Product Brief

Intel® HPC Engines
5th Gen Intel® Xeon® Scalable Processors

Accelerated HPC with 5th Gen Intel® Xeon® Scalable Processors and Intel® HPC Engines

Up to
1.31x higher
LAMMPS performance on 5th Gen Intel Xeon Scalable platform vs. prior gen[1]

Customer success: Intel worked with Taboola to optimize and benchmark their prediction algorithm on Intel® Xeon® processors.

Read the story ›

The Academic Center for Computing and Media at **Kyoto University** uses Intel Xeon processors to accelerate the results of scientific investigations.

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 or other hardware and software solutions. The 5th Generation Intel® Xeon® Scalable processors come equipped with purpose-built accelerators that can elevate HPC workload performance and power efficiency.

High-performance computing with Intel® HPC Engines

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 for obtaining business insights faster and for making critical business decisions — all while lowering costs.

The new and improved features of 5th Gen Intel Xeon Scalable processors with Intel® HPC Engines can increase performance across the fastest-growing workload types, like simulation and modeling. The 5th Gen Intel Xeon Scalable processors deliver improved performance, efficiency and cost savings for targeted workloads — with 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

Machine learning (ML) technologies are making workloads more efficient, effective and insightful. Industry trends are driving customers to use HPC- and AI-powered solutions for improved business results. Intel AMX is designed to boost AI performance, and Intel is sharing its expertise in AI with customers that are utilizing both HPC and AI solutions.

Intel AMX, one of the built-in accelerator engines integrated into 5th Gen Intel Xeon Scalable processors, is Intel's latest 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 multiply operations. Intel AMX also uses a two-dimensional register file to store larger chunks of data. Built to accelerate AI workloads, Intel AMX is critical for delivering performance across workloads where HPC and AI converge.

Intel Storage Engines: Built-in accelerators for storage-specific workloads

Integrating workload accelerator engines into the CPU has three major benefits. First, built-in accelerators resolve the I/O bottlenecks and latency inherent in drop-in accelerator cards and external appliances. Second, they process their specific workloads faster than a CPU alone. Third, they allow the CPU to offload tasks and preserve headroom for the workloads that need higher-performance computing resources.

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.

Fewer steps mean 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. Intel Xeon Scalable processors have up to two fused multiply-add (FMA) units per core to combine multiplication and addition into a single operation and accelerate computation speeds.

Intel QuickAssist Technology

Free up space and reduce costs by offloading compute-intensive workloads with Intel QAT, a built-in accelerator for 5th 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. It also allows each core to serve more clients by improving performance for cryptography and data compression while reducing the data footprint.

Intel Data-Streaming Accelerator

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

Intel DSA, an accelerator integrated into Intel Xeon processors, delivers high performance for storage, networking and 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.

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

With 5th Gen Intel Xeon Scalable processors, HPC acceleration is a native feature

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

The Intel® 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. The Intel® oneAPI Math Kernel Library, meanwhile, offers a highly optimized, fast and complete library of math functions for Intel® CPUs and Intel® GPUs.

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.

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

9.9x higher

real-time Natural Language Processing inference performance (BERT-large) and 7.7x higher performance/watt on 5th Gen Intel Xeon with AMX BF16 vs. 3rd Gen Intel Xeon processors [2]

Up to

2.3x

performance speedup with 5th Gen Intel Xeon Scalable processor vs 3rd Gen Intel Xeon on GPT-J first token latency (int8)[3]

Learn more

Intel AVX-512 ›

AI and HPC convergence ›

AI and deep learning on Intel Xeon Scalable processors ›

Start accelerating HPC workloads now — in the cloud or on your own infrastructure — with 5th Gen Intel Xeon Scalable processors.

Visit intel.com/hpc

---

intel® xeon®

1. See [H14] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
2. See [A19] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
3. See [A1] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.

*Geomean of chat bot, context creation marketing, misc., search, text classification, text generation

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 5th 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.

© 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. 0922/MP/CMD/PDF

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