Solution Brief

Security Accelerated
Intel® Xeon® Scalable Processors

Accelerated AI Inference with Confidential Computing

By optimizing workloads for Intel® Accelerator Engines, Fortanix uses 4th Gen Intel® Xeon® Scalable processors to help secure and accelerate AI inference in the cloud

For Confidential AI, both security and performance matter. Intel 4th Gen Intel® Xeon® Scalable processors are designed to help secure and accelerate AI inference. Intel® Accelerator Engines are purpose-built integrated accelerators on Intel® Xeon® Scalable processors that deliver performance and power efficiency advantages across many of today's fastest-growing workloads.

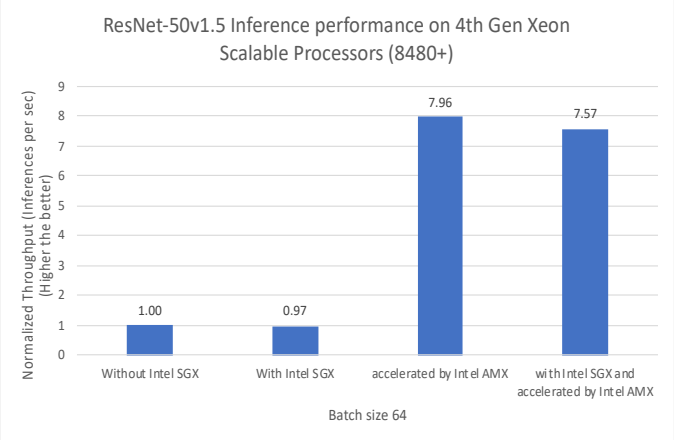
Workloads on the Fortanix Runtime Encryption® (RTE) platform using both Intel Software Guard Extensions (Intel® SGX) and Intel® Advanced Matrix Extensions (Intel® AMX) show up to a 7.57x increase in performance running TensorFlow Resnet50,[1] and up to 5.26x improvement running Bert-Large.[2]

Accelerators like Intel AMX help AI inference workloads achieve outstanding performance, even when combined with hardware-backed security like Intel SGX.

Confidential Computing Helps Secure Cloud-Based AI

Fortanix RTE uses Confidential Computing, powered by Intel® SGX, to enable general purpose computation on encrypted data without exposing plaintext application code or data to the operating system or any other running process. Even if the infrastructure is compromised, or malicious insiders have root passwords, the application remains cryptographically protected.

Confidential Computing allows for the extraction of insights or training of AI models using sensitive data without exposing that data to other software, collaborators, or your cloud provider. This enables business transformation using data that was previously too sensitive or regulated to activate for analytics and other purposes.



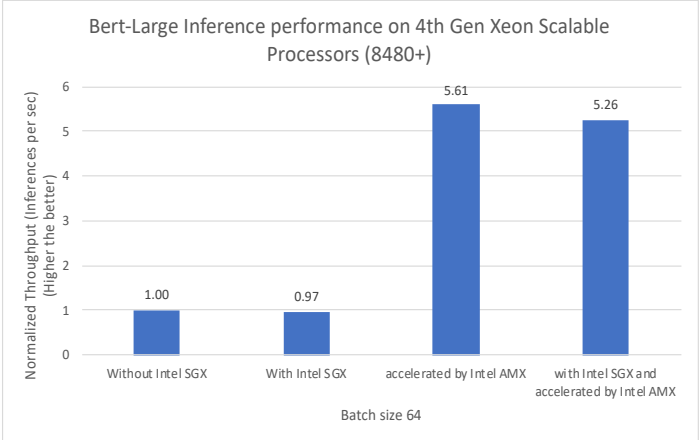


Figure 1. ResNet50 Inference Workload Performance using Intel Xeon Scalable 4th Gen Processors with Intel AMX & Intel SGX.[1]

Figure 2. Bert-Large Inference Workload Performance using Intel Xeon Scalable 4th Gen Processors with Intel AMX & Intel SGX.[2]

Accelerating AI Capabilities with Intel AMX

Intel AMX is a new built-in accelerator that improves the performance of deep-learning training and inference on the CPU and is ideal for workloads like natural-language processing, recommendation systems, and image recognition. Intel advances AI capabilities with 4th Gen Intel Xeon Scalable processors and Intel AMX, delivering higher inference and training performance compared to previous generation Intel Xeon Scalable processors.[3]

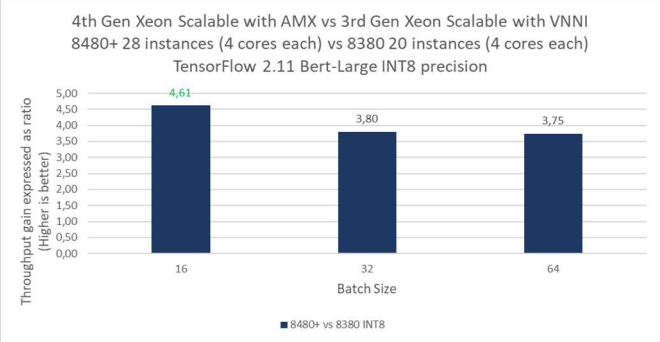


Figure 3. Comparison of Secured DL Inference Workload Performance at INT8 Precision for 3rd Gen Intel Xeon Scalable Processors and 4th Gen Intel Xeon Scalable Processors.[3]

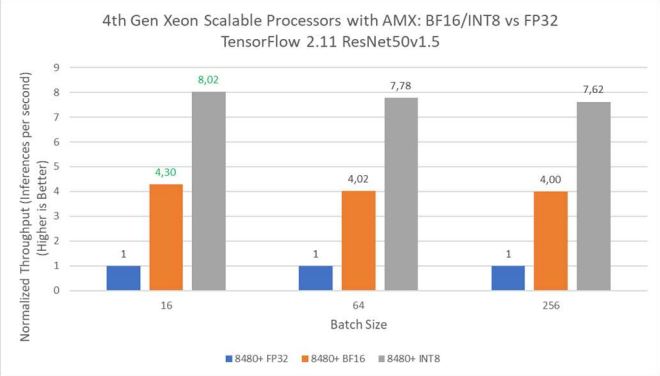


Figure 4. Secured DL Inference Workload Performance for TensorFlow ResNet50 at various batch sizes on 4th Gen Intel Xeon Scalable Processors.[4]

When using Fortanix RTE with both Intel SGX and Intel AMX, performance of Bert-Large inference workload at INT8 precision on 4th Gen Intel Xeon Scalable processors vs 3rd Gen Intel Xeon Scalable processors is up to 4.61x higher.[3]

Get the Most Built-In Accelerators Available

4th Gen Intel Xeon Scalable processors have the most built-in accelerators of any CPU on the market to help improve performance efficiency for emerging workloads, especially those powered by AI.

In addition to performance improvements, 4th Gen Intel Xeon Scalable processors have advanced security technologies to help protect data in an ever-changing landscape of threats, while unlocking new opportunities for business insights. Even when enabling hardware-enabled features designed for security, like Intel SGX, Intel Accelerator Engines like Intel AMX help AI inference workloads achieve outstanding performance.

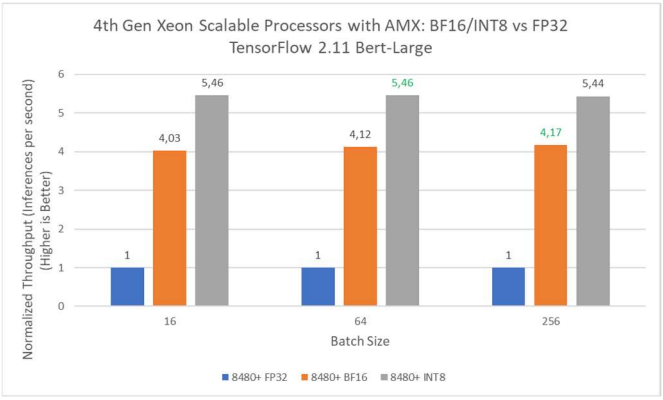


Figure 5. Secured DL Inference Workload Performance for TensorFlow Bert-Large at various batch sizes on 4th Gen Intel Xeon Scalable processors.[5]

With 4th Gen Intel Xeon Scalable processors, performance and security can go hand in hand.

For more information, visit intel.com/xeonaccelerated and fortanix.com/platform/runtime-encryption.

NOTICES AND DISCLAIMERS

1. Up to 7.57x improvement in performance running TensorFlow ResNet50 inference workload on 4th Gen Xeon Scalable processors with Intel SGX and Intel AMX. See configuration details below.
2. Up to 5.26x improvement running Bert-Large inference workload on 4th Gen Xeon Scalable processors with Intel SGX and Intel AMX. See configuration details below.
3. Up to 4.61x improvement in performance running Bert-Large inference workload at INT8 precision on 4th gen Intel Xeon Scalable processors with Intel SGX and Intel AMX vs. Previous Gen. See configuration details below.
4. Up to 8.02x improvement in performance at INT8 precision and up to 4.30x improvement in performance at BF16 precision running TensorFlow ResNet50 inference workload on 4th Gen Intel Xeon Scalable processors with Intel SGX and Intel AMX vs. FP32. See configuration details below.
5. Up to 5.46x improvement in performance at INT8 precision and up to 4.17x improvement in performance at BF16 precision running Bert-Large inference workload on 4th Gen Gen4th Gen Intel Xeon Scalable processors with Intel SGX and Intel AMX vs. FP32. See configuration details below.

CONFIGURATION DETAILS

TEST-1: Test by Intel as of 21 Nov 2022. 1-node, 2x Intel® Xeon® Platinum 8380 CPU @ 2.30GHz , 40 cores, HT Off, Turbo On, Total Memory 512 GB (16x32GB DDR4 3200 MT/s [run @3200 MT/s]), BIOS version SE5C6200.86B.0022.D64.2105220049, ucode version 0xd000375, OS Version Ubuntu 22.04.1 LTS, kernel version 6.0.6-060006-generic , workload/benchmark Deep Learning inferencing in secure enclaves with Fortanix, framework version TensorFlow 2.11 , model name & version ResNet50v1.5/Bert-Large

TEST-2: Test by Intel as of 21 November 2022. 1-node, 2x Intel® Xeon® Platinum 8480+ CPU @ 2.0GHz, 56 cores, HT Off, Turbo On, Total Memory 512 GB (16x32GB DDR5 4800 MT/s [run @4800 MT/s]), BIOS version 3A05, ucode version 0x2b000070, OS Version Ubuntu 22.04.1 LTS, kernel version 6.0.6-060006-generic, workload/benchmark Deep Learning inferencing in secure enclaves with Fortanix, framework version TensorFlow 2.11, model name & version ResNet50v1.5/Bert-Large

Performance varies by use, configuration, and other factors. Learn more on the Performance Index site.
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.