Solution Brief

AI Inference
Intel® Xeon® Scalable Processors

Optimize Inference with Intel® CPU Technology

Use open standards–based tools and the built-in artificial intelligence (AI) accelerators of Intel® Xeon® Scalable processors to unify and enhance your AI pipeline. Enjoy improved inferencing performance and lower overall total cost of ownership (TCO) across an integrated AI platform.

Up to **70% of CPUs installed** for inferencing are Intel Xeon processors.[1]

Intel Xeon Scalable processors are **the only x86 data center CPUs with built-in AI accelerators.**[2]

Intel Xeon Scalable processors provide up to **30% higher AI performance** across 20 workloads than discrete accelerator hardware.[3]

AI deployment: facts versus fiction

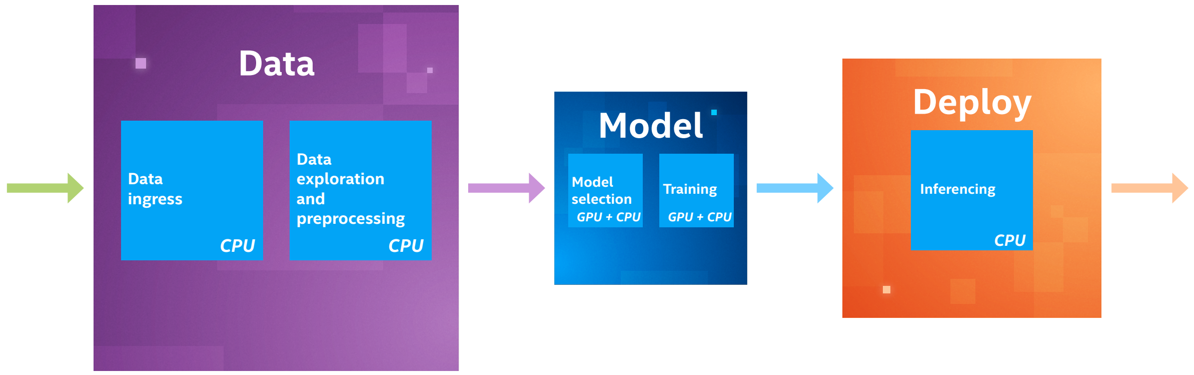
There is a prevailing belief that you need graphics processing units (GPUs) to run all high-performing AI workloads in the data center. If you subscribe to that belief, it might surprise you to learn that today's AI development and deployment pipelines run mostly on Intel Xeon Scalable processors. In 2021, Intel Xeon Scalable processors represented 70 percent of the installed processing units in data centers running AI inferencing workloads.[1] Today's data-centric organizations have a wide variety of business and operational reasons for deploying AI. Yet, as the numbers reveal, their preferred AI solutions converge around using Intel Xeon Scalable processors to boost inferencing performance.

Many organizations use cloud-based AI deployments to remain agile in extremely volatile markets. Utility companies use cloud-based AI to drive prediction models that can more accurately forecast shifting weather conditions, which helps improve the efficiency of their wind-powered generators. These cloud instances run on Intel Xeon Scalable processors and an open software ecosystem that unifies and streamlines their AI pipelines. A unified AI pipeline, one that integrates open software, frameworks, and libraries, facilitates rapid development and deployment of multiple machine learning (ML) model prototypes. This integrated, open ecosystem can help significantly shorten times to solution for complex prediction models. A unified CPU-based platform can also cost less to scale out and is simpler to administer than a platform built on proprietary hardware and software.

Healthcare organizations that use AI for medical-image analysis can count on Intel Xeon Scalable processors to deliver fast processing of large data volumes, near-real-time and highly accurate reporting, and strong data protection for confidential patient information. Intel Xeon Scalable processors, optimized with Intel Deep Learning Boost (Intel DL Boost), provide high-performance inferencing for the deep learning (DL) models used to analyze medical imaging. Intel Xeon Scalable processors can help secure AI workloads with hardened data and system security, which can also help healthcare organizations stay compliant with Health Insurance Portability and Accountability Act (HIPAA) regulations.

Online commerce providers communicate with international customers using AI-driven natural language processing (NLP) and neural machine translation (NMT) interfaces. End users expect an accurate and natural conversational experience, which requires response latencies measured in microseconds.[4] Intel Xeon Scalable processors use 8-bit integer low-precision arithmetic (INT8) to deliver lower latency over 32-bit floating-point precision arithmetic (FP32) workloads, while maintaining high levels of accuracy. Intel Advanced Vector Extensions 512 (Intel AVX-512) with Vector Neural Network Instructions (VNNI) can speed up INT8 inferencing performance even further.[5]

The AI pipeline



The three outer boxes represent AI pipeline stages.
The five inner boxes represent AI workloads.
Box sizes indicate relative levels of processor activity within the AI pipline.

Figure 1. Data-stage workloads are handled almost exclusively by CPUs and produce roughly two-thirds of all processor activity across the AI pipeline

The long-held industry perception has been that you need a GPU to deploy AI in your data center. This perception is built on GPUs being highly performant for DL training. However, DL training occupies a small percentage of model-stage processing. In fact, ML model selection and training, which are CPU-processed workloads, take up the bulk of the model stage. The largest percentage of AI pipeline activity is allocated to the data stage, which, along with inferencing, is a CPU-intensive workload (see Figure 1).

Now that you know the critical role that inference plays, you can use the Intel Xeon Scalable processors in your existing infrastructures to significantly improve overall AI performance. AI platforms built on Intel Xeon Scalable processors support open software environments that can be more cost-effective to purchase, operate, expand, and upgrade than proprietary, training-dedicated solutions. An added benefit is that Intel Xeon Scalable processors can run many non-AI workloads, which can help you extract more functionality and efficiency from your existing data center architecture.

Improve AI performance and lower costs on existing data center infrastructures

Intel Xeon Scalable processors help streamline data workflows and reduce application latency without requiring complicated workarounds. With these improvements, you can build a simplified data center architecture that smoothly connects multiple, diverse AI workloads from end to end across the network.

2nd and 3rd Generation Intel Xeon Scalable processors are the only x86 data center CPUs enhanced with built-in AI accelerators and supported by a broad range of optimized software and tools.[2] Use these AI-enhanced processors to accelerate compute-intensive inference workloads without adding to the complexity of your data infrastructure. Without any additional hardware, 3rd Generation Intel Xeon Scalable processors can be optimized to deliver up to 30 percent higher performance than NVIDIA GPUs across 20 workloads.[3]

Accelerate inference with Intel DL Boost

3rd Generation Intel Xeon Scalable processors deliver significant performance advantages thanks to built-in AI accelerators, such as Intel DL Boost, which includes Vector Neural Network Instructions (VNNI). DL workloads for neural machine translation (NMT), for example, can be significantly accelerated by Intel DL Boost. DL processes language data, taking into account localized usage, changing trends, and idiomatic ambiguities. In order for users to experience real-time interactivity, NMT latency times must be less than 10 µs. Thanks to faster inference processing of DL workloads by their built-in AI accelerators, 3rd Generation Intel Xeon Scalable processors are able to bring NMT latency down to 8.9 µs.[4]

Most DL applications use FP32 precision for inferencing workloads. 2nd Generation Intel Xeon Scalable processors use Intel AVX-512 instructions to accelerate throughput for FP32-based data.[5] Intel DL Boost enables 2nd Generation Xeon Scalable processors to use INT8-based workloads. INT8 convolutions can improve inferencing performance up to 3x faster than FP32—and most importantly, with minimal loss to accuracy.[6] And 3rd Generation Intel Xeon Scalable processors optimized with Intel DL Boost can accelerate INT8-based inferencing performance by as much as 1.56x compared to 2nd Generation Intel Xeon Scalable processors.[7] For AI training workloads that do not require high levels of precision, Intel DL Boost uses brain floating-point format (BF16) to improve performance.[8]

Optimize AI performance with an open software ecosystem

Purpose-built for open software environments, 3rd Generation Intel Xeon Scalable processors facilitate integrating AI applications from edge to cloud for shorter time to solution or to production. Deploying AI in an open software environment can make administration more efficient, which can help lower TCO. IT staff can integrate open source software with the confidence it will operate as intended. They can use familiar tools to optimize AI performance without implementing complicated workarounds or needing specialized knowledge. Open tools let IT staff fine-tune AI performance with "write once, deploy anywhere" efficiency.

3rd Generation Intel Xeon Scalable processors with built-in accelerators can benefit AI frameworks such as TensorFlow and PyTorch.[9] Intel DL Boost can improve DL workload performance for TensorFlow up to 16x, and can improve PyTorch performance up to 53x, compared to unoptimized frameworks.[10] After optimizing with Intel oneAPI Data Analytics Library (oneDAL), scikit-learn performance on 2nd Generation Xeon Scalable processors improved up to 100x.[11]

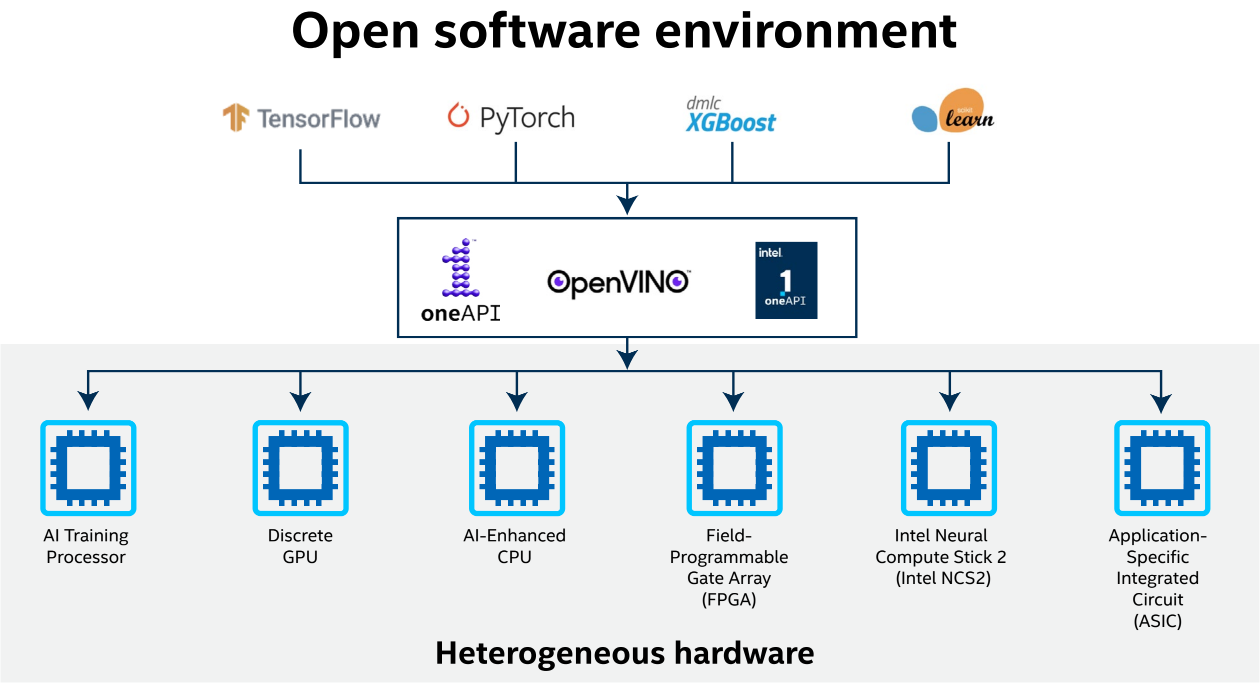

Open software environment
TensorFlow
PyTorch
dmlc XGBoost
scikit learn
oneAPI
OpenVINO
intel. 1 oneAPI
AI Training Processor
Discrete GPU
AI-Enhanced CPU
Field-Programmable Gate Array (FPGA)
Intel Neural Compute Stick 2 (Intel NCS2)
Application-Specific Integrated Circuit (ASIC)
Heterogeneous hardware

Figure 2. Using the Intel Distribution of OpenVINO™ toolkit for integration, Intel Xeon Scalable processors support ML/DL frameworks and libraries across heterogenous hardware: GPUs, CPUs, field-programmable gate arrays (FPGAs), Intel Neural Compute Stick 2 (Intel NCS2), vision processing units (VPUs), and application-specific integrated circuits (ASICs)

Get business insights sooner with the Intel Distribution of OpenVINO toolkit

The Intel Distribution of OpenVINO toolkit helps improve AI performance on Intel Xeon Scalable processors with an easy-to-use computer-vision (CV) library and pre-trained models.[12] You can speed up media analytics by using the Intel Distribution of OpenVINO toolkit to accelerate the AI inference pipeline. Media analytics processes audio and video streams produced by Internet of Things (IoT) sensors and devices. AI transforms these media streams into actionable insights through neural network training, inference, and analytics. The Intel Distribution of OpenVINO toolkit helps 3rd Generation Intel Xeon Scalable processors boost inferencing performance for certain image-classification workloads by as much as 47 percent, compared to 2nd Generation Intel Xeon Scalable processors.[13]

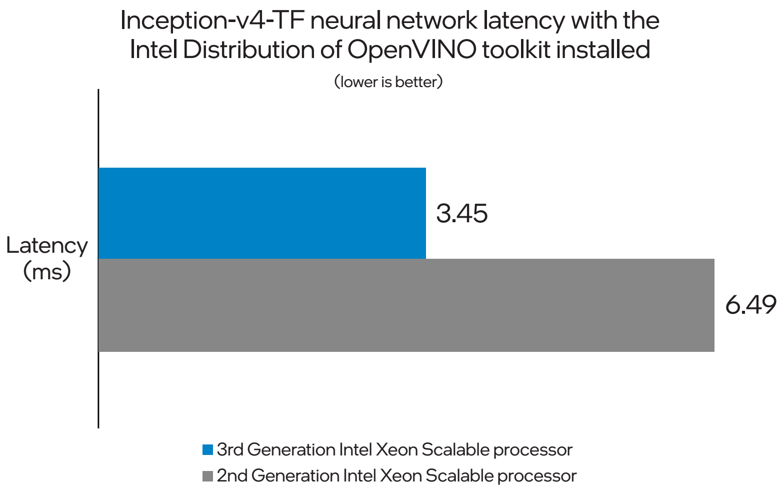


Figure 3. Reduce inferencing latency for Inception-v4-TF models by up to 47 percent with the Intel Distribution of OpenVINO toolkit and 3rd Generation Intel Xeon Scalable processors, compared to 2nd Generation Intel Xeon Scalable processors[13]

The Intel Distribution of OpenVINO toolkit, which includes the Model Optimizer API and Open Model Zoo, lets you automate, optimize, tune, and run AI inferencing with little or no coding knowledge. A built-in inference engine supports CV accelerators across heterogeneous hardware, including CPUs, GPUs, FPGAs, and Intel NCS2.

Test-drive AI applications on Intel DevCloud

Intel DevCloud is a device sandbox that gives you access to everything you need to develop AI applications. Develop, test, and run AI workloads on a cluster with the latest Intel Xeon processors. With Intel DevCloud, you can explore the no-code AI accelerators of the Intel Distribution of OpenVINO toolkit and try out the DL inference samples created for the ONNX framework.

Get more AI accelerator features with 4th Generation Intel Xeon Scalable processors

While GPUs and other discrete accelerators provide outstanding performance for certain AI deployments, innovative AI-accelerated CPUs continue to expand cost-efficient and performant processing capabilities across the AI pipeline. Looking ahead on Intel's CPU roadmap, 4th Generation Intel Xeon Scalable processors will deliver new and improved built-in accelerators. Dual-socket servers using 4th Generation Intel Xeon Scalable processors, for example, can infer more than 24K images/second, comparing CPUs favorably with dedicated processors.[14,15]

Speaking of the future, standardizing AI deployments on Intel hardware can help ensure that upgrading to 4th Generation Intel Xeon Scalable processors can improve inferencing performance *without* disrupting AI workflows or adding to platform complexity. A new, scalar architecture has plenty of capacity for more AI features in future generations of Intel Xeon Scalable processors.

So, what AI innovations can you expect from 4th Generation Intel Xeon Scalable processors?

Optimize efficiency for AI and non-AI processing with Intel Advanced Matrix Extensions

The 4th Generation Intel Xeon Scalable processors will include a new AI accelerator, Intel Advanced Matrix Extensions (Intel AMX).[16] Intel AMX gives INT8 inferencing workloads another boost, up to 8x more operations per clock per core compared to current-generation Intel AVX-512 with VNNI acceleration.[17]

The Intel AMX architecture and instructions function like a systolic array to efficiently process matrix multiplications. In other words, Intel AMX enables a 4th Generation Intel Xeon Scalable processor to handle training workloads and DL algorithms like a GPU does. And unlike accelerators dedicated to training, Intel AMX accelerators also accelerate CPU-preferred processing for other AI stages and non-AI workloads.

4th Generation Intel Xeon Scalable processors can be fine-tuned to peak efficiency by using the Intel oneAPI Deep Neural Network Library (oneDNN). oneDNN is part of the oneAPI toolkit and integrated into TensorFlow and PyTorch AI frameworks and with the Intel Distribution of OpenVINO toolkit. And you can also use the oneAPI toolkit to write instructions that remove the administrative burden of manually assigning the right accelerator to an AI or non-AI workload. This automation enables the 4th Generation Intel Xeon Scalable processor to run all your data pipeline workloads and automatically scale for peak and non-peak cycles.

Benefit from cost-efficient and performant inference now and in the future

Organizations across all industries are discovering how an open, scalable AI platform powered by Intel Xeon Scalable processors with built-in AI accelerators can deliver outstanding performance and TCO. With 3rd Generation Intel Xeon Scalable processors, you can enjoy cost-efficient and performant inference now; and you can expect even better performance and cost savings with 4th Generation Intel Xeon Scalable processors.

All Intel Xeon Scalable processors support open software environments, open AI frameworks and libraries, and open source tools such as the Intel Distribution of OpenVINO toolkit. Count on an open AI platform to efficiently utilize technologies already in place and provide future-proofed compatibility for expansion, scaling, and upgrades.

Learn more about the "Critical Considerations for AI Deployments" at intel.com/content/www/us/en/products/performance/nvidia-ai-facts.html. Or contact your Intel sales representative to find out how to implement Intel Xeon Scalable processors to help lower TCO and raise the performance of your AI deployment.

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

[2] Intel. "3rd Gen Intel® Xeon® Scalable Processors." intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html.

[3] Source: Claim 44 at Intel. "Performance Index - 3rd Generation Intel® Xeon® Scalable Processors." https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/.

[4] Intel. "Break the Latency Barrier for Real-Time Neural Machine Translation." January 2022. https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Breaking-the-Latency-Barrier-for-Real-Time-Neural-Machine/post/1344750.

[5] Intel. "Intel® Advanced Vector Extensions 512 (Intel® AVX-512)." intel.com/content/www/us/en/architecture-and-technology/avx-512-animation.html.

[6] Intel. "Quantizing ONNX Models using Intel® Neural Compressor." February 2022. https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Quantizing-ONNX-Models-using-Intel-Neural-Compressor/post/1355237.

[7] Source: Claim 121 at Intel. "Performance Index—3rd Generation Intel® Xeon® Scalable Processors." intel.com/3gen-xeon-config.

[8] bfloat16 is only supported by 3rd Gen Intel Xeon Scalable processors, Cooper Lake (code name) release. Source: Claims 1 and 9 at Intel. "Performance Index—3rd Generation Intel® Xeon® Scalable Processors." intel.com/3gen-xeon-config.

[9] Intel. "AI Frameworks." intel.com/content/www/us/en/developer/tools/frameworks/overview.html.

[10] Source: Slides 9 and 11 at Intel. "Software AI Accelerators: The Next Frontier." June 2021. slideshare.net/IntelSoftware/software-ai-accelerators-the-next-frontier-software-for-ai-optimization-summit-2021-keynote-249477197. **TensorFlow:** 1-node, 2 x Intel Xeon Platinum 8380 processor with 1 TB (16 slots, 64 GB, 3,200 MHz) total DDR4 memory, ucode: 0xd000280, Intel HT Technology on, Intel Turbo Boost Technology on, Ubuntu 20.04.1 LTS, 5.4.0-73-generic1, 900 GB Intel SSD operating system (OS) drive; ResNet50 v1.5, FP32/INT8, BS=128, https://github.com/IntelAI/models/blob/master/benchmarks/image_recognition/tensorflow/resnet50v1_5/README.md; SSD-MobileNetv1, FP32/INT8, BS=448, https://github.com/IntelAI/models/blob/master/benchmarks/object_detection/tensorflow/ssd-mobilenet/README.md. Software: TensorFlow 2.4.0 for FP32 and Intel-TensorFlow (icx-base) for both FP32 and INT8, tested by Intel on 5/12/2021. **PyTorch:** 1-node, 2 x Intel Xeon Platinum 8380 processor with 1 TB (16 slots, 64 GB, 3,200 MHz) total DDR4 memory, ucode: 0xd000280, Intel HT Technology on, Intel Turbo Boost Technology on, Ubuntu 20.04.1 LTS, 5.4.0-73-generic1, 900 GB Intel SSD OS drive; ResNet50 v1.5, FP32/INT8, BS=128, https://github.com/IntelAI/models/blob/icx-launch-public/quickstart/ipex-bkc/resnet50-icx/inference; DLRM, FP32/INT8, BS=16, https://github.com/IntelAI/models/blob/icx-launch-public/quickstart/ipex-bkc/dlrm-icx/inference/fp32/README.md. Software: PyTorchv1.5 without DNNL build for FP32 and PyTorch v1.5 + IPEX (icx) for both FP32 and INT8, tested by Intel on 5/12/2021.

[11] Source: Key100 Sandra Rivera, AITI001 Pradeep Dubey, slide 21 at Intel. "Performance Index—Innovation Event Claims." https://edc.intel.com/content/www/us/en/products/performance/benchmarks/innovation-event-claims/.

[12] Intel. "Intel® Distribution of OpenVINO™ Toolkit." intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html.

[13] Rated maximum TDP/socket in watts. Source: Intel. "Intel® Distribution of OpenVINO™ toolkit Benchmark Results." March 2022. https://docs.openvino.ai/latest/openvino_docs_performance_benchmarks_openvino.html#doxid-openvino-docs-performance-benchmarks-openvino. **CPU inference engines: 3rd Generation Intel Xeon Scalable processor:** Neural network: Inception-v4-TF, Intel Xeon Platinum 8380 processor at 2.30 GHz, Intel HT Technology on, Intel Turbo Boost Technology on, 16 x 16 GB DDR4 3,200 MHz, OS: Ubuntu 20.04.1 LTS, kernel: 5.4.0-64-generic, BIOS: WLYDCRB1.SYS.0020. P86.2103050636, BIOS release: March 5, 2021, BIOS settings: Select optimized default settings (change power policy to "performance"), batch size: 1, precision: INT8, number of concurrent inference requests: 80, tested as of March 17, 2022. Rated maximum TDP/socket: 270 W. **2nd Generation Intel Xeon Platinum Scalable processor:** Neural network: Inception-v4-TF, Intel Xeon Platinum 8270 processor at 2.70 GHz, Intel HT Technology on, Intel Turbo Boost Technology on, 12 x 32 GB DDR4 2,933 MHz, OS: Ubuntu 20.04.3 LTS, kernel: 5.3.0-24-generic, BIOS: SE5C620.86B.02.01. 0013.121520200651, BIOS release: December 15, 2020, BIOS settings: Select optimized default settings (change power policy to "performance"), batch size: 1, precision: INT8, number of concurrent inference requests: 52, tested as of March 17, 2022, Rated maximum TDP/socket: 205 W.

[14] Source: Key100 Sandra Rivera, AITI001 Pradeep Dubey, slide 37 at Intel. "Performance Index—Innovation Event Claims." https://edc.intel.com/content/www/us/en/products/performance/benchmarks/innovation-event-claims/.

[15] Towards Data Science. "Accelerating ResNet-50 Training on the IPU: Behind our MLPerf Benchmark." January 2022. https://towardsdatascience.com/accelerating-resnet-50-training-on-the-ipu-behind-our-mlperf-benchmark-2cefe43ab2b2.

[16] Intel. "Intel Architecture Day 2021 (Event Replay)." August 2021. youtube.com/watch?v=3jU_YhZ1NQA&t=4155s.

[17] Source: AMX Performance-core statement and details at Intel. "Performance Index: Architecture Day 2021." edc.intel.com/content/www/tw/zh/products/performance/benchmarks/architecture-day-2021/.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.
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.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
All product plans and roadmaps are subject to change without notice.

Printed in USA 0622/SO/PRW/PDF Please Recycle 351247–001US