Case Study

Confidential AI
Intel® Trust Domain Extensions (Intel® TDX)

Ant Group Develops Confidential Computing for Financial SaaS Solutions with Intel Technologies

Through Alibaba Cloud Elastic Compute Service (ECS) g8i instance family, Ant Group leverages Intel® Xeon® processors, Intel® Trust Domain Extensions (Intel® TDX), and Intel® Advanced Matrix Extensions (Intel® AMX) to gain the benefits of confidential computing for its financial software-as-a-service (SaaS) solutions.

"Ant Group has built a complete confidential PaaS (Platform as a Service) product matrix on Alibaba Cloud ECS instances: the confidential computing engine Occlum and the confidential computing service KubeTEE. Based on this confidential PaaS, Ant Group also offers confidential SaaS solutions for financial scenarios, such as the Ant Privacy-Enhancing Data Analytics Platform, Ant Privacy-Enhancing AI Platform, and more."

— Shuang Liu, Ant Group, Confidential Computing Team Lead[1]

Ant Group, one of the world's leading open Internet platforms, is developing large language models (LLMs) to power new financial solutions in China. These include intelligent assistants for both consumers and industry professionals. End customers can even use their data to fine- tune the LLMs, enhancing the value of their assistants.

On Alibaba Cloud Elastic Compute Service (ECS) g8i instances, Ant Group built a confidential platform-as-a-service (PaaS) product matrix. The company used 4th Gen Intel® Xeon® Scalable processors with Intel® Trust Domain Extensions (Intel® TDX), a hardware-based trusted execution environment (TEE) that helps secure customer data and Ant Group AI models while in use.

Building on its PaaS platform, Ant Group further developed confidential SaaS solutions for financial scenarios, including Ant Privacy-Enhancing Data Analytics Platform and Ant Privacy-Enhancing AI Platform. These solutions use Intel® Advanced Matrix Extensions (Intel® AMX), an Intel® Accelerator Engines, to enhance the performance of matrix-oriented operations in training deep neural networks (DNNs) and model inference.

By using Intel Xeon processors and the Intel TDX security engine, Ant Group can migrate its general virtual machines (VMs) to confidential VMs which run an Occlum-based secure operating system with TEE-specific access control mechanisms. This seamless migration capability enables customers to accelerate product launches while reducing costs. The Alibaba Cloud instances with Intel Xeon processors also provide scalability. As potential demand for services increases, companies can easily deploy additional model instances to expand capacity.

Challenge

In order to bring more value to its customers through intelligent financial assistants, Ant Group is exploring various ways to allow customers to fine-tune its LLMs with their own data. To run these fine-tuning and inference processes, which are all AI workloads, in the cloud, there are several challenges for the technical stack, including the need to:

- **Maintain the confidentiality of proprietary and customer data.** Ant Group is committed to protecting the datasets of end customers, while ensuring the secure access to its own proprietary business data. Only authorized individuals or systems are able to access, interact with, or modify data.

- **Protect its intellectual property (IP).** Like all the other companies, Ant Group needs to protect its IP during the fine-tuning and inference processes. For example, those valuable IPs should never be opened to malicious entities for unauthorized access, use, or replication.
- **Strengthen compliance with regulations.** Companies must comply with regulations when handling sensitive data, privacy, and security domestically and globally. One of the most robust of these is the General Data Protection Regulation (GDPR), enacted by the European Union (EU) and enforced since 2018.

Solution

Ant Group confidential PaaS works on the Alibaba Cloud ECS g8i instance family as a confidential computing solution to fine-tune its model for individual customers' datasets and to run inference processes. The Alibaba Cloud solution enables customers to effectively maintain the confidentiality of their data, protect their IP, and comply with regulations. The Alibaba Cloud ECS g8i instance family also offers the performance to deliver a highly responsive user experience. Finally, the solution provides seamless migration and scalability, making it a good solution for rapidly growing businesses.

Confidential AI

Confidential AI means using a confidential computing infrastructure to run AI workloads. In confidential computing, code is executed inside a hardware-hardened, attested TEE to protect data in use. In the TEE, AI models—including model parameters, weights, and training and inferencing data—in addition to user data can only be accessed by authorized users and software.

Alibaba Cloud offers a confidential computing framework using 4th Gen Intel Xeon Scalable processors with Intel TDX. Ant Group confidential PaaS found that the framework supported its financial-services security and privacy requirements by:

- Allowing the model owners to control data
- Maintaining privacy when collaborating on multiparty analysis
- Strengthening compliance
- Providing hardware-based isolation and access controls

Figure 1 illustrates how Alibaba Cloud uses a confidential computing framework to create a confidential AI environment.

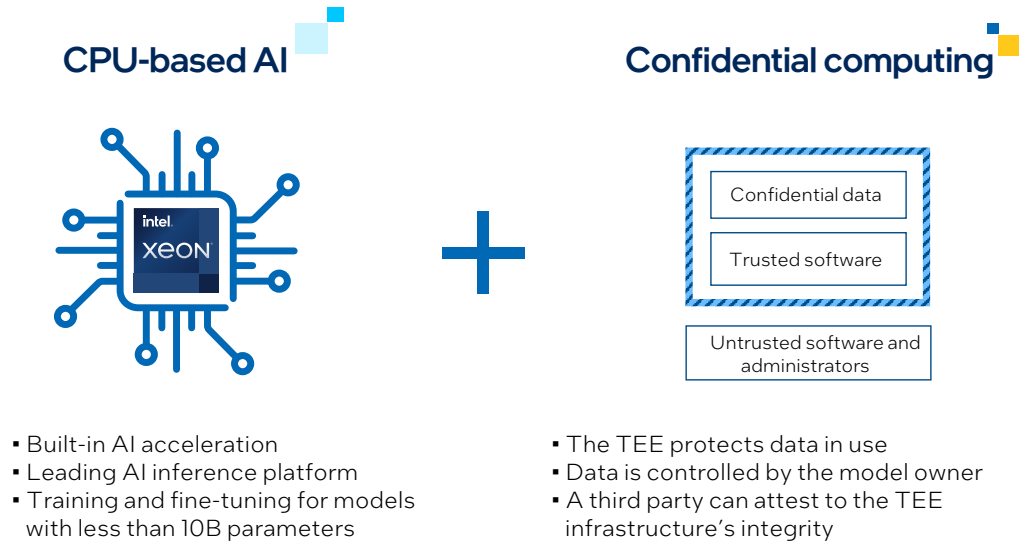

Figure 1. Alibaba Cloud creates a confidential AI environment using Intel TDX.

Data Protection

Typically, cloud data is encrypted both at rest in storage and in transit across networks. However, when data is being actively processed it is often unprotected. Alibaba Cloud VMs use Intel TDX to address this gap, helping to secure data when it's in use by the processor or memory.

Intel TDX processes data in hardware-isolated VMs called trust domains (TDs). These TDs help protect against various threats, including attacks via software, firmware, and the host's cloud stack. Hardware-isolated VMs are especially important when data manipulation and tampering are critical concerns. Occlum TEE operating system may run inside the TD and protect the AI workloads. The blue box in Figure 1 that contains confidential data and trusted software represents the TEE. The blue and black hashed line is the trust boundary.

The trust boundary creates a technological separation between all layers of the software and the admins. In the case of the Alibaba Cloud ECS g8i instance family, the Alibaba Cloud management stack, hypervisor, infrastructure admins, and applications in all other VMs outside the trust boundary cannot access or modify the data inside the trust boundary.

Data Control

Data is encrypted and controlled through several mechanisms. The TEE is designed so that only authorized data owners, authorized software, or trusted organizations can unencrypt and view the data.

TEE Attestation

Attestation is available to verify that the TEE is genuine and that it is updated to comply with current security policies. It can also verify that the software running in the TEE is configured as expected. Attestation is critical in establishing trust with users that the computing platform is security-enabled and that their sensitive data will be safe. KubeTEE provides TEE attestation services and is available on Alibaba Cloud.

Performance

The Alibaba Cloud ECS g8i instance family also features Intel AMX, an accelerator for inferencing and training, which minimizes the need for specialized hardware. Figure 2 shows that 4th Gen Intel Xeon Scalable processors with Intel AMX can deliver up to 5.7–10x higher generation-over-generation real-time inference. By using these Alibaba Cloud instances and their built-in Intel® technologies, Ant Group is able to provide its customers with timely responses to their financial queries.

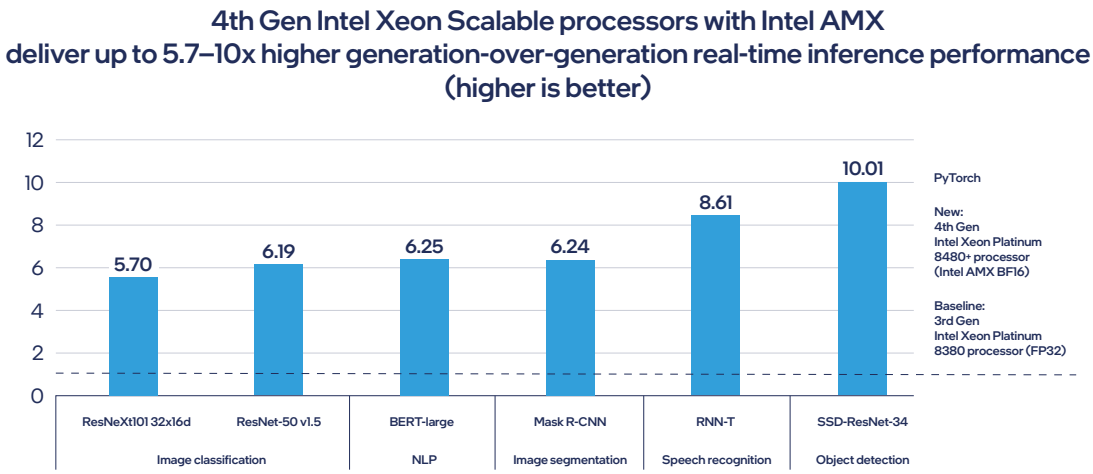

Figure 2. Real-time inference performance with 4th Gen Intel Xeon Scalable processors and Intel AMX.[2]

Simplicity and Scalability

Intel TDX simplifies the migration of AI models from general VMs to confidential VMs in Alibaba Cloud. Ant Group's models were deployed in the Alibaba Cloud ECS g8i instance with a lift-and-shift migration followed by attestation within the new environment. As potential demand increases, additional instances can be deployed easily.

Results

Through working with Alibaba Cloud to run its fine-tuning and inference processes within a confidential AI environment, Ant Group gains the benefits of confidential computing for its financial software-as-a-service (SaaS) solutions. The Alibaba Cloud ECS g8i instance family running on 4th and 5th Gen Intel Xeon processors helps in four ways:

- **Security:** Intel TDX creates a TEE to help protect the security and privacy of data and AI models, enabling customers to comply with regulations more easily.
- **Performance:** Intel AMX accelerates performance to help provide a responsive customer experience.
- **Low cost:** Intel TDX helps to enable seamless migrations from general VMs to confidential VMs, reducing setup time and cost.
- **Scalability:** As demands for services grow, organizations can scale instances on Alibaba Cloud.

Deploy confidential AI today

Learn more about deploying confidential AI on Alibaba Cloud with Intel Xeon processors, Intel Confidential Computing, Intel TDX, and Intel AMX.

Ready to get started? Try out Alibaba Cloud g8i instances today.

[1] Intel. Ant Group interview. February 2024.
[2] See [A17] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Printed in USA 0424/AC/HBD/PDF Please recycle 358250-002US