

Case Study

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

Ant Group Selects Alibaba Cloud with Intel for Confidential AI

Ant Group uses the Alibaba Cloud Elastic Compute Service (ECS) g8i instance family with Intel® Xeon® processors, Intel® Trust Domain Extensions (Intel® TDX), and Intel® Advanced Matrix Extensions (Intel® AMX) to gain the benefits of confidential AI 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 Big Data Analytics Platform, Ant Privacy AI Platform, and more.”

— Shuang Liu, Ant Group,
Confidential Computing Team Lead¹

Ant Group, a technology leader based in China, has developed financial large language models (LLMs) to power new financial solutions. These include intelligent financial assistants for both consumers and industry professionals. End users can even use their data to fine-tune the LLM, enhancing the value of their assistants.

Working with Alibaba Cloud, a global cloud computing and AI leader, Ant Group built a confidential platform-as-a-service (PaaS) product matrix on Alibaba Cloud Elastic Compute Service (ECS) g8i instances. 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 developed confidential SaaS solutions for financial scenarios, including Ant Privacy Big Data Analytics Platform and Ant Privacy AI Platform. These solutions use Intel Advanced Matrix Extensions (Intel AMX), an Intel® Accelerator Engine, to enhance the performance of matrix-oriented operations in training deep neural networks (DNNs) and AI inference.

By using Intel Xeon processors and the Intel TDX security engine, Ant Group can migrate its general virtual machines (VMs) to confidential VMs based on Alibaba Cloud instances without modifications. This seamless migration capability enabled Ant Group to accelerate its launches while reducing costs. The Alibaba Cloud instances with Intel Xeon processors also provide scalability. As customer demand for Ant Group’s services increases, the company can easily deploy additional model instances to expand capacity.

Challenge

Ant Group wanted to maximize the value its customers could realize from its intelligent financial assistants. To this end, the company wanted customers to be able to fine-tune its LLMs with their data. Ant Group planned to run these fine-tuning and inference processes in the cloud. However, to run AI workloads in the cloud, the organization faced several challenges, including the need to:

- **Maintain the confidentiality of proprietary and customer data.** Ant Group needed to secure access to its proprietary data. The company also wanted to protect the datasets of the end customer. The company wanted only authorized individuals or systems to be able to access, interact with, or modify data.
- **Protect its intellectual property (IP).** Ant Group sought to protect its IP during the fine-tuning and inference processes. The company wanted to protect its IP from competitors and malicious entities’ unauthorized access, use, or replication.

- **Strengthen compliance with regulations.** Ant Group must comply with regulations regarding handling sensitive data, privacy, and security domestically and globally. The most robust of these is the General Data Protection Regulation (GDPR), enacted by the European Union (EU) and enforced since 2018.
- **Protect its brand.** Ant Group needed a security-enabled solution to protect its brand. Any security or privacy issues might tarnish the company’s reputation.

Solution

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

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 found that the confidential AI environment supported its AI-based financial-services security and privacy requirements by:

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

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

Confidential AI helps protect data and models in use

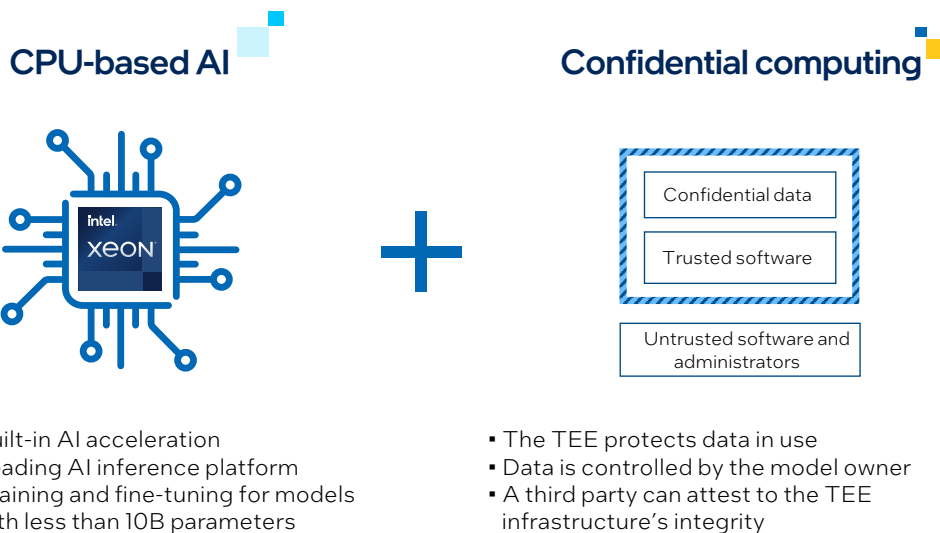


Figure 1. Alibaba Cloud creates a confidential AI environment using Intel TDX

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. 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.

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.

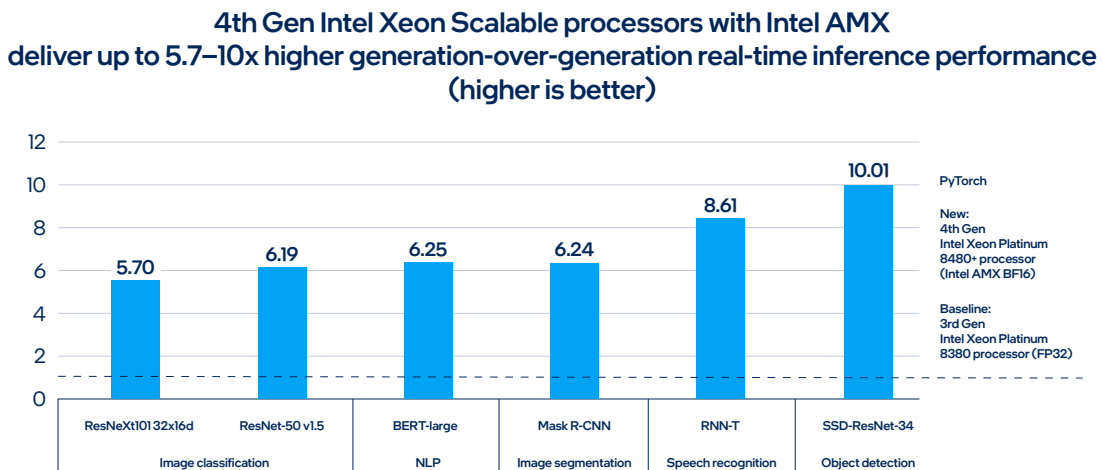


Figure 2. Real-time inference performance with 4th Gen Intel Xeon Scalable processors and Intel AMX²

Simplicity and scalability

Ant Group found that Intel TDX simplified 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 customer demand for Ant Group’s intelligent financial assistants increases, additional instances can be deployed.

Results

By selecting Alibaba Cloud to run its fine-tuning and inference processes within a confidential AI environment, Ant Group is on track for success with its intelligent financial-assistant services. Ant Group benefits from its choice of the Alibaba Cloud ECS g8i instance family running on 4th and 5th Gen Intel Xeon processors in four ways:

- **Security:** Intel TDX creates a TEE to help protect customer data and Ant Group AI models. This enables Ant Group to comply with privacy and security regulations more easily and protect its brand.
- **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 demand for Ant Group's business grows, the organization 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.



¹Intel, Ant Group interview, February 2024.

²See [A17] at [intel.com/processorclaims](https://www.intel.com/processorclaims): 4th Gen Intel Xeon Scalable processors. Results may vary.

Performance varies by use, configuration, and other factors. Learn more at www.intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for additional 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.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.