Case Study

3rd Gen Intel® Xeon® Scalable Processors
Intel® IPP-Cryptography Library
Intel® AVX-512 IFMA Instruction Set
Federated Learning
Homomorphic Encryption/Partial Homomorphic Encryption

Accelerating Secure Computing for Federated Learning

3rd Gen Intel® Xeon® Scalable Processors empowers WeBank federated learning open-source framework FATE by accelerating partial homomorphic encryption with Intel® IPP- Cryptography library

WeBank 微众银行

"Federated learning open-source framework is designed to help users build federated modeling solutions more efficiently and quickly, to create better-performing AI models, using rich multi-source data. The modular exponentiation operation of partial homomorphic encryption in our FATE (Federated AI Technology Enabler) framework has been enhanced significantly through the introduction of the multi-buffer function provided by Intel® IPP- Cryptography library, helping improve the overall efficiency of user scenarios as well as reducing TCO."

Qian Xu
Vice General Manager of AI Department
WeBank

Large-scale high-quality data sourced from different providers have been proven to effectively enhance the application efficiency of Artificial Intelligence (AI). However, for data security and privacy protection considerations, the shared modeling of multi-source data requires more efficient and safer privacy computing solutions. To this end, WeBank, which has been focusing on exploring and promoting federated learning, uses the leading FATE (Federated AI Technology Enabler) open source platform to help users quickly build federated learning solutions.

Homomorphic encryption (HE) is a commonly used privacy technology in federated learning. HE enables computation directly on the encrypted data, without use of a secret key. HE thereby ensures data security, though at a significant computational overhead compared to unencrypted computation. In order to improve the computational efficiency of the solution, WeBank partnered with Intel to accelerate the modular exponentiation operation of Partial Homomorphic Encryption (PHE) by using the multi-buffer functions provided by Intel® Integrated Performance Primitives Cryptography (Intel® IPP- Cryptography) library. Enabled for the 1-socket and 2-socket server-oriented 3rd Gen Intel® Xeon® Scalable Processors, with the Integer Fused Multiply-Add (IFMA) instruction set newly added by Intel® Advanced Vector Extensions 512 (Intel® AVX-512) at its core. By doing so, the overall operating efficiency of federated learning solution based on the FATE framework was improved significantly. So far, this optimization solution has been successfully tested and verified, and it is planned for rollout to users. This will help users improve the efficiency of federated learning solutions, while effectively reducing the Total Cost of Ownership (TCO).

WeBank Provides a Premium Federated Learning Open-source Framework

With algorithms becoming increasingly sophisticated and computing power continually enhanced, a growing number of commercial AI applications have gained momentum in various industries. In the financial industry, for example, some banks are managing credit risks through AI-based intelligent risk control models, to reduce the rate of their non-performing loans. However, this requires the support of abundant and rich sourced data for such AI applications to yield better efficiency.

Multi-source data collaboration and sharing are by no means easy. On the one hand, when it comes to business sensitive information, most organizations are very cautious when it comes to data sharing. Even for different departments in the same organization, information transfer requires multiple levels of approval. On the other hand, data security and privacy protection have drawn increased public attention. This brings

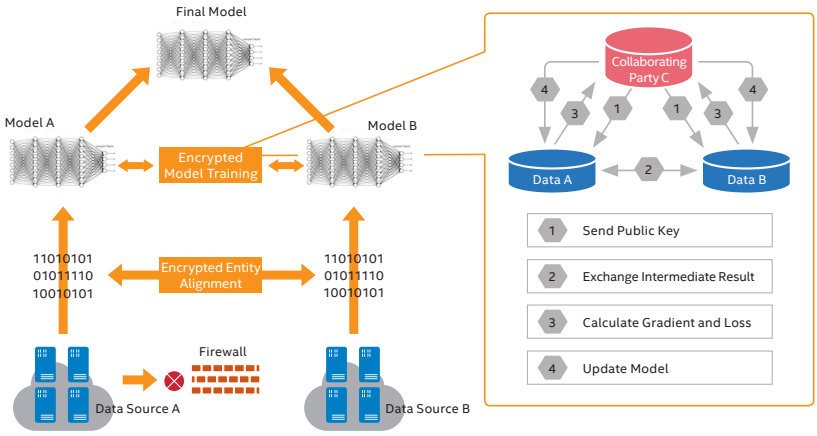

Figure 1 A Basic Federated Learning Infrastructure

ever increasing scrutiny to data security issues during the interaction, transfer, and aggregation of multi-source data.

Therefore, the AI teams in every organization are seeking privacy computing solutions with higher security and higher efficiency.This enhances multi-source data collaboration when modeling, to provide larger-scale and higher-quality data for AI model training.

Among them, the federated learning method is a privacy computing method that has gained widespread attention in recent years. Compared with previous Secure Multi-Party Computation, Distributed Machine Learning and Deep Learning methods, the federated learning method has the following features:

- The data from each party resides locally with the participant during the training. The basic process of a federated learning method is illustrated in Figure 1 below.
- With all data from each party participating in the training process, the loss of accuracy for federated learning is more manageable compared to the model trained with all the data merged.
- Such a training process does not compromise privacy and security, and each party can participate in and facilitate the optimization of the AI model and share the final model, without having to disclose raw data.

As an active explorer and participant in federated learning, WeBank has led the way by releasing its industrial federated learning open-source framework, called FATE, to help improve the convenience and efficiency of building a federated learning solution and to quickly integrate with user application scenarios. As a distributed secure computing framework, FATE supports privacy technologies like homomorphic encryption (including full homomorphic encryption and partial homomorphic encryption) and covers different federated learning models such as Horizontal and Vertical Federated Learning as well as Federated Transfer Learning, thus providing high-performance secure computing for machine learning, deep learning, and so on.

Intel has been consistently providing exceptional foundational capabilities for all types of privacy technologies in federated learning solutions, with advanced software/hardware products and technologies. The various homomorphic encryption algorithms supported by the WeBank FATE framework need to perform high-density computing tasks, which can be significantly boosted in performance by taking advantage of the Intel® Architecture Platform.

Now, the advent of 3rd Gen Intel® Xeon® Scalable Processors with 1-socket and 2-socket, has equipped various federated learning tasks with extraordinary computing power. Its integrated Intel® AVX-512 IFMA instruction set, combined with unmatched technologies and products such as Intel® IPP- Cryptography library, has provided specific accelerating capabilities for the partial homomorphic encryption computation.

Accelerating Modular Exponentiation Operation in Partial Homomorphic Encryption with Multi-buffer Function

Homomorphic encryption is one of the most commonly used privacy technologies in federated learning. The so-called homomorphism means that through a certain encryption method, computation performed on the encrypted data will, when decrypted, yield the same result as the computation performed on the unencrypted data. As illustrated in Figure 2, assuming original data A and B came from Party X and Y of federated learning, respectively. Both parties plan to obtain result C through collaborative computing, but neither party wants the counterparty to know its result. In a federated learning environment that uses homomorphic encryption, data A and B can be encrypted separately and perform the operations in the encrypted environment, and the obtained result can be decrypted to get C. Neither party X nor party Y can see each other's the original data during the computing process, as the whole collaborative computing process has been conducted in an encrypted environment. Therefore, the security of the original data is guaranteed.

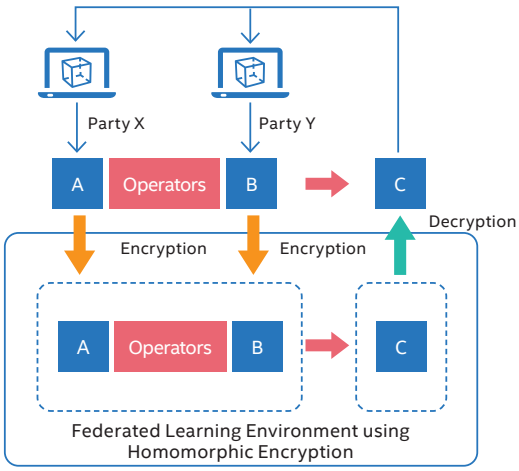

Figure 2 Basic Method for Homomorphic Encryption Protocol

Here is the typical work flow for homomorphic encryption protocol-based federated learning:

- **Key:** Public keys are generated by an arbiter and distributed to each party.
- **Encryption:** encrypt each party's original data and upload it to the federated learning environment.
- **Operation:** to perform operations (e.g., training a model) on encrypted data in a federated learning environment.
- **Decryption:** the resulting optimization model is returned to participants, and then decrypted.

The process of training a model through federated learning using homomorphic encryption requires not only key generation but also encryption and decryption procedures. All kinds of calculations required by the model itself (such as gradient computation) need to be performed in an encrypted environment as well. Based on the encryption strength, the length of integer precision after encryption varies from 1024 bits to 4096 bits, requiring a significant amount of computing resources.

Homomorphic encryption has sustainable and distinct advantages among all sorts of common privacy technologies. However, its high computational complexity leads to enormous computing and time resources overhead for users. As a result, when deploying federated learning solutions, users need to purchase more server resources and spend longer modeling time. This can lead to increased cost and reduced efficiency. For an optimal solution, users need to weigh security against computational complexity, and make a trade-off between homomorphic encryption and other privacy technologies.

For this reason, many users prefer a partial homomorphic encryption solution which has lower computing overhead than leveled or fully homomorphic encryption. But even so, the entire encryption process can be extremely time-consuming if there are hundreds of millions of dataset samples involved in training a model.

To meet the practical needs of different business scenarios, WeBank plans to further accelerate the Paillier algorithm-based partial homomorphic encryption in its FATE framework. Working with Intel, WeBank identified that modular exponentiation is required during the processes of key generation, encryption/decryption, and the multiplication/addition operation of the encrypted data in partial homomorphic encryption, as shown in the encryption/decryption process below:

Encryption: Computing Ciphertext $c = g^m r^n \mod n^2$
Decryption: Computing Plaintext $m = L(c^\lambda \mod n^2) \times \mu \mod n$

In the formulas above, encryption and decryption require 1 and 2 modular exponentiation operations, respectively.

Simply put, the modular exponentiation operation is X^Y mod c. Although the computing itself is not complex, as is described above, the data in a partial homomorphic encryption solution is encrypted, and the length of integer precision reaches thousands of bits. Performing modular exponentiation on such large numbers consumes big amount of computing resources. Typically, modular exponentiation operations can account for a large portion of the cost of the entire compute pipeline.

The FATE framework by default adopts the open-source mathematical operation library, GNU MP Bignum Library (GMP), to perform the modular exponentiation operation. Although the GMP library can support arbitrary-precision mathematical operations, it still uses "serial" computing requests, even if the length of the data involved in the operation reaches thousands of bits.

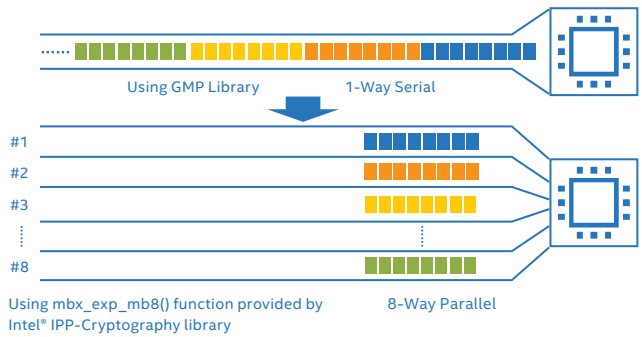

Figure 3 New Computing Request Method Provided by Intel® IPP-Cryptography Library

From Intel's perspective, this has a huge room for optimization. The new Intel® AVX-512 IFMA instruction set integrated into the 1-socket and 2-socket server-oriented 3rd Gen Intel® Xeon® Scalable Processors provides support for Intel® IPP-Cryptography library, which brings an additional function that implements generalized modular exponentiation operation.

As Intel's powerful instruction set for Single Instruction Multiple Data (SIMD), Intel® AVX-512 supports 512 bits vector

calculation—meaning the processor can compute 8 channels of 64-bit integer processing simultaneously. The newly added IFMA subset in the 1-socket and 2-socket-oriented 3rd Gen Intel® Xeon® Scalable Processors enables the processors to perform fused integer multiplication and addition, thereby significantly improving computating efficiency. Using Intel® AVX-512 IFMA, the Intel® IPP-Cryptography library can initiate concurrent 8 parallel computing requests to the processors through the multi-buffer function mbx_exp_mb8() in the modular exponentiation operation, as shown in Figure 3. There is a significant improvement in the computing efficiency of the modular exponentiation operation through the change in computing request methods from serial to parallel, leading to improvements in the overall computing performance of partial homomorphic encryption using modular exponentiation.

FATE Framework Efficiency Improved Leading to Dramatic Reduction in Deployment Cost

Intel and WeBank jointly conducted multi-faceted verification tests, to verify the performance improvement of enabling the multi-buffer function provided by Intel® IPP-Cryptography under the partial homomorphic encryption in the FATE framework.

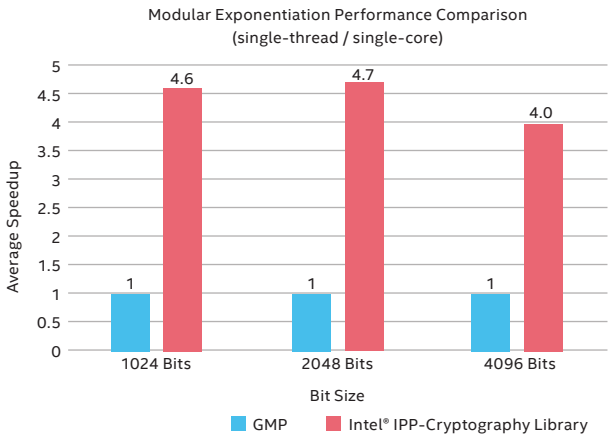

Figure 4 Performance Boosted by Introducing Intel® IPP- Cryptography Library (Normalized Data Comparison)

The test result is shown in Figure 4. By using the multi-buffer function provided by Intel® IPP-Cryptography library, the performance of the modular exponentiation operation needed by the partial homomorphic encryption algorithm has substantial improvement in the various lengths of integer precision, compared to the original GMP library. Notably, the performance increased 4.7X when the length of integer precision is 2048 bits[1].

The substantial increase in performance implies that users require fewer server resources when deploying partial homomorphic encryption for federated learning solutions, thereby effectively reducing the overall TCO. At the same time, thanks to the performance increase of 1-socket and 2-socket server-oriented 3rd Gen Intel® Xeon® Scalable Processors, users can double their productivity when this is used as the computing platform for their federated learning solutions, thereby doing more with less.

Vision

Further advancement in computational efficiency is ensuring users to enjoy faster and easier access to WeBank's FATE—federated learning open-source framework. Using multi-source data enables legal and compliant creation of AI models with excellent performance, thereby allowing various industries to benefit from mature and accessible AI solutions.

Facing the future, WeBank and Intel will continue to team up and collaborate in the following aspects and promote the FATE framework to provide powerful capability support for users to build federated learning solutions:

- Through Intel® IPP-Cryptography library, further optimize partial homomorphic encryption for federated learning solutions.
- Through Intel® IPP-Cryptography library and other hardware/software products and technology, carry out the optimization of federated learning solutions with full homomorphic encryption.

It is worth mentioning that Intel provided a fully homomorphic encryption acceleration library with its open-source project Intel® Homomorphic Encryption Acceleration Library (Intel® HEXL). Bottlenecks in the computation of encrypted data in the solutions can be optimized and eliminated by Intel® HEXL, through the implementation of various functionalities such as forward and inverse negacyclic Number-Theoretic Transform (NTT), element-wise vector--vector modular operations. All these functionalities are being implemented efficiently using the Intel® AVX-512 IFMA Instruction Set. With increasing collaboration between two parties, Intel and WeBank will further explore applications of Intel® HEXL in homomorphic encryption solutions.

FATE Framework Empowers Federated Auxiliary Diagnosis

Federated learning-based smart healthcare not only empowe less capable hospitals to provide high-quality medical testing results and attract more patients, it also assists physicians in diagnosis reduces their workloads. At the same time, localized treatment can also alleviate the burden on patients and their families.

Taking stroke detection as an example, by introducing horizontal federated learning, hospitals with fewer patients can increase the testing accuracy by 10%~20% while maintaining patients' privacy, compared to using only the hospital's patients as training samples[2]. The accuracy of the model will increase with the increasing number of the hospital's case samples added into the federated learning training. Taking China's nearly 2 million patients who receive allopathic medical treatment (strictly limited to those receiving cross-provincial medical treatment) as an example, it is estimated that 200 million RMB in savings can be achieved annually for the pre-disease confirmation phase by covering 10% of the population using federated learning-based disease prediction3[3].

FATE Framework Enforces Federated Credit Risk Control

To address the lack of data for credit rating and approval of credit for small and micro enterprises, the FATE—federated learning framework helps users establish a multi-source data fusion mechanism, to assist relevant financial associations obtain multi-dimensional data, to enrich the feature set. In the process of establishing the feature set, the FATE framework-based federated learning solution can ensure privacy and security for data providers, thus improving the effectiveness of the model.

Statistics suggest that by using multi-dimensional federated data modeling, the effectiveness of the risk control model can be increased by 12%. This will help consumer finance organizations effectively reduce credit approval costs, with total cost savings estimated to reach 5%–10%, thereby further enhancing risk control capability[4].

Recommended Configuration

Thanks to its robust federated modeling and trusted data security safeguards, WeBank's FATE framework has been selected by various organizations, institutes, and research organizations that are looking to explore and deploy federated learning solutions. Based on rich practical experience and proven test results, WeBank and Intel recommend the following hardware configuration to help users achieve efficient modeling in practice and develop better AI models.

Name	Specification
Number of Nodes	2
Processor	3rd Gen Intel® Xeon® Scalable Processors
HT	On
Turbo	On
Memory	256 G (32 G DDR4 3200 x 8)

Test Configuration:
[1] Test conducted by Intel on 3 Sept. 2021, with 2-socket Intel® Xeon® Platinum 8360Y Processor of 2.4GHz, 36 cores, HT on, Turbo Boost on; total memory of 512 GB (64 GB * 8) 3200MT/s; BIOS version: SE5C6200.86B.0022.D64.2105220049 (ucode:0xd0002b1); Operating System: Ubuntu 20.10, 5.8.0-33-generic; Intel® oneAPI DPC++/C++ Compiler version: 2021.3.0 (2021.3.0.20210619); gcc Compiler version: (Ubuntu 10.3.0-1ubuntu1~20.10) 10.3.0
[2.3.4] Data cited from "Federated Learning Whitepaper V2.0", by WeBank, National Engineering Laboratory of Electronic Commerce and Electronic Payment, Peng Cheng Laboratory, Ping An Technology, Tencent Research Institute, Cloud Computing and Big Data Research Institute, and CMG Fintech: https://cn.fedai.org/