Product Brief

AI Accelerated
Intel® Xeon® Scalable Processors

Built-in Accelerators in Intel® Xeon® Scalable Processors Boost Performance for the Entire AI Pipeline

70%
of data center AI inferencing **runs on Intel® Xeon® processors**[1]

9 out of 10
enterprise applications **will include AI** by 2025[2]

AI spans a wide range of workloads and use cases from data analysis and classical machine learning to language processing and image recognition. Intel® Xeon® Scalable processors combine flexible computing performance for the entire AI pipeline plus built-in accelerators for specific AI workloads in data science, model training, and deep learning inference.

AI is bigger than deep learning, and it's only getting bigger

AI is in its early stages and growing rapidly on every front. Classic machine learning algorithms and deep learning models are becoming basic building blocks of how business gets done, from core enterprise applications to automated voice attendants. Putting AI to work at scale depends on a lengthy development pipeline that flows from data science to training, validation, and finally deployment. Each step has its own development toolchains, frameworks, and workloads—all of which create unique bottlenecks and place distinct demands on computing resources. Intel Xeon Scalable processors feature built-in acceleration that can help break through these barriers and increase AI performance across the board.

AI is really just math. Lots and lots of math.

Massive volumes of mathematical calculations are at the core of every AI task and operation. Many data science operations—like modeling data and machine learning algorithms—run on statistics, algebra, and complex vector mathematics. Deep learning AI requires vast amounts of matrix multiplication. All of these AI applications are brute force operations involving large datasets and extensive processing resources including CPUs, GPUs, FPGAs, and workload-specific, custom-manufactured ASICs.

Intel® Advanced Vector Extensions 512 (Intel® AVX-512)—the math cheat that speeds up AI

Intel Xeon cores can hash SSL encryption for websites, crunch massive databases, and run simulations for pharmaceutical research, chip design, or Formula 1 engines. They are all-around workhorses, but they are not as fast, natively, at deep learning training—a subset of the overall AI pipeline—as dedicated accelerators. This is because CPUs process operations sequentially, one calculation at a time. Other processor types can process operations in parallel, which means multiple calculations at the same time.

Intel® AVX-512 overcomes the architectural limitations of a CPU by packing more operations into each clock cycle. This allows the CPU to cheat and work more like a parallel processor.

Customer success—Real-world acceleration on Intel® Xeon® Scalable processors

Tencent Cloud delivers real-time speech synthesis with 3rd Gen Intel® Xeon® Scalable processors.

Get the details ›

BeeKeeperAI develops clinical AI algorithms while helping preserve data privacy.

Read the story ›

Complicated CPU instructions, simple strategy: Work smarter, do more in every cycle

The extensions in Intel AVX-512 are instruction sets that tell the CPU what to do and how to do it. How they work is very complex, but the basic logic of AVX-512 is pretty simple. First, condense multiple steps into fewer operations whenever possible. Second, help the CPU do more operations with every clock cycle.

Fewer steps means faster processing

Math can be very smart—and very elegant. Intel AVX-512 uses a lot of smart, beautiful math to 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. AVX-512 takes that logic and applies it to hundreds of workload-specific operations, including some of the toughest operations in AI.

Counting by eight is a lot faster than counting by one

The "512" in AVX-512 refers to the second way that these instructions increase the number of bits at the CPU's disposal with every clock cycle. Forty years ago, a 16-bit PC was pretty impressive. Soon, 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. AVX-512 expands the number of registers to—can you guess?—512 bits. When an application takes advantage of Intel AVX-512, it runs up to 8x faster than the CPU's base 64-bit speed simply by expanding the number of registers. It's like counting to 96 by one, two, three … vs. eight, 16, 24.

Intel® Deep Learning Boost (Intel® DL Boost)—smarter math for neural networks

Deep learning AI uses massive amounts of matrix multiplication to train neural network models and apply those models to real-world tasks using a method called inference. During inference, a computer compares incoming data (for example, an audio signal containing speech) to a model (in this case a speech recognition model) and infers what the data means. Inference is used in object recognition, image segmentation, text recognition, and practically every other deep learning AI task.

Training deep learning models can take hours or days of computing power. Deep learning inference can take fractions of a second to minutes, depending on the model's complexity and how accurate the results need to be. When you scale training or inferencing up to data center–level computing, the time, energy, and performance budgets become immense.

Intel DL Boost uses several Intel AVX-512 instructions to accelerate deep learning workloads. It combines three operations into one Vector Neural Network Instructions (VNNI) set, reducing the number of operations per clock cycle. Intel DL Boost also accelerates deep learning workloads using INT8 precision.

Upcoming advancements will boost AI performance even further

4th Gen Intel® Xeon® Scalable processors will include a built-in accelerator dedicated to the matrix multiplication at the heart of deep learning workloads. Intel® Advanced Matrix Extensions (Intel® AMX) combines a new instruction set that turns large matrices into a single operation with two-dimensional register files that store larger chunks of data for each core.

Faster AI is practically automatic with Intel Xeon processors

AI acceleration on Intel Xeon Scalable processors is built into the CPU's instruction set architecture (ISA). This means they are ready and available for any piece of software that can take advantage of them. We don't expect data scientists and AI developers to recode their tools and recompile them for Intel AVX-512—we do it for them.

Intel software engineers are constantly optimizing open source AI toolchains and passing those optimizations back to the community. For example, TensorFlow 2.9 ships with Intel® oneAPI Deep Neural Network Library (Intel® oneDNN) optimizations by default. Download the latest edition, and TensorFlow will automatically take advantage of Intel optimizations.

For other applications in the AI pipeline, data scientists and developers can download free open-source Intel distributions, libraries, and development environments that take advantage of every built-in accelerator in our ISA for 3rd Gen Intel Xeon Scalable processors.

Basically, faster AI on Intel hardware can be as easy as downloading the Intel version of the tools you already use and getting to work.

Learn more

AI and deep learning on Intel Xeon Scalable processors ›

Intel AVX-512 ›

Intel Deep Learning Boost ›

Intel® AI Analytics Toolkit ›

Software optimization gains for applications in the AI pipeline

~38x–200x
faster scikit-learn with the Intel® Extension for scikit-learn[3]

~90x
faster pandas with the Intel® Distribution of Modin[3]

Up to
3x faster
TensorFlow using Intel® oneDNN[3]

AI acceleration on 3rd Gen Intel® Xeon® Scalable processors

Speed boosts for deep learning AI workloads

Up to
1.74x
higher INT8 batch inference throughput on BERT-Large SQuAD with Intel® DL Boost on 3rd Gen Intel® Xeon® Scalable processors vs. prior generation[4]

Up to
1.59x
higher INT8 real-time inference throughput with Intel® DL Boost on 3rd Gen Intel Xeon Scalable processors vs. prior generation[5]

Up to
4.5x
more images per second at INT8[6] and up to 6x more images per second at BF16[7] object detection (SSD-ResNet-34) using Intel® AMX on upcoming 4th Gen Intel® Xeon® Scalable processors

Start accelerating AI workloads now—in the cloud or on your own infrastructure—with Intel optimizations for AI and machine learning.

Learn more ›

intel XEON®

1. Based on Intel market modeling of the worldwide installed base of data center servers running AI inference workloads as of December 2021.
2. "IDC FutureScape: Worldwide IT Industry 2020 Predictions," October 2019. Doc #US45599219. idc.com/getdoc.jsp?containerId=US45599219.
3. "One-Line Code Changes to Boost pandas, scikit-learn, and TensorFlow Performance," July 2021. intel.com/content/www/us/en/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html
4. See [123] at intel.com/3gen-xeon-config. Results may vary.
5. See [122] at intel.com/3gen-xeon-config. Results may vary.
6. See [41] at edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/. Results may vary.
7. See [42] at edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/. Results may vary.

0922/MP/CMD/PDF