Better Together: Privacy-Preserving Machine Learning Powered by Intel® SGX and Intel® DL Boost

Zongmin Gu
Hongliang Tian
Qing Li
Chunyang Hui
Ant Group

Qiyuan Gong
Dongjie Shi
Wesley Du
Yabai Hu
Jack Chen
Yuan Wu
Ban Hsu
Intel

Introduction

Machine Learning (ML) and Deep Learning (DL) are increasingly important to many real-world applications. ML and DL models are first trained on known data and then deployed to interpret new data, including classifying images, and recommending content. In general, increased data results in a superior ML/DL model. However, stockpiling vast amounts of data also conveys inherent privacy, security, and regulatory risks.

Privacy-Preserving Machine Learning (PPML) helps address these risks. Using techniques such as cryptography differential privacy, and hardware technologies, PPML aims to protect the privacy of sensitive user data and of the trained model as it performs ML tasks.

Ant Group has collaborated with Intel to build a PPML platform on top of Intel® Software Guard Extensions (Intel® SGX) and Occlum, Ant Group's memory-safe, multi-process library OS for Intel® SGX. This blog provides an overview of the solution, which runs on Analytics Zoo. We also show the solution's performance benefits when accelerated with Intel® Deep Learning Boost (Intel® DL Boost) on 3rd Gen Intel® Xeon® Scalable Processors.

Intel® SGX and Occlum

Intel® SGX is Intel's Trusted Execution Environment (TEE), offering hardware-based memory encryption that isolates specific application code and data in memory. Intel® SGX enables user-level code to allocate private regions of memory, called enclaves, which are designed to be protected from processes running at higher privilege levels. (See Figure 1.)

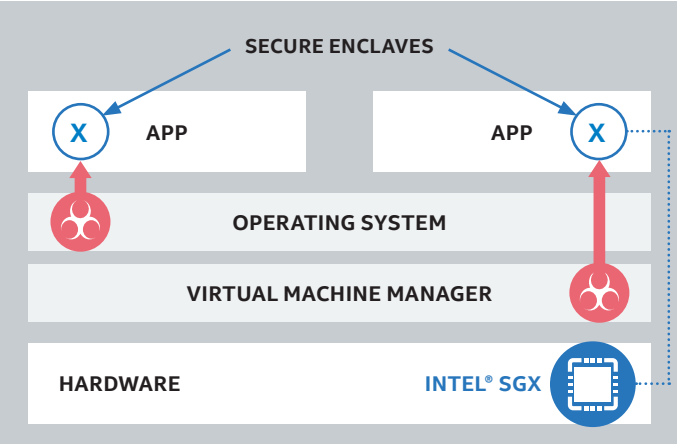

Figure 1. Increased protection through Intel® SGX

Going beyond homomorphic encryption and differential privacy, Intel® SGX protects data against software attacks even if the operating system, drivers, BIOS, virtual machine manager, or system management model are compromised. This enables Intel® SGX to help increase protections for sensitive data and keys even when an attacker has full control of the platform. The 3rd Gen Intel® Xeon® Scalable Processor is available with secure enclaves of up to 512GB per CPU, enabling Intel® SGX to provide an outstanding foundation for PPML solutions..

Ant Group, formally established in 2014, serves over one billion users and is one of the world's leading fintech companies. An active explorer in PPML, Ant Group has initiated an open-sourced project named Occlum, a memory-safe, multi-process library operating system (LibOS) for Intel® SGX. Using Occlum, ML workloads and others can run on Intel® SGX with minimal to zero modifications of source code, thus protecting the confidentiality and integrity of user data transparently. Figure 2 shows the Occlum architecture for Intel® SGX.

End-to-end PPML Solution Built with Analytics Zoo

Analytics Zoo is a unified data analytics and AI platform for distributed TensorFlow, Keras and PyTorch on Apache Spark, Flink, and Ray. With Analytics Zoo, analytics frameworks , ML/DL frameworks, and Python libraries can run as an integrated piece in the Occlum LibOS in a protected manner. Analytics Zoo also provides secure data access, secure gradient and parameter management and other security features that help enable PPML use cases such as federated learning. Figure 3 illustrates the end-to-end Analytics Zoo PPML solution.

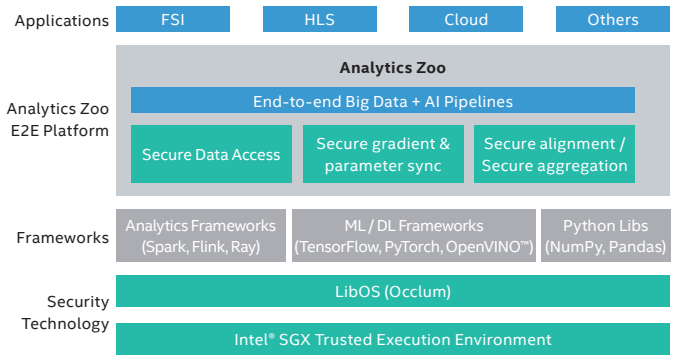

Figure 3. End-to-end PPML solution for secure distributed computing in financial services, healthcare, cloud services, and other applications

With the Analytics Zoo PPML platform, Ant Group has worked with Intel to build a more secure, end-to-end, and distributed inference service pipeline (Figure 4). We constructed the inference service pipeline with Analytics Zoo Cluster Serving, a lightweight distributed, real-time serving solution that supports a range of deep learning models, including TensorFlow, PyTorch, Caffe, BigDL and OpenVINO™ models. Analytics Zoo Cluster Serving components include a web

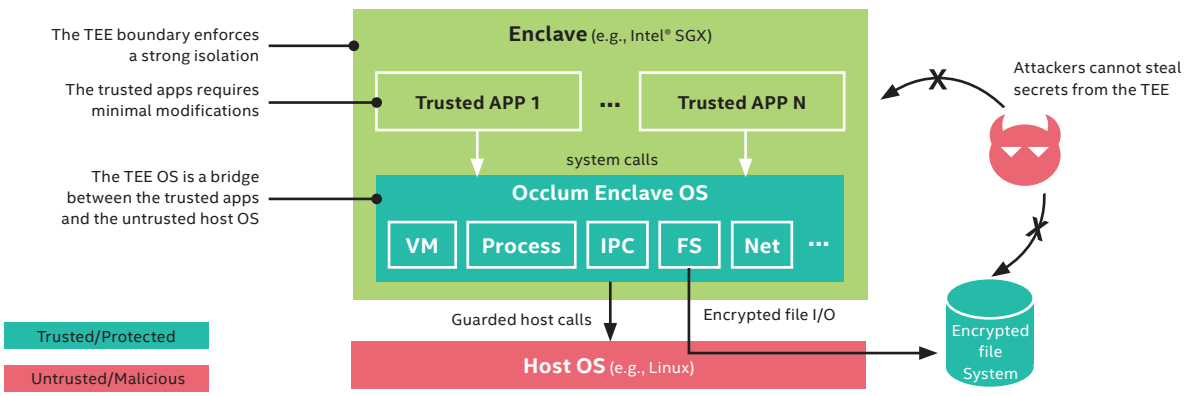

Figure 2. Occlum architecture for Intel® SGX (Image Source: Occlum · GitHub)

front end; Redis, the in-memory data structure store, and an inference engine such as Intel® Optimization for TensorFlow or Intel® Distribution of OpenVINO™ Toolkit. Components also include distributed streaming frameworks such as Apache Flink.

The inference engine and streaming frameworks run on top of Occlum and inside Intel® SGX enclaves. The web front end and Redis are encrypted by the Transport Security Layer (TLS) protocal. As a result, the data in the inference pipeline, including the user data and model, are more protected whether in-store, in transit or in use.

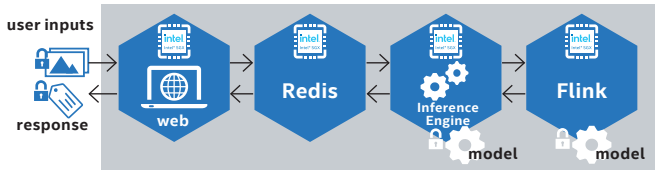

Figure 4. Inference service pipeline

Better Together: End-to-end PPML Solution Accelerated by Intel® DL Boost

The solution implements an end-to-end inference pipeline as following:

1. The RESTful http APIs receive user inputs, and the Analytics Zoo pub/sub APIs feed the user inputs into an input queue managed by Redis. User data is protected by encryption.

2. Analytics Zoo consumes the user inputs from the input queue. It conducts inference using an inference engine on a distributed streaming framework such as Apache Flink. The inference engine and distributed streaming framework are protected by Intel® SGX using Occlum. The Intel® oneAPI Deep Neural Network Library (oneDNN) takes advantage of the Intel® DL Boost with INT8, increasing performance of the distributed inference pipeline.

3. Analytics Zoo collects the inference output from the distributed environment before sending it back to the output queue managed by Redis. Then, the solution uses the RESTful http APIs to return the inference results as predictions to the user. Data in the output queue as well as the http communication are encrypted.

Performance Analysis

Intel and Ant Group validated the performance of the Analytics Zoo PPML solution on a system with 3rd Gen Intel® Xeon® Scalable Processors and other technologies shown in Table 1.

Server	• 3rd Gen Intel® Xeon® Scalable Processors-based servers (2) with 1024GB memory, Intel® SSD and Intel® 10GbE network
System Software	• Occlum Library OS • Ubuntu OS
Application Software	• Analytics Zoo • Intel® Distribution of OpenVINO™ Toolkit • Intel® oneAPI oneDNN 0.19 • Redis • Apache Flink
Workload	• ResNet-50 deep learning model

Table 1. Test Configuration

Figure 5 shows the results of our tests. When the inference solution is protected by Intel® SGX, the ResNet50 inference pipeline experiences a minor loss in throughput compared to an inference pipeline not protected by Intel® SGX. Meanwhile, upon application of the Intel® DL Boost with INT8, the Intel® SGX protected inference pipeline demonstrates a 2x increase in throughput.

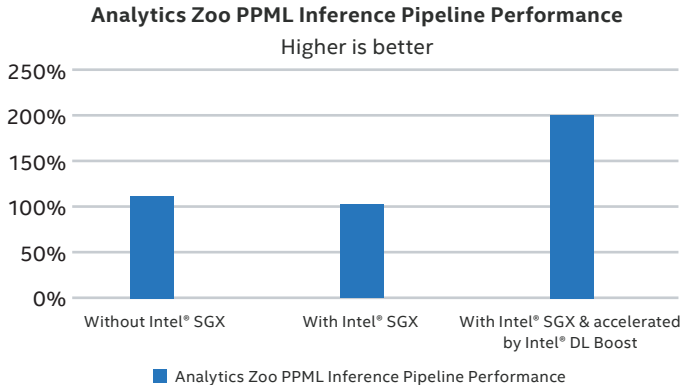

Figure 5. High performance security capabilities with Intel® SGX, Intel® DL Boost, and 3rd Gen Intel® Xeon® Scalable processors

Building on Intel® SGX, the Analytics Zoo PPML solution inherits the merits of a trusted execution environment or TEE. Compared to other data security solutions, it achieves outstanding performance on the security level and data utility level, with only a minor performance impact compared to plain text. Intel® DL Boost and oneDNN further Increase performance for the Analytics Zoo PPML inference solution. Table 2 summarizes the strengths of the solution (TEE) compared to homomorphic encryption (HE), differential privacy (DP), secure multi-party computation (MPC), and plain text.

	TEE	HE	DP	MPC	Plain Text
Security Level	★★★★★	★★★★★	★★★★	★★★★★	NA
Performance	★★★★	★	★★★★	★★★	★★★★★
Data Utility	★★★★★	★★★★	★	★★★	★★★★★

Table 2. Comparing the Analytics Zoo PPML Solution (TEE) to Other Approaches

Summary

In the increasingly complex legal and regulatory environment, it is more important than ever for organizations to safeguard customers' data privacy. PPML allows organizations to continue to explore powerful AI techniques while working to minimize the security risks associated with handling large amounts of sensitive data.

Analytics Zoo PPML, built with Occlum, Intel® SGX , Intel® DL Boost, and Analytics Zoo, establishes a solution platform to help ensure data security and performance for big data AI workloads. Ant Group and Intel have jointly implemented and verified the PPML solution and will continue to explore the best practice in in AI and data security.

Test Configurations

System Configuration: 2-node, Intel® Xeon® Platinum 8369B Processor, 2 sockets, 32 cores per socket, HT On, Turbo ON, Total Memory 1024 GB (16 slots/ 64GB/ 3200 MHz), EPC 512GB, SGX DCAP Driver 1.36.2, Microcode: 0x8d05a260, Ubuntu 18.04.4 LTS, 4.15.0-112-generic kernel. Tested by Intel on 3/20/2021.

Software Configuration: LibOS Occlum 0.19.1, Flink 1.10.1, Redis 0.6.9, OpenJDK 11.0.10, Python 3.6.9;

Workload Configuration: Model: Resnet50, Deep Learning Framework: Analytics Zoo 0.9.0, OpenVINO™ 2020R2, Dataset: Imagenet, BS=16 per instance, 16 instances/2 socket, Datatype: FP32/INT8.

All performance data is tested in lab environment.

Learn More

- Intel® SGX
- Intel® DL Boost
- 3rd Gen Intel® Xeon® Scalable Processors
- Occlum
- Analytics Zoo