# Product Brief

**HPC Accelerated**
**Intel® Xeon® Scalable Processors**

# Getting More Out of Every High Performance Computing Core

**Accelerated HPC**

Up to
**2.4x faster**
**on Nanoscale Molecular Dynamics (NAMD)**
with Intel® AVX-512
vs. without[1]

High performance computing (HPC) is essential to scientific discovery, engineering simulations, and the modeling of complex systems. Intel® workload acceleration technologies help get more performance out of Intel® Xeon® Scalable processors today. Our next-generation Intel Xeon Scalable processors have built-in accelerators for key HPC workloads, freeing the CPU cores for other tasks.

## High performance computing is entering a new era

Historically, HPC has had three bottlenecks: compute speed, access, and cost. Over the years, HPC speed has increased exponentially. However, cost and access have remained barriers until very recently when hyperscalers began delivering cloud-based HPC you can consume by the core hour. HPC owners have used this shift to augment their on-premises supercomputers with cloud-based resources, creating hybrid supercomputing models. This new era of hybrid HPC and consumption-based pricing promises to lower the cost of HPC, expand HPC to more diverse use cases, and increase the speed of innovation.

To meet these new demands for HPC, Intel continues to develop the workload-specific acceleration we started with Intel® AVX2 and Intel® AVX-512—instruction sets that help boost performance on current HPC infrastructure. Our next generation of Intel® Xeon® Scalable processors will introduce hardware accelerators that will multiply Intel AVX2 and Intel AVX-512 gains even further.

## 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 AVX2 and Intel AVX-512, 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. A primitive example: you could instruct a CPU to calculate 3x3x3x3x3, 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). Each 3rd Gen Intel® Xeon® Scalable processor has two FMA units per core, which combine multiplication and addition into a single operation and accelerate computation speeds.

### 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® Deep Learning Boost.

Read the story ›

Natural language processing (NLP):
**1.74x**
**better performance**[4]

Image classification:
Up to
**1.93x**
**higher training performance**[5]

For workloads and configurations, visit intel.com/3gen-xeon-config. Results may vary.

### Deep learning performance

3rd Generation Intel Xeon Scalable processors vs. 2nd Gen Intel® Xeon® Scalable processors

### Counting by 512 is a lot faster than counting by one

The "512" in AVX-512 refers to the second way these instructions help the CPU do more with every clock cycle by increasing the number of bits at the CPU's disposal. Forty years ago, a 16-bit PC was impressive. Then, 32-bit machines took over. Today, your smartphone runs at 64 bits. Bit counts refer to the number of registers—the memory slots where the CPU holds data—that the CPU can address per clock cycle.

Intel AVX-512 expands the number of registers for optimized software. Applications can pack 32 double-precision and 64 single-precision floating point operations per clock cycle within the 512-bit vectors, as well as eight 64-bit and 16 32-bit integers, with up to two 512-bit FMA units, thus doubling the width of data registers, doubling the number of registers, and doubling the width of FMA units, compared to Intel® Advanced Vector Extensions 2 (Intel AVX2). It's like counting by 512, 1,024, 1,536 vs. one, two, three.

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

Machine learning and deep learning inference are expanding the capabilities of HPC by accelerating processing speeds, increasing accuracy, and creating entirely new methods for modeling and analysis. Intel DL Boost combines three operations into a single Intel AVX-512 instruction called Vector Neural Network Instructions (VNNI), which reduces the number of operations per clock cycle. Intel DL Boost also allows inferencing to run at int8 precision.

## Looking forward—next generation of integrated accelerators for HPC

4th Gen Intel® Xeon® Scalable processors include integrated, hardware-based accelerators to help you optimize some of the most compute-intensive HPC workloads. These built-in accelerators boost performance in two ways. First, they use custom engines to run specific workloads faster than general-purpose CPU cores. Second, offloading workloads to an accelerator frees the CPU cores to do other tasks.

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

Intel AMX is a built-in accelerator and custom instruction set dedicated to the matrix multiplication at the heart of deep learning workloads. Intel AMX transforms large matrix math calculations into a single operation and uses a two-dimensional register file to store larger chunks of data.

### Intel® Data Streaming Accelerator (Intel® DSA)

Moving data in and out of memory, storage, and networking subsystems places a major burden on the CPU. Intel Data Streaming Accelerator speeds data copying and transformation. It shoulders almost all data movement operations, including checksum, memory compare, and checkpointing. In an example open virtual switch (Open vSwitch) use case, Intel DSA reduced CPU utilization by nearly 40 percent and delivered a 2.5x improvement in data-movement performance.[2] This resulted in nearly doubling the effective performance for the Open vSwitch workload.

### Intel® QuickAssist Technology (Intel® QAT)

Intel QuickAssist Technology is a mature data compression and encryption accelerator developed for Intel® Ethernet controllers and drop-in accelerator cards. With 4th Gen Intel Xeon Scalable processors, Intel QAT will debut as a built-in accelerator for on-the-fly data compression/decompression and cryptographic workloads. Intel QAT frees up to 98 percent of core capacity for other workloads while markedly reducing compressed data footprints.[3]

## With Intel Xeon Scalable processors, HPC acceleration is built in

The core foundation for HPC acceleration—Intel AVX-512 and Intel DL Boost—is baked into every Intel Xeon Scalable processor and available for virtually any software to leverage. We don't expect—or want—data scientists, financial analysts, and engineers recoding their tools and recompiling them for Intel AVX-512. We do it for them. Intel software engineers are constantly optimizing HPC applications and toolchains, and all our distributions, libraries, and tools are free to use.

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® acceleration technologies 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.

## 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 ›**

## Start accelerating HPC workloads now—in the cloud or on your own infrastructure—with 3rd Gen Intel Xeon Scalable processors.

**Visit intel.com/hpc**

intel XEON

1. 2.43x better performance for Nanoscale Molecular Dynamics (NAMD) with software optimizations: See [107] at intel.com/3gen-xeon-config. Results may vary.
2. On an open virtual switch use case with up to four instances of Intel® Data Streaming Accelerator (Intel® DSA), we see a nearly 40 percent reduction in CPU utilization and a 2.5x improvement in data movement performance. This results in nearly doubling the effective core performance for this workload. See edc.intel.com/content/www/us/en/products/performance/benchmarks/architecture-day-2021/. Results may vary.
3. With the Zlib L9 compression algorithm, we see a 50x drop in CPU utilization (i.e., a 98 percent decrease in expected core utilization) while also speeding up the compression by 22 times. Without Intel® QAT, this level of performance would require upward of 1,000 Performance-cores to achieve. See edc.intel.com/content/www/us/en/products/performance/benchmarks/architecture-day-2021/. Results may vary.
4. 1.74x higher INT8 batch inference throughput on BERT-Large SQuAD: See [123] at intel.com/3gen-xeon-config. Results may vary.
5. Up to 1.93x higher AI training performance: See [9] at intel.com/3gen-xeon-config. Results may vary.



0922/MP/CMD/PDF