Solution Design Brief

Data Centers | Confidential Computing

Intel Accelerated Solution

# Securely Use Confidential Data with Intel® Software Guard Extensions and Fortanix Confidential AI

**Business Challenge:** How do you safeguard sensitive data, valuable intellectual property and competitive insights without slowing AI performance or creating data silos?

Up to

7.57x

Improved Inference Throughput

using 4th Gen Intel Xeon Scalable processors with Intel AMX and Intel SGX

## Solution Overview and Summary

**Solution:** Artificial intelligence (AI) can unlock actionable insights hidden in data. With more data available, more insights are possible. That is why, for example, banks want to use AI solutions on shared data to better detect and prevent financial fraud. Retailers want to securely run AI and analytics in the cloud. Drug researchers want to collaborate to uncover better healthcare options. AI on shared data is a powerful tool. However, each of these scenarios raises concerns about protecting code and data. For example, increasingly strict data regulations can result in fines if data is misused or improperly protected. AI algorithms and models are valuable intellectual property that need to be safeguarded. And the business insights generated through AI inferencing are themselves fuel for a competitive edge. Until now, data could be encrypted at rest and in transit, but not during use. Plus, encrypting data is often perceived as being accompanied by an unacceptable performance penalty.

The Intel® Security Solution for Fortanix Confidential AI platform combines Fortanix Confidential AI, Fortanix Confidential Computing Manager, and new 4th Generation Intel® Xeon® Scalable processors with Intel® Software Guard Extensions (Intel® SGX) and Intel® Advanced Matrix Extensions (Intel® AMX). This powerful blend of technologies enables enterprises to better secure and accelerate AI inference in the cloud and provides a unique capability to encrypt data and code in memory during use. Using this platform, it is possible to deploy unmodified AI applications in secure enclaves, making it easier to share sensitive data and gain more insight while still keeping that data, code and insights safe in any cloud environment. What's more, this high level of protection is delivered with minimal performance overhead.

The platform is a turnkey service, designed for immediate provisioning that's ready to use without requiring extensive AI skills. It also maintains model integrity by securing data and models in secure enclaves. This simplifies protection across all stages of the data security lifecycle so that organizations can easily get up and running with AI.

**Results:** Our testing shows that 4th Gen Intel Scalable processors with Intel AMX and Intel SGX can empower AI inferencing and collaboration by improving inference throughput by up to 7.57x, compared to a system with that doesn't use Intel AMX or Intel SGX, with only a very small decrease in performance compared a system that doesn't use Intel SGX.[1]

### Testing Details

The following approach was used for each published result:

- Run the test three times.
- Take the mean of those three results.

## Test Methodology

The AI inferencing tests used the Intel® Optimization for TensorFlow AI framework. Two models were tested: ResNet50v1.5 for a computer vision use case and BERT-Large for a natural language processing (NLP) use case. Each test varied the number of model instances from a single small instance assigned to a few cores of the processor running on bare metal to many small instances utilizing the full resources of the system. The batch size was also varied from one to several hundred samples, and several precisions were tested, including FP32 and INT8. All the tests were run with and without Intel SGX secure enclaves to characterize the impact of Intel SGX on the workload.

## Results

Figures 1 and 2 illustrate the performance gain provided by using 4th Gen Intel Xeon processors with Intel AMX for the two models. Intel AMX by itself (without Intel SGX) increases ResNet50 inference throughput by almost 8x and BERT-Large inference throughput by up to 5.61x (INT8 precision), compared to a system that doesn't use Intel AMX (FP32 precision).[1]

More importantly, these performance metrics also prove that Intel SGX does not introduce a significant performance penalty—meaning enterprises get the security and performance they need for their AI projects. Using the combination of Intel AMX and Intel SGX, throughput was only 5% less for ResNet50 and 6% less for BERT-Large—a small price to pay for enhanced security and the ability to share data in secure enclaves across on-premises, hybrid and multicloud environments.

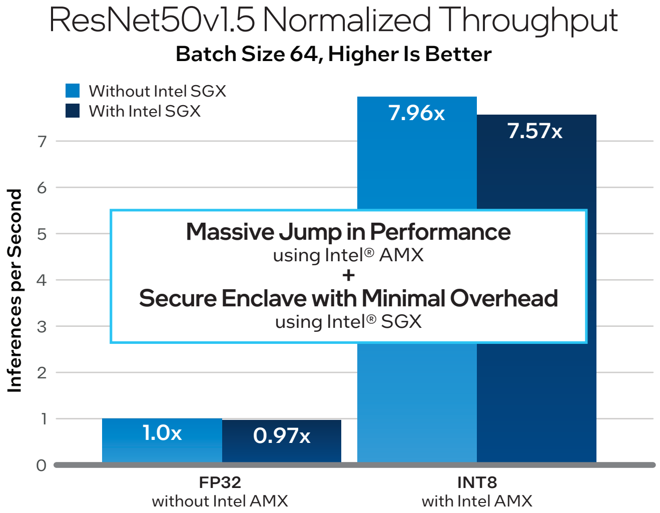

**Figure 1.** For ResNet50, combining Intel® AMX and Intel® SGX supports sharing data in secure enclaves with only 5% overhead.[1]

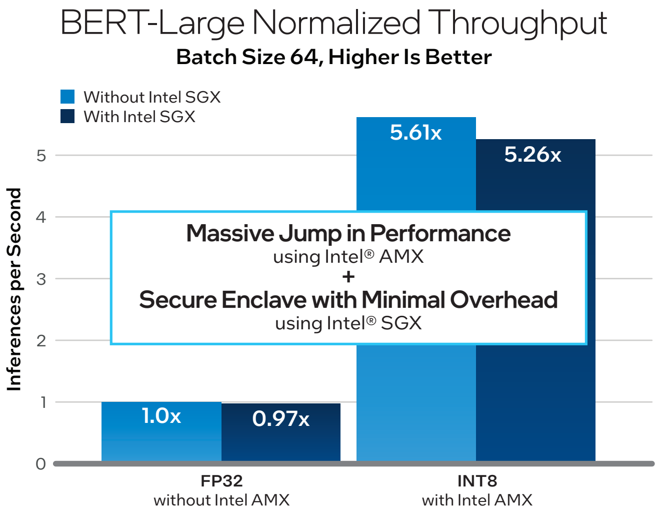

**Figure 2.** For BERT-Large, combining Intel® AMX and Intel® SGX supports sharing data in secure enclaves with only 6% overhead.[1]

## Configuration Details

The following tables provide information about components and settings of the infrastructure used for performance analysis and characterization testing.

| Component | Hardware Configuration |
|---|---|
| Processor | 2x Intel® Xeon® Platinum 8480+ processor (56 cores, 2.0 GHz) |
| Memory | 512 GB (16x 32 GB DDR5 @4800 MT/s) |
| Network Card | Intel® Ethernet Network Adapter X710 |
| Capacity Storage | NAND SSD (at least 2 TB, 2.5" U.2 NVMe, TLC) |

| Important System Settings | |
|---|---|
| Number of Nodes | 1 |
| Intel® Hyper-Threading Technology | Off |
| BIOS CPU Setting | Performance |

| Software Versions | |
|---|---|
| OS | Ubuntu 22.04.1 LTS |
| Workload | ▪ Fortanix Confidential AI<br>▪ Fortanix Confidential Computing Manager |
| Framework | Intel® Optimization for TensorFlow 2.11 |
| Models | ▪ ResNet50v1.5<br>▪ BERT-Large (ww30 gold release) |
| Libraries | ▪ Docker version 20.10.21, build baedalf<br>▪ oneAPI Deep Neural Network Library – oneDNN (included in Intel Optimization for TensorFlow)<br>▪ Fortanix Enclave OS Runtime Encryption Platform 1.20.devel |

| Accelerator Technologies Enabled |
|---|
| Intel® Software Guard Extensions (Intel® SGX) |
| Intel® Advanced Matrix Extensions (Intel® AMX) |
| Intel® Turbo Boost Technology |

## Profiles and Workloads

Residual Network (ResNet) is a popular deep learning model for image recognition. The computer vision inference testing used ResNet50v1.5 with synthetic data. Bidirectional Encoder Representations from Transformers (BERT-Large) is a transformer model that is pretrained on a large corpus of English book texts and Wikipedia data in a self-supervised fashion. BERT has become a ubiquitous baseline for handling various NLP tasks. The NLP inference testing used the pretrained BERT-Large model with the Stanford Question Answering Dataset (SQuAD).

Along with model parameters like batch size, precision and compute allocation (such as number of instances and cores per instance) as described in the Test Methodology section, the tests included variations of BIOS settings like Turbo and Hyper-Threading to arrive at our ultimate configuration recommendation for running these workloads with Intel SGX.

ResNet50v1.5 instances used four cores per instance; BERT-Large instances used 14 cores per instance. Tests were conducted in containers using Docker.

## Conclusion

Until now, the risks of sharing data for better AI hampered collaborative efforts. Because of the strong working relationship between Intel and Fortanix, users can now expect robust inference performance while maintaining security in their infrastructure. Whether your solution is on-premises, in the cloud, or a hybrid, 4th Gen Intel Xeon Scalable processors deliver the latest processor technologies to protect data, code and intellectual property while accelerating time to insight via accelerator technologies built into the processor.

## More Information

- Intel and Fortanix Confidential Computing Manager Joint Solution Brief
- Data Security Manager with Intel® Software Guard Extensions White Paper
- Intel® Security Solution for Fortanix Confidential AI Solution Snapshot
- Intel® Software Guard Extensions
- Fortanix Confidential AI
- 4th Gen Intel® Xeon® Scalable Processors
- Intel® oneAPI
- Intel® Distribution of OpenVINO™ toolkit
- Intel® Optimization for TensorFlow

Learn more about Fortanix Confidential AI.

Contact your Intel representative to learn more about this solution.

**Solution Provided By:**

intel. Fortanix

[1] Testing by Intel as of November 21, 2022. 2x Intel® Xeon® Platinum 8480+ processor (56 cores, 2.0 GHz), 512 GB (16x 32 GB DDR5 @4800 MT/s), BIOS = 3A05, microcode = 0x2b000070, Intel® Ethernet Network Adapter X710, storage = SSDSC2KG96. OS/software: Ubuntu 22.04.1, Fortanix Confidential AI and Fortanix Confidential Computing Manager, Intel® Optimization for TensorFlow 2.11, ResNet50v1.5, and 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. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.

0623/JCAP/KC/PDF