intel.
XEON

How to get the most out of Intel® Xeon® Scalable processors with built-in accelerators

E-Guide

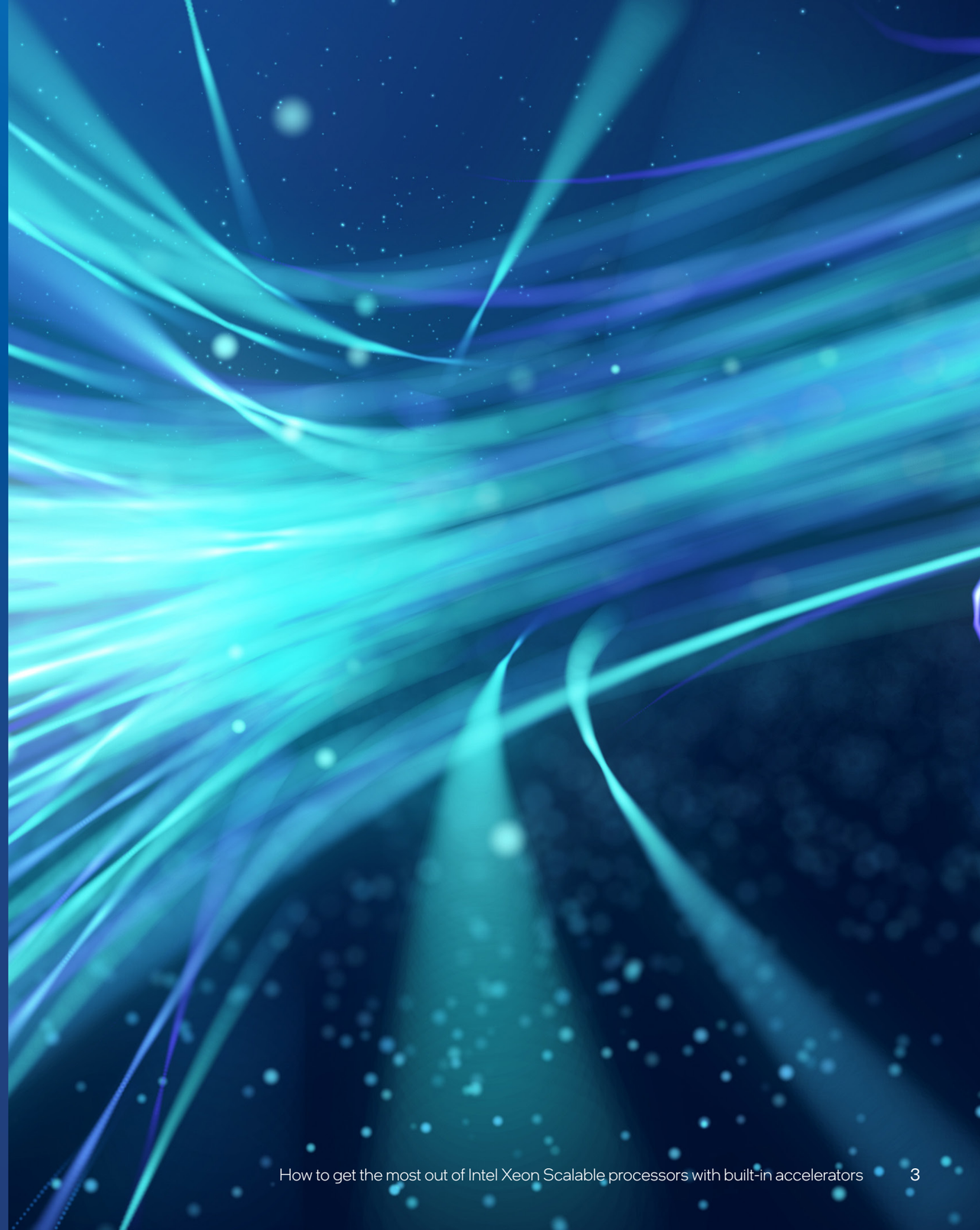
Table of contents

What is built-in acceleration and why should you use it?

What if, instead of buying new equipment every time you need to establish new capabilities, you could rely on the technology already built into your CPU? With Intel® Xeon® Scalable processors, you can. These CPUs contain features known as built-in accelerators to deliver enhanced benefits to your workloads.

Every Intel Xeon Scalable processor features the broadest and most unique set of integrated accelerators to help boost performance and efficiency, reducing the need for additional specialized hardware. In both cloud and on-prem environments, these purpose-built features support today's most common and demanding workloads spanning AI, security, HPC, analytics, storage and networking.

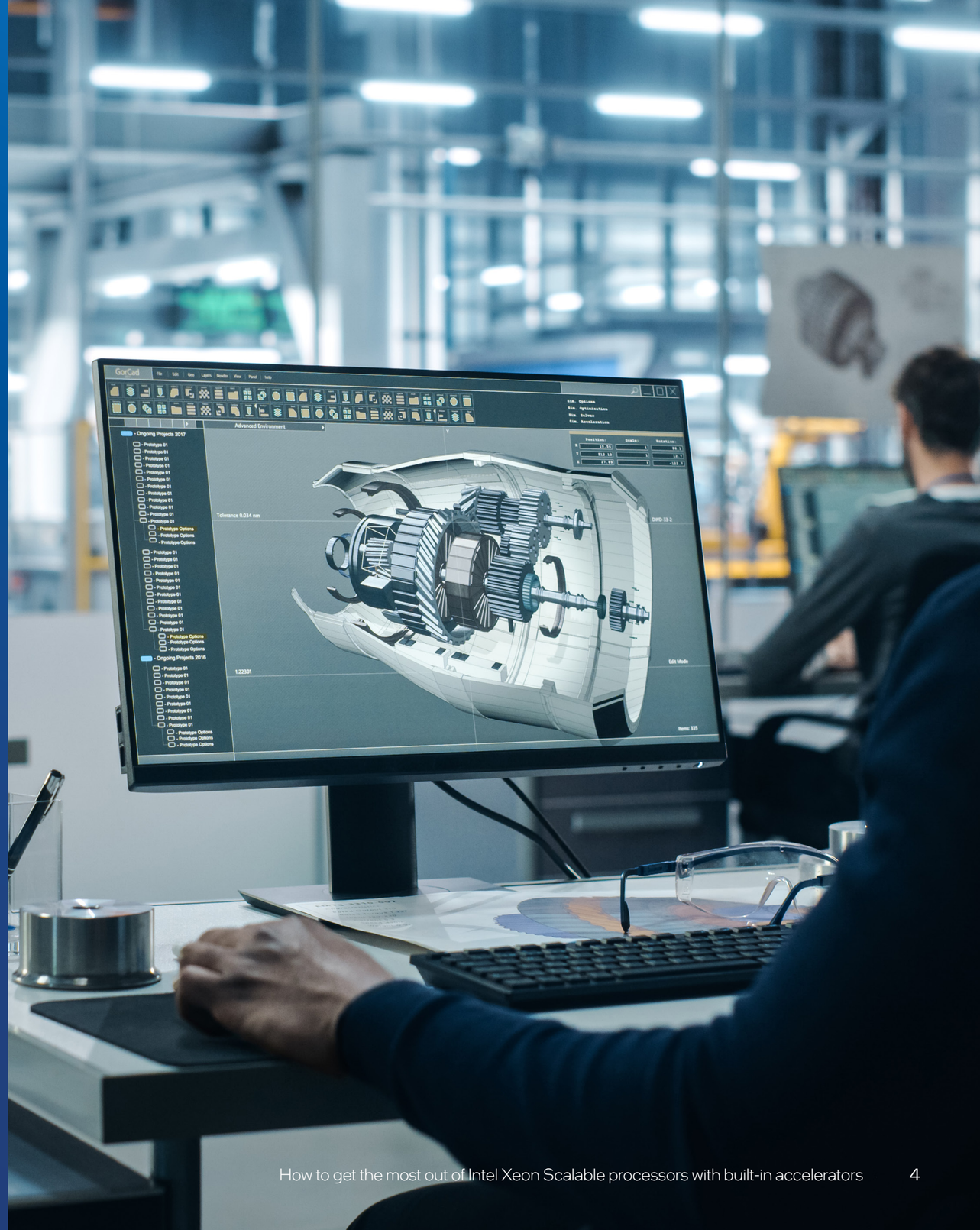
In this guide, we'll focus on AI, HPC and security.



The real-world benefits of Intel built-in accelerators

Whether you're using Intel Xeon Scalable processors for your workloads on prem, in the cloud or at the edge, our integrated accelerators can help your business reach new heights. They provide a range of benefits including faster security processing, stronger data protection and better infrastructure utilization. Above all, these built-in accelerators provide **increased application performance, reduced costs and improved energy efficiency:**

Performance	As a purpose-built component, Intel built-in accelerators can often deliver higher performance for a targeted workload.[1]
Cost savings	Intel built-in accelerators allow you to improve performance without having to purchase additional specialized hardware.
Energy savings	Intel built-in accelerators help users improve energy efficiency by sparing users from having to include additional cores to the server rack.



Which Intel built-in accelerators are right for your business?

While Intel Xeon Scalable processors have a full array of accelerators built in, certain accelerators are better suited for specific tasks/workloads. To help you decide which Intel technologies can best support your business, let's take a deep dive into three of our key offerings in the AI, HPC and security segments.

AI

Intel® Deep Learning Boost (Intel® DL Boost) brings significantly accelerated performance to inferencing and training across common AI and HPC workloads.[2]

HPC

Intel® Advanced Vector Extensions 512 (Intel® AVX-512) is purpose-built to accelerate performance for the most demanding computational workloads in science, business and beyond.

Security

Intel® Software Guard Extensions (Intel® SGX) helps protect data in use via unique application-isolation technology.

AI — Intel Deep Learning Boost

What is Intel Deep Learning Boost?

Intel DL Boost is an accelerator designed to enhance performance and provide greater efficiency for AI and deep learning-related tasks and workloads.

Intel Xeon Scalable processors introduced purpose-built AI acceleration in 2019 with Intel Vector Neural Network Instructions (VNNI), now Intel DL Boost.

Based on the Intel AVX-512 accelerator, the VNNI component of Intel DL Boost can combine three instruction sets into one, drastically reducing the amount of time it takes to complete a task.

What are the most common use cases for Intel DL Boost?

Intel DL Boost accelerates various AI inferencing tasks such as image classification, language translation and object detection.

How are companies using Intel DL Boost in the real world?

Software company **rinf.tech** used Intel DL Boost to provide quicker and more accurate image analysis to support better real-time decision-making in retail, automotive, video surveillance and business intelligence use cases. Inference performance was up to 7.4 times faster than the baseline.[3]

Huiyi Huiying Medical Technology (HYHY) used Intel DL Boost to help optimize the performance of full-cycle AI medical imaging solutions. Thanks to the advantages brought by synergistic software-hardware acceleration, the company saw significant improvements in inference speed for image analysis scenarios like COVID-19 screening and breast cancer detection.[4]

Konfoong Biotech International Co., Ltd. (KFBIO), which specializes in digital pathology system development and production, used Intel DL Boost to complete scanning and diagnosis of M. tuberculosis specimens up to 11.4 times faster than the baseline.[5]

What performance advantages does Intel DL Boost offer?

Customers who use Intel® optimization for TensorFlow and Intel DL Boost will **gain over**

11x more AI inference performance

on 3rd Gen Intel Xeon Scalable processors compared to 2nd Gen Intel Xeon Scalable processors.[6]

Intel DL Boost can also deliver more performance per watt for AI workloads, allowing organizations to reduce costs and energy consumption.

HPC — Intel Advanced Vector Extensions 512

What is Intel Advanced Vector Extensions 512?

Intel AVX-512 is a general-purpose performance-enhancing accelerator with a wide range of uses. With ultrawide 512-bit vector operations capabilities, Intel AVX-512 is especially suited to handle the most demanding computational tasks commonly encountered in the HPC segment.

What are the most common use cases for Intel AVX-512?

Intel AVX-512 is used by organizations across educational, municipal, financial, enterprise, engineering and medical industries for a wide range of complex tasks. These include AI, scientific simulations, 3D modeling and analysis, financial analytics, audio and video processing, cryptography, and data compression.

Intel AVX-512 enables real-time analytics for financial service workloads to improve customer experience, compliance and data security.

Intel AVX-512 enables you to run complex workloads on existing hardware, accelerating performance for tasks like 3D modeling and simulation.

How are companies using Intel AVX-512 in the real world?

The University at Buffalo's Center for Computational Research uses Intel AVX-512 to offer rich computing resources to businesses in Western New York like Marion Surgical, a company that uses virtual reality (VR) and augmented reality (AR) to teach surgeons complex procedures.[7]

The Broad Institute of MIT and Harvard uses Intel AVX-512 to improve processing speed and reduce costs for genomics workloads on Google Cloud N1 and N2 instances.[8]

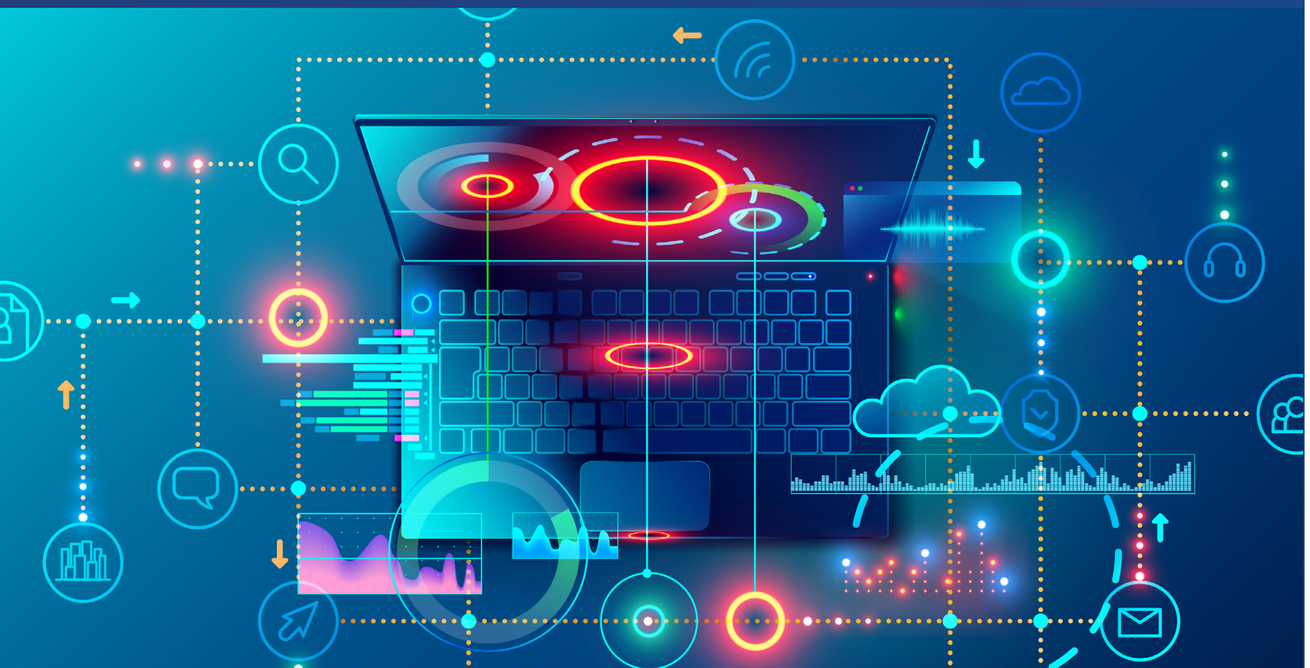
Researchers studying fundamental particles at **CERN**, the European Organization for Nuclear Research, used Intel AVX-512 to accelerate simulation workloads via quantization. It offered **1.8 times higher performance** and slightly improved accuracy.[9]

What performance advantages does Intel AVX-512 offer?

Because HPC workloads involve massive amounts of data, CPU performance is crucial to ensuring accurate results in a timely manner.

Intel AVX-512 helps improve the vector processing power of Xeon CPUs compared to previous-gen solutions, allowing organizations to tackle intense workloads with greater speed. 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 fused-multiply add (FMA) units, thus doubling the width of data registers, number of registers and width of FMA units compared to Intel® Advanced Vector Extensions 2 (Intel® AVX2).

Compared to AMD solutions, 3rd Gen Intel Xeon Scalable processors working with Intel AVX-512 offered superior performance across 12 HPC-related benchmarks and real-world applications.[10]



When running the biomolecular simulation algorithm NAMD, 3rd Gen Intel Xeon Scalable processors offered

1.27x higher performance

compared to **AMD Milan.**[11]

When running the RELION algorithm used for 3D imaging in structural biology, 3rd Gen Intel Xeon Scalable processors offered

1.32x higher performance

compared to **AMD Milan.**[12]

When running the financial investment simulation algorithm Monte Carlo FSI, 3rd Gen Intel Xeon Scalable processors offered

1.50x higher performance

compared to **AMD Milan.**[13]

When running the modeling and simulation algorithm LINPACK, 3rd Gen Intel Xeon Scalable processors offered

1.18x higher performance

compared to **AMD Milan.**[14]

Security — Intel Software Guard Extensions

What is Intel Software Guard Extensions?

Intel SGX offers a hardware-based security solution that helps protect data in use via application-isolation technology. By protecting selected code and data from inspection or modification, developers can run sensitive data operations inside enclaves to help increase application security or protect data confidentiality. This adds another layer of defense by helping to reduce the attack surface of the system.

What are the most common use cases for Intel SGX?

Intel SGX enables confidential computing solutions to better protect data on prem, at the edge and in the cloud. This accelerator supports organizations looking to protect sensitive data and code to help ensure compliance with regulations related to data privacy, sovereignty and confidentiality.

How are companies using Intel SGX in the real world?

British financial institution **Nationwide Building Society** used Intel SGX to build "Know Your Customer," a system **that allows multiple confidential data sets to be more safely processed inside of an enclave,** in compliance with data regulations.[15]

Swiss Re Group, one of the world's largest reinsurance providers, is successfully exploring additional **protection of confidential data** from multiple parties with a confidential computing proof of concept using Intel SGX.[16]

The **University of California San Francisco** used Intel SGX to develop privacy-preserving analytics that accelerate the development and validation of clinical algorithms. The platform will provide a "zero trust" environment to help **protect both the intellectual property of an algorithm and the privacy of health care data.**[17]

What security advantages does Intel SGX offer?

Intel SGX provides a crucial building block for confidential computing by restricting access to sensitive data and code while in use, helping keep it safe from inspection or corruption by other software. Only Intel SGX has the flexibility to support virtualized, bare-metal and cloud-native container deployments.

The next generation of Intel built-in accelerators

Intel built-in accelerators are currently featured within Intel Xeon Scalable processors — and the next generation of built-in accelerators is on the horizon.

The 4th Generation Intel Xeon Scalable processor will include Intel® Advanced Matrix Extensions (Intel® AMX), Intel® QuickAssist Technology (Intel® QAT) and Intel® Data Streaming Accelerator (Intel® DSA), among other built-in features. Here's a quick glimpse at what you can expect.

Intel AMX

Our next-generation DL Boost to advance deep-learning performance, Intel AMX features a set of matrix multiplication instructions that will significantly advance AI inference and training, with up to 4.5 times INT8 image inference per second compared to the prior generation.[18]

Intel QAT

Now a part of Intel Xeon Scalable processors, it will offer users even faster data encryption and more efficient data compression for applications from networking to enterprise, cloud to storage and content delivery to database.

Intel DSA

Intel DSA is a high-performance accelerator designed to optimize streaming data movement and transformation operations common in networking, data processing-intensive applications and high-performance storage.

Conclusion

Intel's long history of innovation and integration has uniquely positioned us to develop this new category of integrated accelerators with Intel Xeon Scalable processors. There are various workloads for which purpose-built integrated circuitry will deliver greater business value to customers.

Whether you're looking to increase performance, support sustainability initiatives or help ensure the protection of your most sensitive data, Intel's family of built-in accelerators offers a wide range of solutions without the need for additional hardware. As our own internal testing along with real-world use cases have shown, these accelerators deliver unmatched value compared to other CPU options on the market.

Learn more about Intel Xeon Scalable processors by visiting https://www.intel.com/xeonscalable.

[1]See [123] at https://www.intel.com/3gen-xeon-config
[2]See [121] at https://www.intel.com/3gen-xeon-config
[3]See page 12 at https://www.intel.com/content/dam/www/public/us/en/documents/product-overviews/dl-boost-product-overview.pdf
[4]See https://www.intel.com/content/www/us/en/customer-spotlight/stories/hyhy-customer-story.html
[5]See https://www.intel.com/content/www/us/en/customer-spotlight/stories/kfbio-ai-customer-story.html
[6]See [118] at https://www.intel.com/3gen-xeon-config
[7]See https://www.intel.com/content/www/us/en/customer-spotlight/stories/university-at-buffalo-customer-story.html
[8]See https://www.intel.com/content/www/us/en/newsroom/news/broad-institute-intel-google-advance-biomedical-research.html
[9]See https://www.intel.com/content/www/us/en/customer-spotlight/stories/cern-inference-customer-story.html
[10]See [104] at https://www.intel.com/3gen-xeon-config
[11]See [36] at https://www.intel.com/3gen-xeon-config
[12]See [38] at https://www.intel.com/3gen-xeon-config
[13]See [37] at https://www.intel.com/3gen-xeon-config
[14]See [39] at https://www.intel.com/3gen-xeon-config
[15]See https://www.intel.co.uk/content/www/uk/en/customer-spotlight/stories/nationwide-building-society-customer-story.html
[16]See https://www.intel.com/content/www/us/en/customer-spotlight/stories/swiss-re-customer-story.html
[17]See https://www.intel.com/content/www/us/en/newsroom/news/ucsf-propel-medical-device-innovations.html
[18]See Session Benchmark #41 and #42 at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/. Results may vary.