their costs.

The new and improved features of 4th Gen Intel Xeon Scalable processors with Intel® HPC engines can increase performance across the fastest-growing workload types, like simulation and modeling. The 4th Gen Intel Xeon Scalable processors will deliver improved performance, efficiency and cost savings for targeted workloads of built-in accelerators: Intel® Advanced Matrix Extensions (Intel® AMX), Intel® Data Streaming Accelerator (Intel® DSA) and Intel® QuickAssist Technology (Intel® QAT).

Intel® Advanced Matrix Extensions (Intel® AMX) AI accelerator

Machine learning (ML) technologies are making workloads more efficient, effective and insightful. Industry trends are driving customers to HPC- and AI-powered solutions to benefit their businesses. Intel has introduced a new accelerator designed to boost AI performance. In doing so, Intel is sharing its expertise in AI with customers that are utilizing both HPC and AI solutions.

Intel AMX, one of the new built-in accelerator engines integrated into 4th Gen Intel Xeon Scalable processors, is Intel's next-generation advancement for deep-learning inference and training performance. Intel extended the built-in AI acceleration capabilities of earlier Intel Xeon Scalable processors, enabling Intel AMX to transform the large matrix math calculations that are at the heart of deep-learning workloads into a single operation. Intel AMX also uses a two-dimension register file to store larger chunks of data. Built to accelerate AI workloads, Intel AMX will be critical to delivering performance across workloads where HPC and AI converge.

Intel® Advanced Vector Extensions 512 (Intel® AVX-512) — the foundation for faster HPC

Every x86 CPU shares a common instruction set architecture (ISA). Intel has extended the base x86 instructions to new workloads and expanded their capabilities generation after generation, starting with Intel® Advanced Vector Extensions (Intel® AVX) in 2011. Today those original Intel AVX instructions — plus their descendants, Intel AVX-512® and Intel AVX2® — accelerate general computing, AI processing and mathematically intense HPC workloads. For instance, by providing higher performance per core, 3rd Gen Intel Xeon Scalable processors delivered significantly better performance for computer-aided engineering (CAE) applications than previous generations.[2]

Fewer steps means faster processing

The "extensions" in Intel AVX-512 condense, combine and fuse common computing operations into fewer steps. As a primitive example, you could instruct a CPU to calculate 3 x 3 x 3 x 3 x 3, which would take five clock cycles. Or you could create an instruction for 3^5 that the CPU can do in one cycle. Intel AVX-512 takes that logic and applies it to hundreds of task-specific operations, like fused multiply-add (FMA). Intel Xeon Scalable processors have two FMA units per core to combine multiplication and addition into a single operation and accelerate computation speeds.

Intel® Deep Learning Boost (Intel® DL Boost) — neural network acceleration for HPC

Machine learning along with deep-learning inference and training are expanding the capabilities of HPC by accelerating processing speeds, increasing accuracy and creating entirely new methods for modeling and analysis. Intel DL Boost includes Vector Neural Network Instruction (VNNI), which is a combination of three AVX-512 instructions designed to maximize compute resources and cache utilization and reduce the number of operations per clock cycle. VNNI can also overcome potential bandwidth bottlenecks to accelerate deep-learning inference workloads. Intel DL Boost supports INT8 (VNNI) and BF16 precision data types. By combining three instruction sets into one, Intel DL Boost can help deliver higher average AI performance across workloads.

Intel DL Boost extensions help accelerate AI performance without expanding memory requirements to support AI convergence on HPC systems.

Intel® QuickAssist Technology (Intel® QAT)

Free up space and reduce costs by offloading compute-intensive workloads with Intel QAT, a new built-in accelerator for 4th Gen Intel Xeon Scalable processors. Intel QAT reduces system resource consumption by providing accelerated cryptography, key protection and data compression. In doing so, it benefits customers by offering more Gbps and Ops/Sec performance in big-data and database applications.

Intel QAT reduces the overhead often associated with encryption and compression, ultimately playing a significant role in improving cluster performance.

Intel QAT allows each core to serve more clients by improving performance for cryptography and data compression while also reducing the data footprint.

Intel® Data Streaming Accelerator (Intel® DSA)

The journey of data — traveling in and out of memory, storage and networking subsystems — can prove burdensome on the CPU.

Intel DSA, an accelerator integrated into Intel Xeon processors, delivers high performance for storage and networking and for data-intensive workloads by improving streaming-data movement and transformation operations. Intel DSA helps speed up data movement across the CPU, memory and caches, and across all attached memory, storage and network devices.

Intel DSA can improve performance, and it reduces latency for Open vSwitch (OVS) to enable more effective network automation while freeing up CPU cores for other high-value tasks. Intel DSA delivers high bandwidth and low latency for data movement, plus improved power efficiency for high-volume use cases like OVS.

Customers can further improve performance and optimize CPU efficiency by offloading OVS to an Intel® Infrastructure Processing Unit (Intel® IPU).

With Intel Xeon Scalable processors, HPC acceleration is built in

The core foundation for HPC acceleration is baked into every Intel Xeon Scalable processor and is available for virtually any software to leverage. HPC customers can gain the benefits of this technology with little to no effort.

The Intel® oneAPI HPC Toolkit is an add-on to the Intel® oneAPI Base Toolkit for building HPC applications using the latest techniques in vectorization, multithreading, multinode parallelization and memory optimization. The toolkit includes cluster analysis and tuning tools based on the open message passing interface (Open MPI) library.

Accelerated performance for the next era of HPC

As HPC becomes more accessible and less expensive, the relative value of supercomputing resources will increase exponentially. Computing power that was once limited to national labs and global manufacturers is becoming available via cloud instances and hybrid HPC clusters. Intel® HPC engines can improve HPC performance across the board so that more organizations can access the computing resources they need to make new discoveries, innovate and get to market faster.

Conquer the most demanding computational tasks with Intel HPC engines that are built into Intel Xeon processors.

Learn more

[Intel high-performance computing ›](#)

[Intel AVX-512 ›](#)

[Intel Deep Learning Boost ›](#)

[AI and HPC convergence ›](#)

[AI and deep learning on Intel Xeon Scalable processors ›](#)

The 'deep' impact on performance with Intel AMX

4th Gen Intel Xeon Scalable processors with Intel AMX vs. 3rd Gen Intel Xeon Scalable processors

Up to

7.3x higher

real-time Natural Language Processing inference performance[3]

Up to

4.1x higher

TensorFlow INT8 batch inference performance[4]

Start accelerating HPC workloads now — in the cloud or on your own infrastructure — with 4th Gen Intel Xeon Scalable processors.

Visit [intel.com/hpc](#)

[1] See [H1] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary

[2] Intel® Xeon® Processors HPC Performance: Manufacturing

[3] See [A19] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

[4] See [A18] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Notices and disclaimers

Performance varies by use, configuration, and other factors. Learn more at 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.

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

Intel® Advanced Vector Extensions (Intel® AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause, a) some parts to operate at less than the rated frequency and, b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration, and you can learn more at intel.com/content/www/us/en/architecture-and-technology/turbo-boost/intel-turbo-boost-technology.html.

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

Your costs and results may vary.

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel® products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

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