Product Brief

Artificial Intelligence (AI)
Intel® Advanced Matrix Extensions (Intel® AMX)

Accelerate AI Workloads with Intel Advanced Matrix Extensions (Intel AMX)

Intel advances AI capabilities on the latest Intel® Xeon® processors and Intel AMX.

Intel Xeon 6980P processors with Intel AMX deliver up to 2x better ResNet-50 performance compared to AMD EPYC 9755 processors.[2]

Optimizing the AI pipeline

Businesses can benefit from applying AI in various scenarios. These range from recommender systems for books and movies to retail digital software that drives large e-commerce sites to natural language processing (NLP) for chatbots and machine translation. The attributes that make AI so valuable—making sense of complex environments and massive datasets and solving previously impenetrable problems—can potentially further revolutionize business.

To optimize AI pipelines, organizations can turn to the latest Intel Xeon processors with Intel Advanced Matrix Extensions (Intel AMX), a built-in AI accelerator. As part of Intel® AI Engines, Intel AMX was designed to balance inference, the most prominent use case for a CPU in AI applications, with more capabilities for training. With Intel Xeon Scalable processors representing 65 percent of the processor units (installed base) that are running AI inference workloads in the data center, selecting the latest Intel Xeon processors with Intel AMX for new AI deployments is an efficient and cost-effective approach to accelerating AI workloads.[1]

Built-in acceleration on the latest Intel Xeon Scalable processors

The latest Intel Xeon processors with Intel AMX and Intel® Deep Learning Boost (Intel® DL Boost) change the game for AI deployments. The latest Intel Xeon processors, which include 4th and 5th Gen Xeon Scalable processors and now Intel Xeon 6 processors, allow IT teams to meet customer service-level agreements (SLAs) today.

Intel Xeon 6 processors introduce a new modular x86 architecture that allows data center architects to configure and deploy infrastructures that are purpose-built for each organization’s unique needs and workloads. Intel AMX is available on Intel Xeon 6 processors with Performance-cores (P-cores), which are optimized for high performance per core and excel at the widest range of workloads in the data center. Intel Xeon 6 processors also are available with a second type of CPU architecture, Efficient-cores (E-cores), which provide the right amount of performance and efficiency across a wide range of workloads.

Intel AMX on Intel Xeon 6 processors helps deliver better performance compared to previous generations and other processors. For example, Intel Xeon 6980P processors deliver up to 2x better ResNet-50 performance and up to 1.85x better BERT-Large performance compared to AMD EPYC 9755 processors.[2,3]

What is Intel AMX?

Intel AMX is an accelerator built into the latest Intel Xeon processors. Intel AMX improves the performance of deep learning (DL) training and inference, making it ideal for workloads like NLP, recommender systems, and image recognition. Imagine an automobile that could excel at city driving and quickly shift to deliver Formula 1 racing performance. The latest Intel Xeon processors deliver this type of flexibility. Developers can code AI functionality to take advantage of the Intel AMX instruction set, and they can code non-AI functionality to use the processor instruction set architecture (ISA). Intel has integrated the Intel® oneAPI Deep Neural Network Library (oneDNN), its oneAPI DL engine, into popular open source tools for AI applications, including TensorFlow, PyTorch, PaddlePaddle, and ONNX.

Intel AMX architecture

Intel AMX architecture consists of two components (see Figure 1):

- The first component is tiles. Tiles consist of eight two-dimensional registers, each 1 kilobyte in size. They store large chunks of data.
- The second component is Tile Matrix Multiplication (TMUL). TMUL is an accelerator engine attached to the tiles that performs matrix-multiplication computations for AI.

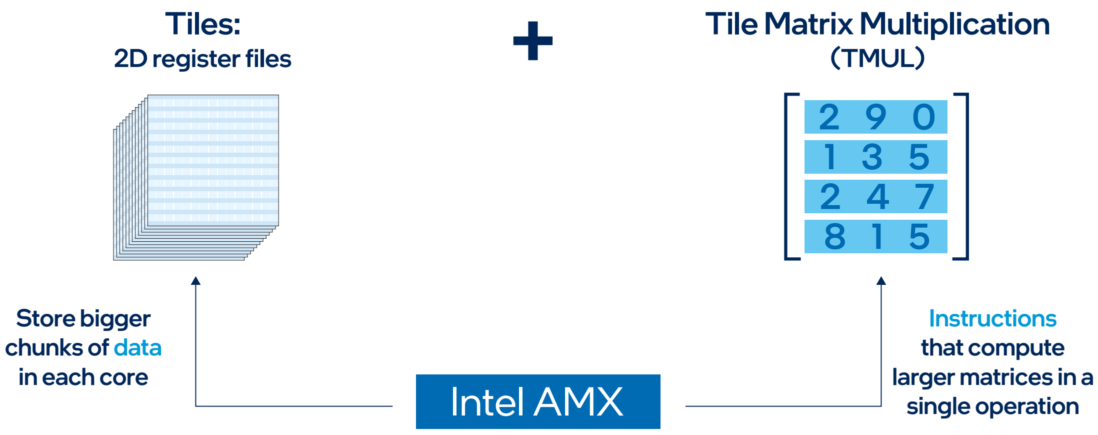
Tiles:
2D register files
+
Tile Matrix Multiplication
(TMUL)
2 9 0
1 3 5
2 4 7
8 1 5
Store bigger chunks of data in each core
Intel AMX
Instructions that compute larger matrices in a single operation

Figure 1. Intel AMX architecture consists of 2D register files (tiles) and TMUL

Intel AMX supports INT8 and BF16 data types:

- INT8 is a data type used for inferencing when the precision of FP32, a single-precision floating-point format often used in AI, isn't needed. Because the INT8 data type is lower precision, more INT8 operations can be processed per compute cycle, which is ideal for real-time applications and matrix multiplication tasks for which speed and efficiency are a priority.
- BF16 is a data type that delivers sufficient accuracy for most training. It can also deliver higher accuracy for inferencing if needed. It enables the training of machine learning (ML) models with nearly the same accuracy as achieved with FP32, yet it incurs only a fraction of the computational cost.
- With this new tiled architecture and support for INT8 and BF16 data formats, Intel AMX generation-on-generation performance gains are significant. Compared to 3rd Gen Intel Xeon Scalable processors running Intel® Advanced Vector Extensions 512 Vector Neural Network Instructions (Intel® AVX-512 VNNI), 4th Gen Intel Xeon Scalable processors running Intel AMX can perform 2,048 INT8 operations per cycle, rather than 256 INT8 operations per cycle. They can also perform 1,024 BF16 operations per cycle,[4] as compared to 64 FP32 operations per cycle, as shown in Figure 2.

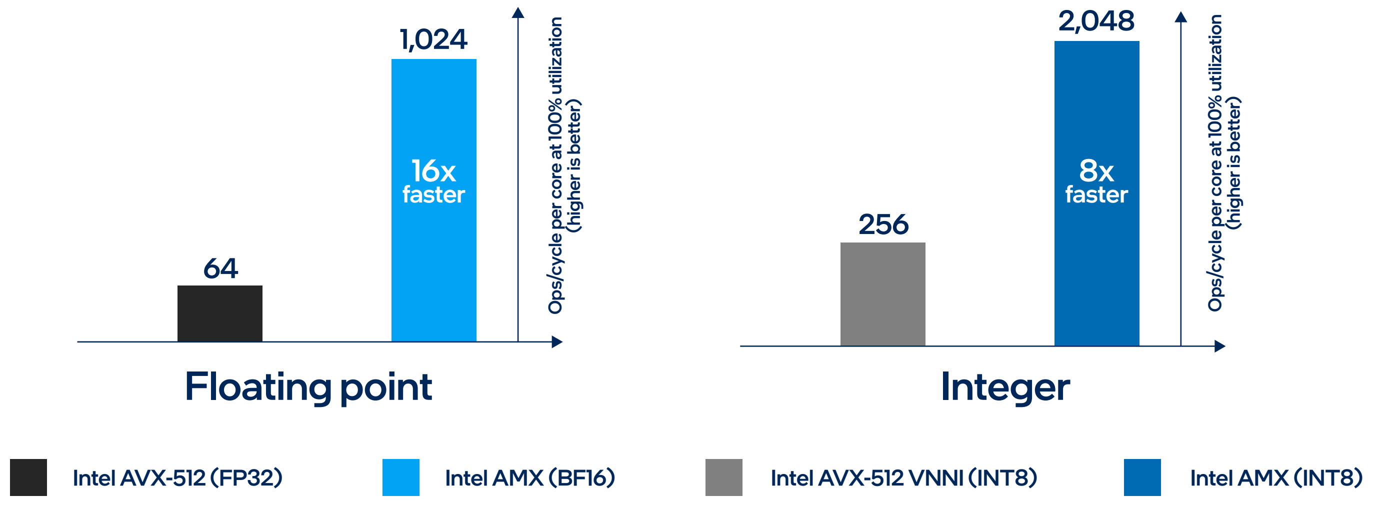

Figure 2. Intel AMX offers better performance than Intel AVX-512 VNNI for INT8 and BF16 data types[4]

AI use cases

The latest Intel Xeon processors with Intel AMX can be deployed in various AI use cases.

Recommender systems
Deliver a customized end-user experience, whether recommending movies and books or showing targeted ads. Create a DL-based recommender system that accounts for real-time user behavior signals and context features such as time and location.

Natural language processing (NLP)
With a global market projected to reach 80.68 billion USD by 2026, NLP applications, including chatbots and sentiment analysis, are critical for businesses to support and scale various functions.[5]

Large language models (LLMs)
Intel AMX accelerates matrix operations by providing dedicated hardware support and is particularly beneficial for applications that rely heavily on matrix computations, such as DL inference. The Numenta Platform for Intelligent Computing (NuPIC) maps neuroscience-based concepts to the Intel AMX instruction set, making it possible to deploy LLMs at scale on Intel CPUs. On Intel-based Amazon Web Services (AWS) instances, inference performance throughput improved up to 70x better than on AMD processor–based instances.[6]

Retail e-commerce software solutions
Supercharge revenue growth and create unparalleled customer experiences by slashing transaction times and effortlessly handling peak demands with DL inferencing and training. Run these workloads on AI-optimized frameworks like PyTorch and TensorFlow.

Figure 3 illustrates how Intel AMX delivers up to 1.8x–2.6x higher real-time inference performance on 5th Gen Intel Xeon processors than on AMD EPYC 9654 processors.[7,8,9] Figure 4 illustrates how Intel AMX delivers up to 1.6x–2.3x higher real-time inference performance per watt on 5th Gen Intel Xeon processors than on AMD EPYC 9654 processors.[7,8,9]

4th and 5th Gen Intel Xeon Processor **Real-Time** Inference Performance[7,8,9]

INT8 inference performance relative to the AMD EPYC 9654 processor (96C)

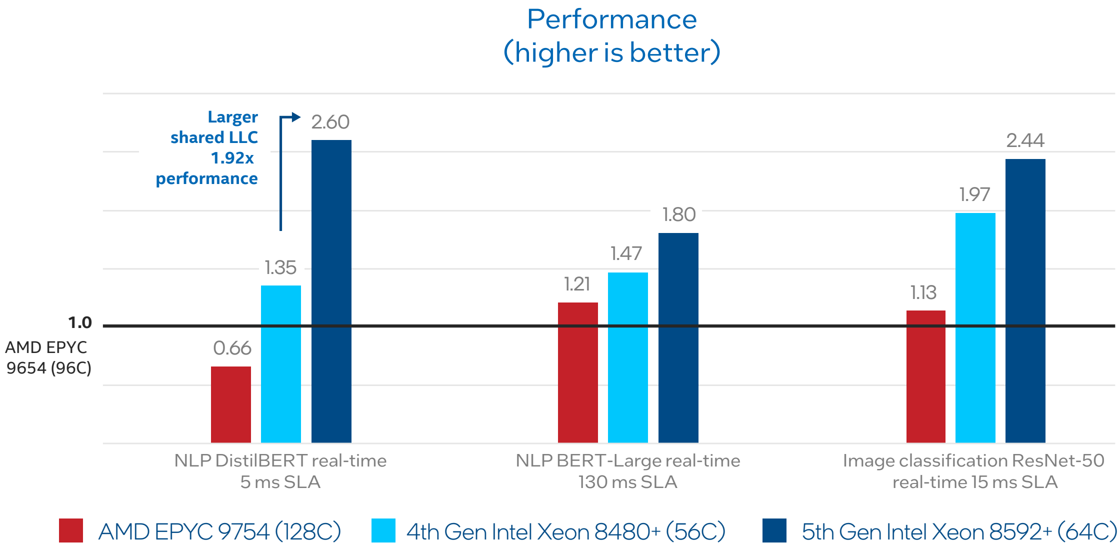

Figure 3. Real-time inference performance[7,8,9]

4th and 5th Gen Intel Xeon Processor **Real-Time** Inference Performance per Watt[7,8,9]

INT8 inference performance relative to the AMD EPYC 9654 processor (96C)

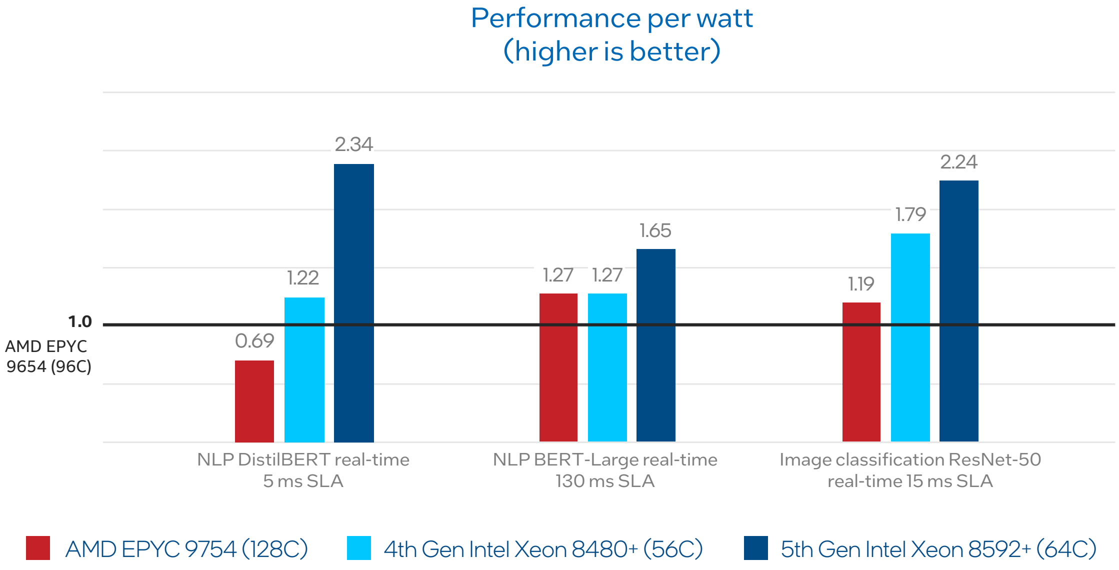

Figure 4. Real-time inference performance per watt[7,8,9]

Figure 5 highlights the TCO advantage that 5th Gen Intel Xeon processors have over 4th Gen AMD EPYC 9554 processors in the context of servers. For the AI-NLP DistilBERT workload, one would need 50 4th Gen AMD EPYC 9554 processor–based servers to achieve the same AI training performance as 15 5th Gen Intel Xeon processor–based servers.[10] Using fewer servers means consuming less energy and generating lower CO_2 emissions. In this scenario, organizations save 62 percent of costs.[10]

5th Gen Intel Xeon Processor TCO Advantages over AMD Processors

A comparison against 50 AMD EPYC 9554 processor–based servers[10]

	AI–NLP DistilBERT
5th Gen Intel Xeon servers	15 servers
Fleet energy saved*	1,496.5 MWh
Reduced CO_2 emissions*	634,428 kg
TCO savings*	$1,300K
TCO delta	**62% savings**

*Estimated over 4 years.

Figure 5. TCO advantages of 5th Gen Intel Xeon processors over AMD EPYC 9554 processors[10]

Figure 6 shows how Intel AMX delivers performance proportionally greater for the incrementally larger core counts in successive generations of Intel Xeon processors, starting with 1st Gen Intel Xeon Scalable processors.

Moore's Law and Accelerators

Targeting the right compute engine for the right workload

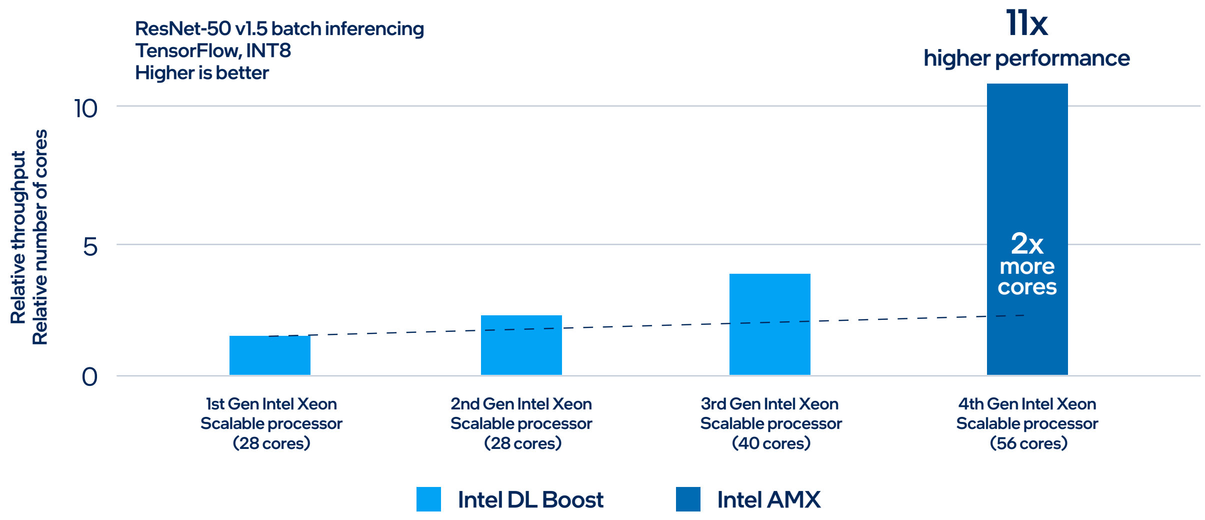

Figure 6. Using the 1st Gen Intel Xeon Scalable processor as a baseline, Intel AMX delivers a non-linear performance improvement compared to previous generations[11]

Get started with Intel AMX

Near-zero effort is required to improve performance with Intel AMX. This is because default frameworks are optimized with Intel oneDNN. Windows and Linux operating systems, kernel-based virtual machines (KVMs), and popular hypervisors expose the Intel AMX instruction set. INT8 and BF16 operations are automatically optimized in open source frameworks like TensorFlow and PyTorch. The Intel® OpenVINO™ toolkit allows developers to automate, optimize, tune, and run AI inferencing with little or no coding knowledge. The only thing developers need to do is to quantize training models to the INT8 data type using the Intel® Neural Compressor.

Intel AMX implementation and developer tools

Intel offers a large set of tools and resources to help developers implement and deploy Intel AMX.

- AI on 4th Gen Intel Xeon Scalable processors tuning guide: Recommendations for tuning 4th Gen Intel Xeon Scalable processors for the best performance in most situations.
- Intel AMX quick-start guide: A document with information and links to the latest Intel-optimized AI libraries and frameworks.
- AI frameworks: Learn more about popular DL and ML frameworks from Intel, including TensorFlow and PyTorch optimizations.
- AI reference kits: Discover AI reference kits for the open source community, with examples of how the downloadable kits are used in real-world applications with tutorials.
- AI and ML development tools: Intel developer resources for every stage of the AI workflow.

Accelerate AI with Intel Xeon Scalable processors

Harness the untapped potential of AI for business by moving to the latest Intel Xeon processors with Intel AMX. Experience exceptional AI training and inference performance with all-new accelerated matrix-multiplication operations while building on the broad foundation of Intel Xeon Scalable processors that are already deployed in the data center.

Learn how Intel AMX can help improve performance for AI workloads: intel.com/ai.

Learn more about 4th Gen Intel Xeon Scalable processors, 5th Gen Intel Xeon processors, Intel Xeon 6 processors, and Intel® Accelerator Engines.

[1] Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2022.

[2] See [9A221] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

[3] See [9A222] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

[4] Based on peak architectural capability of matrix multiply + accumulate operations per cycle per core assuming 100-percent CPU utilization. As of August 2021. For full workloads and configuration details, visit intel.com/processorclaims (search on "Architecture Day 2021"). Results may vary.

[5] Fortune Business Insights. "Natural Language Processing (NLP) Market to Reach USD 80.68 billion by 2026: Increasing Demand for Enhanced Algorithms to Boost Growth, says Fortune Business Insights™." *PR Newswire*. January 2020. prnewswire.com/news-releases/natural-language-processing-nlp-market-to-reach-usd-80-68-billion-by-2026-increasing-demand-for-enhanced-algorithms-to-boost-growth-says-fortune-business-insights-300984381.html.

[6] See [P11] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

[7] 4th and 5th Gen Intel Xeon processor performance with ResNet-50 v1.5 configurations: **Intel Xeon Platinum 8592+ processor–based configuration**: 1-node, 2 x Intel Xeon Platinum 8592+ processor, 64 cores, Intel® Hyper-Threading Technology (Intel® HT Technology) on, Intel® Turbo Boost Technology on, NUMA 2, 1,024 GB total memory (16 x 64 GB DDR5 5,600 megatransfers per second [MT/s]), BIOS 2.0, microcode 0x21000161, 2 x Intel® Ethernet Controller X710 for 10GBASE-T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. TensorFlow = Intel Optimization for TensorFlow 2.13, oneDNN = 3.2, Python 3.8, AI model = ResNet-50 v1.5 (https://github.com/IntelAI/models/), INT8-AMX, real time (BS=1) results while maintaining 15 ms latency service-level agreement (SLA), tested by Intel as of 10/10/2023. **Intel Xeon Platinum 8480+ processor–based configuration**: 1-node, 2 x Intel Xeon Platinum 8480+ processor, 56 cores, Intel HT Technology on, Intel Turbo Boost Technology on, NUMA 2, 1,024 GB total memory (16 x 64 GB DDR5 4,800 MT/s), BIOS 2.0, microcode 0x2b0004d0, 1 x Ethernet interface, 2 x Intel Ethernet Controller X710 for 10GBASE-T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, TensorFlow = Intel Optimization for TensorFlow 2.13, oneDNN = 3.2, Python 3.8, AI model = ResNet-50 v1.5 (https://github.com/IntelAI/models/), INT8-AMX, real time (BS=1) results while maintaining 15 ms latency SLA, tested by Intel as of 10/25/23. **AMD EPYC 9654 processor–based configuration**: 1-node, 2 x AMD EPYC 9654 processor, 96 cores, Simultaneous Multithreading (SMT) on, turbo on, NUMA 2, 1,536 GB total memory (24 x 64GB DDR5 4,800 MT/s), BIOS 1.5, microcode 0xa10113e, 2 x 10 gigabit (Gb) Intel Ethernet Controller X550T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, ZenDNN 4.1, TensorFlow = 2.12.1, Python 3.8, AI model = ResNet-50 v1.5 (https://github.com/IntelAI/models/), INT8, real-time (BS=1) results while maintaining 15 ms latency SLA, tested by Intel as of 09/11/23. **AMD EPYC 9754 processor–based configuration**: 1-node, 2 x AMD EPYC 9754 processor, 128 cores, SMT on, turbo On, NUMA 2, 1,536 GB total memory (24 x 64 GB DDR5 4,800 MT/s),BIOS 1.5, microcode 0xaa00212, 2 x 10 Gb Intel Ethernet Controller X550T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, ZenDNN 4.1, TensorFlow = 2.12.1, Python 3.8, AI model = ResNet-50 v1.5 (https://github.com/IntelAI/models/), INT8, real time (BS=1) results while maintaining 15 ms latency SLA, tested by Intel as of 10/26/23.

[8] 4th and 5th Gen Intel Xeon processor performance with BERT-Large configurations: **Intel Xeon Platinum 8592+ processor–based configuration**: 1-node, 2 x Intel Xeon Platinum 8592+ processor, 64 cores, Intel HT Technology on, Intel Turbo Boost Technology on, NUMA 2, 1,024 GB total memory (16 x 64 GB DDR5 5,600 MT/s), BIOS 2.0, microcode 0x21000161, 2 x Intel Ethernet Controller X710 for 10GBASE-T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, framework = PyTorch 2.0, IPEX = 2.0, Python 3.8, AI model = BERT-Large (https://github.com/IntelAI/models/), INT8-AMX, real time (BS=1) results while maintaining 130 ms latency SLA, tested by Intel as of 10/10/2023. **Intel Xeon Platinum 8480+ processor–based configuration**: 1-node, 2 x Intel Xeon Platinum 8480+ processor, 56 cores, Intel HT Technology on, Intel Turbo Boost Technology on, NUMA 2, 1,024 GB total memory (16 x 64 GB DDR5 4,800 MT/s), BIOS 2.0, microcode 0x2b0004d0, 1 x Intel Ethernet Controller I225-LM, 11 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = BERT-Large (https://github.com/IntelAI/models/), INT8-AMX, real time (BS=1) results while maintaining 130 ms latency SLA, tested by Intel as of 09/05/2023. **AMD EPYC 9654 processor–based configuration**: 1-node, 2 x AMD EPYC 9654 processor, 96 cores, SMT on, turbo on, NUMA 2, 1,536 GB total memory (24 x 64 GB DDR5 4,800 MT/s), BIOS 1.5, microcode 0xa10113e, 2 x 10 Gb Intel Ethernet Controller X550T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = BERT-Large (https://github.com/IntelAI/models/), INT8, real time (BS=1) results while maintaining 130 ms latency SLA, tested by Intel as of 09/11/23. **AMD EPYC 9754 processor–based configuration**: 1-node, 2 x AMD EPYC 9754 processor, 128 cores, SMT on, turbo on, NUMA 2, 1,536 GB total memory (24 x 64 GB DDR5 4,800 MT/s), BIOS 1.5, microcode 0xaa00212, 2 x 10 Gb Intel Ethernet Controller X550T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = BERT-Large (https://github.com/IntelAI/models/), INT8, real time (BS=1) results while maintaining 130 ms latency SLA, tested by Intel as of 10/26/23.

[9] 4th and 5th Gen Intel Xeon processor performance with DistilBERT configurations: **Intel Xeon Platinum 8592+ processor–based configuration**: 1-node, 2x Intel Xeon Platinum 8592+ processor, 64 cores, Intel HT Technology on, Intel Turbo Boost Technology on, NUMA 2, 1,024 GB total memory (16 x 64 GB DDR5 5,600 MT/s), BIOS 2.0, microcode 0x21000161, 2 x Intel Ethernet Controller X710 for 10GBASE-T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = DistilBERT (https://github.com/IntelAI/models/), INT8-AMX, real time (BS=1) results while maintaining 5 ms latency SLA, tested by Intel as of 10/10/2023. **Intel Xeon Platinum 8480+ processor–based configuration**: 1-node, 2 x Intel Xeon Platinum 8480+ processor, 56 cores, Intel HT Technology on, Intel Turbo Boost Technology on, NUMA 2, 1,024 GB total memory (16 x 64 GB DDR5 4,800 MT/s), BIOS 2.0, microcode 0x2b0004d0, 1 x Intel Ethernet Controller I225-LM, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = DistilBERT (https://github.com/IntelAI/models/), INT8-AMX, real time (BS=1) results while maintaining 5 ms latency SLA, tested by Intel as of 09/05/2023. **AMD EPYC 9654 processor–based configuration**: 1-node, 2 x AMD EPYC 9654 processor, 96 cores, SMT on, turbo on, NUMA 2, 1,536 GB total memory (24 x 64 GB DDR5 4,800 MT/s), BIOS 1.5, microcode 0xa10113e, 2 x 10 Gb Intel Ethernet Controller X550T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = DistilBERT (https://github.com/IntelAI/models/), INT8 real time (BS=1) results while maintaining 5 ms latency SLA, tested by Intel as of 09/11/23. **AMD EPYC 9754 processor–based configuration**: 1-node, 2 x AMD EPYC 9754 processor, 128 cores, SMT on, turbo on, NUMA 2, 1,536 GB total memory (24 x 64 GB DDR5 4,800 MT/s), BIOS 1.5, microcode 0xaa00212, 2 x 10 Gb Intel Ethernet Controller X550T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = DistilBERT (https://github.com/IntelAI/models/), INT8 real time (BS=1) results while maintaining 5 ms latency SLA, tested by Intel as of 10/26/23.

[10] Based on 5th Gen Intel Xeon processors delivering up to 3.49x faster performance than 4th Gen AMD EPYC processors while running real-time NLP inference (DistilBERT), which can drive a fleet a reduction from 50 to 15 servers; over four years, this could save: 1,496.5 MWH of energy, 634,428 kg CO2 emissions, and $1,300K. **Testing configurations: Intel Xeon Platinum 8592+ processor–based configuration**: 1-node, 2 x Intel Xeon Platinum 8592+ processor, 64 cores, Intel HT Technology on, Intel Turbo Boost Technology on, NUMA 2, 1,024 GB total memory (16 x 64 GB DDR5 5,600 MT/s), BIOS 2.0, microcode 0x21000161, 2 x Intel Ethernet Controller X710 for 10GBASE-T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = DistilBERT (https://github.com/IntelAI/models/), INT8-AMX, real time (BS=1) results while maintaining 5 ms latency SLA, tested by Intel as of 10/10/2023. **AMD EPYC 9554 processor–based configuration**: 1-node, 2 x AMD EPYC 9554 processor, 64 cores, SMT on, turbo On, NUMA 2, 1,536 GB total memory (24 x 64 GB DDR5 4,800 MT/s), BIOS 1.5, microcode 0xa10113e, 2 x 10 Gb Intel Ethernet Controller X550T, 1 x 1.7 TB Samsung MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, framework = PyTorch 2.0, IPEX=2.0, Python 3.8, AI model = DistilBERT (https://github.com/IntelAI/models/), INT8, real time (BS=1) results while maintaining 5 ms latency SLA, tested by Intel as of 09/11/23. **Cost calculations for a 50 server fleet with AMD EPYC 9554 processors, estimated as of October 2023**: capital expenditure (CapEx) costs: $1.36; operating expense (OpEx) costs over four years, including power and cooling utility costs and infrastructure and hardware maintenance: $749.7K; energy use in kWh (over four years, per server): 46,573, PUE 1.6; other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394. **Cost calculations for a 15 server fleet with 5th Gen Intel Xeon Platinum 8592+ processors as of October 2023**: CapEx costs: $572K, OpEx costs over four years, including power and cooling utility costs and infrastructure and hardware maintenance costs: $238.3K; energy use in kWh (over four years, per server): 55,475, PUE 1.6; other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394. Costs based on Intel estimates and information from thinkmate.com as of October 2023.

[11] **Software configuration for INT8 measurements**: TensorFlow ResNet-50 v1.5, inference: BS=116 (INT8), 1 instance/socket. oneDNN v2.7, Intel optimized TensorFlow 2.10. Tested by Intel on 10/24/2022. (3rd and 4th Gen Intel Xeon Scalable processors) and 7/19/2022 (2nd and 1st Gen Intel Xeon Scalable processors). **Hardware configurations: 4th Gen Intel Xeon Scalable processor hardware configuration (measured)**: Pre-production platform with 2S Intel Xeon Platinum 8480 processor (56 cores, 350 W thermal design power [TDP]) with 1 TB (8 channels/64 GB/4,800 MHz) total DDR5 memory, using BKC 01, using Intel AMX/INT8 and BF16, CentOS Stream 8, Intel AMX kernels (5.15), measurements will vary. **3rd Gen Intel Xeon Scalable processor hardware configuration (measured)**: 1 node, 2 x Intel Xeon Platinum 8380 processor (40 cores/2.3 GHz, 270 W TDP) processor with 1 TB (8 slots/64 GB/3,200 MHz) total DDR4 memory, ucode 0xd0002f2, Intel HT Technology on, Intel Turbo Boost Technology on, Ubuntu 20.04.2 LTS (Focal Fossa), 5.4.0-73-generic, 1 x Intel SSDSC2CW480A3 OS drive. **2nd Gen Intel Xeon Scalable processor hardware configuration (measured)**: 1 node, 2-socket Intel Xeon Platinum 8280 processor, 28 cores, Intel HT Technology on, Intel Turbo Boost Technology on, 384 GB total memory (12 slots/32 GB/2,933 MHz), BIOS: SE5C620.86B.02.01.0013.12152020065 (ucode: 0x500320a), CentOS Stream 8, 4.18.0-383.el8.x86_64. **Intel Xeon Scalable processor hardware configuration (measured)**: 1 node, 2-socket Intel Xeon Platinum 8180 processor, 28 cores, Intel HT Technology on, Intel Turbo Boost Technology on, 384 GB total memory (12 slots/32 GB/2,666 MHz), BIOS: SE5C620.86B.0X.01.0117.021220182317 (ucode: 0x2006b06), Ubuntu 20.04.2 LTS, 5.4.0-73-generic.

Performance varies by use, configuration, and other factors. Learn more at intel.com/processorclaims.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration 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.

Printed in USA 0325/DR/PRW/PDF Please Recycle 360690-001US