Product Brief

AI Accelerated
Intel® Xeon® Scalable processors

Intel® AI Engines for Intel® Xeon® CPUs boost performance of the entire AI pipeline

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

Up to
10x higher
PyTorch performance for both real-time inference and training workloads with built-in Intel AMX BF16 vs. prior generation with FP32[2]

Real-Time Inference Performance
Up to
6.2x higher
real-time Natural Language Processing inference performance (BERT) on 4th Gen Intel Xeon Platinum 8480+ with Intel AMX BF16 vs. prior generation with FP32[3]

AI spans a wide range of workloads and use cases, from data pre-processing and classical machine learning to deep learning models such as language processing and image recognition. Intel® Xeon® Scalable processors with Intel® AI Engines combine powerful compute performance for the entire AI pipeline plus built-in accelerators for specific AI workloads in machine learning, data analysis and deep learning.

Built-in power for AI across the enterprise

AI is pervasive and stretches across diverse and critical workloads. Classic machine learning (ML) 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 pre-processing to training to 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 accelerators that can be used to run the entire pipeline right out of the box and increase AI performance across the board. Intel® Accelerator Engines are purpose-built integrated accelerators that support the most demanding emerging workloads.

Accelerate deep learning with Intel® Advanced Matrix Extensions (Intel® AMX)

Intel AMX is Intel's next-generation advancement for deep-learning training on 4th Gen Intel® Xeon® Scalable processors. Ideal for workloads like natural-language processing, recommendation systems and image recognition, Intel AMX extends the built-in AI acceleration capabilities of previous Intel Xeon Scalable processors while also offering significant performance gains.[4]

Intel AMX provides workload boost for AI models and can help customers improve total cost of ownership (TCO) by aggregating specific AI workloads onto the CPU instead of offloading them to a discrete accelerator.

Intel AMX also improves tiled multiply performance with a higher max throughout (Ops/Cycle) compared to Intel® Advanced Vector Extensions 512 (Intel® AVX-512) on CPU cores.[5]

Improving natural-language processing and recommender systems

4th Gen Intel Xeon Scalable processors and Intel AMX offer a big performance boost for natural-language processing — and without additional hardware. Libraries are already integrated into TensorFlow and PyTorch, giving developers the benefits of built-in AI acceleration without the extra work. Developers can also more easily migrate code from different hardware environments — a process that can be both lengthy and costly.

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

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

CERN, home to the world's largest particle accelerator, uses built-in Intel® DL Boost to speed up inferencing — and not at the expense of accuracy.

By accelerating deep-learning inferencing and training, the 4th Gen Intel Xeon Scalable processor featuring Intel AMX helps you customize the user experience while balancing TCO. It does this with a deep-learning-based recommender system that factors in real-time user behavior signals and additional context features like time and location.

Future innovation is here with 4th Gen Intel Xeon Scalable processors and accelerator engines

Whether you're using Intel Xeon processors for your workloads on prem, in the cloud or at the edge, Intel Accelerator Engines can help your business reach new heights. They provide a range of benefits including faster security processing, stronger data protection and better infrastructure utilization.

Intel Accelerator Engines can also help increase virtual and physical CPU utilization and minimize per-core solution licensing.

Above all, these built-in accelerators provide increased application performance, reduced costs and improved platform-level efficiency.

Intel® Advanced Vector Extensions 512 (Intel® AVX-512) for faster machine learning

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 can complete deep-learning training workloads even faster with the help of the AVX-512 accelerator.

Improved over multiple generations, Intel AVX-512 allows Intel Xeon Scalable processors to pack more operations into each clock cycle and offer performance that compares with parallel processing. 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. Here's 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. 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 count refers 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 eight times faster than the CPU's base 64-bit speed simply by expanding the number of registers. It's like counting to 96 by 1, 2, 3 ... versus 8, 16, 24.

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

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 by using both INT8 and BF16. It combines three operations into one Vector Neural Network Instructions (VNNI) set, reducing the number of operations per clock cycle while simultaneously providing the full computing potential of Intel Xeon Scalable processors. VNNI accelerates DL inference by using INT8 precision.

The introduction of 4th Gen Intel Xeon scalable processors also promises an even bigger lift in performance. With Intel AMX's working alongside AVX-512 on 4th Gen Intel Xeon Scalable processors, tile multiply performance has even greater max throughout (Ops/Cycle) compared to 3rd Gen Intel Xeon Scalable processors.[6]

Engines that require less power to run a more powerful AI

Because Intel Xeon Scalable processors featuring Intel AI Engines require fewer hardware resources, they offer a more powerful and energy-efficient solution for running AI workloads.

Intel Xeon Scalable processors with built-in accelerator engines can help provide improved workload results, like lowering TCO and delivering better return on investment (ROI) for today's demanding AI workloads.

For instance, on average, systems with Intel Xeon Scalable processors cost 17% less than comparable systems requiring GPU integration.[7]

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 it's ready and available for any piece of software that can take advantage of it. 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 Intel Xeon Scalable processors.

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.

Organizations today need to get more workload performance out of their infrastructure — and do so with more power efficiency and at lower costs. The purpose-built Intel AI Accelerator Engines of Intel Xeon Scalable processors will help you get the maximum out of the AI workloads that matter most to your business.

Learn more about what Intel Xeon Scalable processors with built-in Intel Accelerator Engines can accomplish for the AI workloads that matter most to your business.

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

[1]Based on Intel market modeling of the worldwide installed base of data center servers running AI inference workloads as of December 2021.

[2]See [A16] and [A17] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

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

[4]3.5x to 10x higher PyTorch Training performance on 4th Gen Intel Xeon Scalable processor with built in Intel AMX (BF16) vs. prior generation (FP32) See [A16] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

[5]https://edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/, Session Benchmark #41 and #42. Results may vary

[6]https://edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/, Session Benchmark #41 and #42. Results may vary

[7]See [100] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/

2038008-0223