Technology Brief

High-Performance Computing (HPC)
Intel® Advanced Vector Extensions 512 (Intel AVX-512)

Get Outstanding Computational Performance without a Specialized Accelerator

Built-in acceleration from the Intel Advanced Vector Extensions 512 (Intel AVX-512) instruction set in 3rd Gen Intel® Xeon® Scalable processors can help meet your most demanding computation challenges.

Ultra-wide 512-bit vector-operations capabilities mean that Intel AVX-512 can handle the most demanding high-performance computing (HPC) and data center tasks.

Accelerate existing workloads without having to modify applications.

Available in Intel-based public-cloud instances to support hybrid environments.

The need for greater computing performance continues to grow across industry segments. Organizations need powerful hardware that can meet modern computational needs. And in order to increase performance efficiency, they also want to maximize utilization of infrastructure by running computationally intensive workloads, such as as those common in high-performance computing (HPC), alongside other cloud and data center workloads.

Enter Intel AVX-512

Intel Advanced Vector Extensions 512 (Intel AVX-512) is a set of instructions that can accelerate performance for vector processing–intensive workloads. Vector processing performs an arithmetic operation on a large array of integers or floating-point numbers in parallel. Examples of applications in which vector processing can be highly intensive include scientific simulations and 3D modeling.

With ultra-wide 512-bit vector-operations capabilities, Intel AVX-512 can handle your most demanding computational tasks. 3rd Gen Intel Xeon Scalable processors are specifically built with the flexibility to run computationally intensive workloads on the same hardware as your existing workloads, so that you do not need to invest in additional hardware to run your demanding computational workloads.

A deeper technical dive into Intel AVX-512

Intel AVX-512 is a "single instruction, multiple data" (SIMD) instruction set built on x86 data center processors. In contrast to traditional "single instruction, single data" instructions, a SIMD instruction allows for the execution of multiple data operations with a single instruction, which makes computation more efficient.

The CPU register is the small cache in which processors store instructions and data for computation. Intel AVX-512 has a register width of 512 bits. This large register supports 32 double-precision and 64 single-precision floating-point numbers, in addition to eight 64-bit and sixteen 32-bit integers. Intel AVX-512 provides up to two 512-bit fused-multiply add (FMA) units; doubling the width of the vector processing doubles the number of registers compared to Intel AVX2, which preceded Intel AVX-512. This increased computational capacity enables Intel AVX-512 to process twice the data with a single instruction, which can boost performance for applications ranging from HPC to artificial intelligence (AI) and deep learning (DL) to cryptographic hashing.

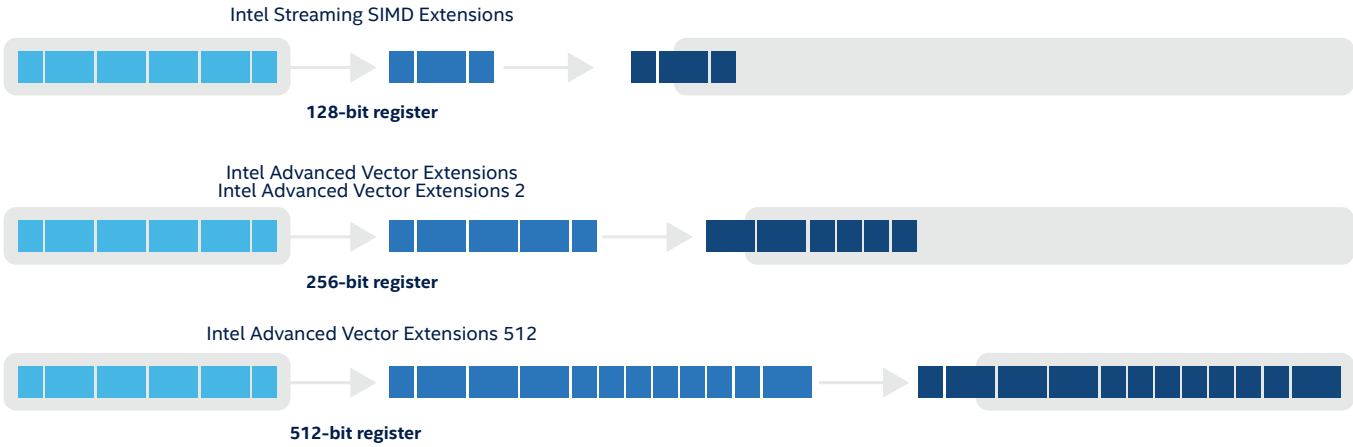

Figure 1. Illustration of the differences in register size and compute efficiency between Intel Streaming SIMD Extensions (Intel SSE), Intel AVX2, and Intel AVX-512

Accelerate data-center performance with Intel AVX-512

3rd Gen Intel Xeon Scalable processors with Intel AVX-512 have a built-in performance advantage for computationally intensive workloads, whereas AMD EPYC processors are currently only available with AVX2, which can lead to significant performance advantages for Intel Xeon Scalable processors in cases such as image-classification, HPC, AI, and web encryption.[1]

A key difference between Intel AVX-512 and AVX2 in AMD processors is the auto-vectorizing compilers that Intel offers. The Intel compilers automatically enable applications to use vector instructions for Intel AVX-512. When coupled with profiling tools that help find high-impact vectorization opportunities to safely speed up applications, Intel compilers are fundamental to making efficient use of the wider vectors in Intel AVX-512.

The performance enhancements from Intel AVX-512—including faster workload speeds and more efficient data processing—are available immediately to your applications. Business applications do not need to be modified to take advantage of the performance improvements made possible by Intel AVX-512. In fact, your applications running on servers powered by 3rd Gen Intel Xeon Scalable processors might already be benefiting from the performance gains provided by Intel AVX-512. And for applications running in the cloud, Intel AVX-512 is also available in Intel-based public-cloud instances to support hybrid environments.

Some application use cases that benefit the most from Intel AVX-512 include:

- HPC: scientific simulations, DNA sequencing, 3D advanced modeling, and financial analytics
- Cryptography and data compression, working with Intel Crypto Acceleration and Intel QuickAssist Technology (Intel QAT)
- Image and audio/video processing
- AI and DL, working with Intel Deep Learning Boost (Intel DL Boost)

As examples of the performance boost that Intel AVX-512 can supply for HPC and other compute-intensive workloads such as these, compared to an AMD EPYC 7763 processor, Intel AVX-512 on an Intel Xeon Platinum 8380 processor can provide 1.50x higher Monte Carlo FSI performance for finance,[2] 1.32x higher RELION performance for life sciences,[3] and 1.27x higher NAMD performance for scientific simulation.[4]

Commercial and open source software applications that utilize Intel AVX-512 include:

- **Intel oneAPI Deep Neural Network Library (oneDNN)**, which works in conjunction with Intel AVX-512 to reduce processor calls to memory and enhance the performance of DL frameworks.[5]
- **SHA-256 SIMD**, which accelerates SHA-256 cryptographic computations and experiences when coupled with Intel AVX-512.[6]
- **MD5 SIMD**, which accelerates MD5-hashing performance with Intel AVX-512.[7]

Customer use cases

Organizations across all industry segments understand the value of data-driven insights in helping them reach their operational and enterprise goals. Here are two examples of organizations using Intel AVX-512 and Intel Xeon Scalable processors to accelerate HPC and cluster-computing workloads.

The U.S. National Oceanic and Atmospheric Administration (NOAA) requires ever-increasing HPC capacity to advance its numerical weather-prediction models. NOAA is developing and prototyping its next-generation Rapid Refresh Forecast System (RRFS) on Intel Xeon Scalable processor–based cloud instances at Amazon Web Services (AWS) that make use of Intel AVX-512 to speed prediction-model calculations.

The University at Buffalo's (UB) Center for Computational Research (CCR) offers unique opportunities to Western New York's many businesses. CCR provides dedicated converged HPC and AI capabilities with a new HPC + AI compute cluster. The cluster, powered by Intel Xeon Gold 6330 processors, enables a large community of customers to develop innovative solutions through simulation, modeling, and machine learning (ML), all on the same generalized hardware.

Boost computational performance without a specialized accelerator

Intel AVX-512 provides increased computing performance across a variety of use cases, without the need for additional, specialized hardware accelerators. Moreover, the benefits of Intel AVX-512 are generally available without the need to modify applications to use them. Finally, Intel AVX-512 is widely available through all major public cloud service providers (CSPs).

To learn more, visit the Intel AVX-512 home page: intel.com/content/www/us/en/architecture-and-technology/avx-512-overview.html.

intel XEON

[1] Intel. "Why Intel is the Right Partner for Business vs. AMD." intel.com/content/www/us/en/products/performance/amd-cloud-facts.html.

[2] As measured by Monte Carlo FSI Kernel testing as of March 2021 of a 3rd Gen Intel Xeon Platinum 8380 processor versus an AMD EPYC 7763 processor. For full workloads and configuration details, visit www.intel.com/PerformanceIndex (3rd Generation Intel Xeon Scalable processors, claim 37). Results may vary.

[3] As measured by RELION Plasmodium Ribosome testing as of March 2021 of a 3rd Gen Intel Xeon Platinum 8380 processor versus an AMD EPYC 7763 processor. For full workloads and configuration details, visit www.intel.com/PerformanceIndex (3rd Generation Intel Xeon Scalable processors, claim 38). Results may vary.

[4] As measured by NAMD STMV testing as of March 2021 of a 3rd Gen Intel Xeon Platinum 8380 processor versus an AMD EPYC 7763 processor. For full workloads and configuration details, visit www.intel.com/PerformanceIndex (3rd Generation Intel Xeon Scalable Processors, claim 36). Results may vary.

[5] Intel. "Accelerate AI with oneDNN." https://www.intel.com/content/www/us/en/developer/articles/technical/accelerate-ai-with-onednn.html.

[6] GitPlanet. "minio/Sha256 Simd." https://gitplanet.com/project/sha256-simd.

[7] GitPlanet. "minio/Md5 Simd." https://gitplanet.com/project/md5-simd.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for additional details.

No product or component can be absolutely secure.

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

Your costs and results may vary.

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

Printed in USA 0722/DR/PRW/PDF Please Recycle 351851-001US