

Huawei Computing Infrastructure Security Technical White Paper (HCIST) ——From Device-Pipe-Cloud Perspective

Version 1.0
Date 2025-09-18



Copyright © Huawei Technologies Co., Ltd. 2025. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: <https://www.huawei.com/>

Email: support@huawei.com

Contents

1 Challenges in the AI Era	1
1.1 LLM-Driven Transformation	1
1.2 Cloud AI Privacy Protection Dilemma	2
2 HCIST Architecture and Key Technologies	4
3 Computing Infrastructure Security in Computing Scenarios	7
3.1 Kunpeng Computing Infrastructure Security	8
3.2 Ascend Computing Infrastructure Security	9
3.3 Heterogeneous Device Confidential Passthrough with PCIPC	13
4 Computing Infrastructure Security in Storage Scenarios	16
4.1 Security Challenges of Storage Architecture	16
4.2 Storage Hardware Identity	18
4.3 Storage Link Encryption	19
4.4 Dual-Layer Authentication	20
4.5 Storage Passthrough	21
5 Computing Infrastructure Security in Huawei Cloud	22
5.1 Huawei Cloud QingTian Architecture	22
5.2 Components of the Huawei Cloud QingTian Architecture	23
5.3 Huawei Cloud QingTian Confidential Computing	26
5.4 Prospect of Huawei Cloud QingTian Heterogeneous Architecture	30
6 Key Technologies of AI Platform Security	32
6.1 A+K Heterogeneous Confidential Computing Acceleration Platform	32
6.2 Confidential Container	34
7 Pipe Security	38
7.1 Device-Cloud Protocol Security	38
7.2 Intrinsic Security of Network Devices	39
7.3 Communication Security in Scale Up and Scale Out Scenarios	41
7.4 Physical Layer Communication Security	42
8 Typical Applications	44
8.1 On-Device and Cloud Collaborative LLM Inference	44
8.2 Comprehensive Data Protection	46

8.3 LLM Lifecycle Protection	46
8.4 Cloud-Native Cryptographic Applications	47
9 Future Outlook	49

AI technologies are rapidly reshaping the global industrial landscape. As they transform the way we live, learn, and work, they are also driving significant changes in global digital transformation. "Computing power + data + network connectivity" has emerged as the focus for businesses undergoing digital transformation. Over the past 24 months, GPU shipments have surged around the world, and the scale of AI training clusters has expanded from tens of thousands to hundreds of thousands of GPUs. The construction of computing infrastructure is accelerating at an unprecedented pace.

As computing power becomes the primary productivity of businesses in the digital era, it has also emerged as a key target of cyber attacks. Any deceleration or compromise of computing resources can lead to service disruptions. A breach in a GPU cluster may result in tampering with AI training models, delays in financial transactions, and disruptions to manufacturing execution systems. As a core production system of digitalization, computing infrastructure is now classified as a critical information asset. Take auditing as an example. The audit on digital enterprises has expanded beyond financial oversight to encompass comprehensive risk management. Auditors must evaluate enterprise operations through business data, with a focus on computing integrity to ensure that outcomes are both accurate and trustworthy. Some national regulators have required the integrity and traceability of system algorithms, parameters, and logs to be essential components of the audit process. An enterprise's compliance governance capability on computing infrastructure directly influences key business metrics such as stock price, insurance premiums, and financing rates. Without embedding security into the planning, construction, and operation of computing infrastructure, an enterprise's most valuable digital productivity will become the greatest source of risk exposure.

In response to these challenges, Huawei remains committed to its core principle of building open, secure, and trusted computing infrastructure. Through deep integration of its device-cloud synergy architecture and advanced technologies such as confidential computing and confidential storage, Huawei Computing Infrastructure Security Technologies (HCIST) have been developed. HCIST delivers full-stack intrinsic security across chips, firmware, software, single nodes, multi-node systems, clusters, and cloud environments. It strictly follows the principle of "data available but invisible" to safeguard user data and model assets throughout their lifecycle. In computing scenarios, HCIST leverages Kunpeng and Ascend processors to establish hardware-level trust assurance. In storage scenarios, confidential storage technology enables end-to-end (E2E) protection of data at rest and in transit. In cloud environments, the QingTian architecture offers a highly secure and strongly isolated confidential computing framework. Additionally, HCIST ensures the authenticity and trustworthiness of platform identities and running environments through remote attestation and security verification mechanisms. Looking ahead, Huawei will continue to advance research in heterogeneous hardware protection, cross-device secure channel protocols, and compliance frameworks, driving the evolution of computing infrastructure toward enhanced security and higher efficiency.

HCIST will help digital enterprises better incorporate computing integrity into enterprise risk management, ensuring the verifiability and traceability of compute nodes while enabling continuous security monitoring for effective risk control and resolution. We believe HCIST not only lays a robust foundation for data processing in the AI era, but also drives new momentum for the sustainable growth of the global digital economy.



Sean Yang

Global Cyber Security & Privacy
Officer, Huawei

AI has entered a new phase of accelerated advancement from the rapid evolution of foundation models to the widespread deployment of intelligent applications across industries. We are witnessing a profound technological transformation. This revolution is not only reshaping how people work and live, but is also redefining the global industrial landscape. As the "engine" behind AI, computing power is emerging as a core production element of the digital economy.

Yet, as intelligent transformation accelerates, many challenges have emerged. On one hand, the demand for computing power continues to surge, and the imbalance between supply and demand has become a critical bottleneck for large-scale adoption of large language models (LLMs). On the other hand, data—as a key production element of AI—faces growing security and privacy risks during cross-organization and cross-scenario data flows. Unlocking the full value of data while ensuring its security and trustworthiness has become a major challenge for the industry.

Huawei is committed to building secure, trusted, and open computing infrastructure for the intelligent world. To this end, we have established HCIST. Designed with openness at its core, HCIST supports adaptation to diverse heterogeneous environments and enables on-demand integration and collaborative operation of resources from different parties. Vertically, HCIST establishes full-stack security capabilities, spanning from chips, firmware, operating systems (OSes), and clusters to clouds. Horizontally, it harnesses the strengths of device-pipe-cloud synergy to deliver E2E protection for both data and model assets throughout their lifecycle.

With intrinsic security mechanisms and an open architecture, HCIST offers high-performance, scalable, easy-to-migrate, and trustworthy computing infrastructure for a wide range of AI applications. In real-world deployments, HCIST ensures comprehensive security protection from CPUs to NPUs in complex heterogeneous computing environments. In data management, it offers confidential storage, privacy-preserving computation, remote attestation, and other key capabilities. At the network level, it supports secure protocols and encrypted transmission across different devices and regions, establishing a trusted environment for data circulation. With continuous technological innovation and Huawei's full-stack hardware and software capabilities, HCIST has evolved into a systematic security framework, enabling everything from individual devices to ultra-large clusters and addressing diverse security and privacy requirements.

Looking ahead, Huawei will continue to work with industry partners to advance the development of security architectures and standards for heterogeneous and cross-domain synergy, fostering the sustainable and reliable development of global intelligence.



Xiaoyong Zhu

Huawei 2012 Labs

Director of the Trustworthiness Theory,
Technology & Engineering Lab

Executive Summary

AI is rapidly transforming the global industrial landscape. The wave of technologies represented by ChatGPT is not only accelerating LLM adoption across industries, but also elevating AI computing power and data security to the strategic frontier in global technological competition. However, this transformation brings two structural challenges. First, the imbalance between computing power supply and demand significantly limits the large-scale deployment and adoption of LLMs. Second, security risks in cross-organization and cross-region data flows impede collaboration and innovation.

In response to these challenges, Huawei remains committed to its core principle of building open, secure, and trusted computing infrastructure. By leveraging device-cloud synergy architecture and integrating confidential computing, confidential storage, trusted networks, and other security capabilities, HCIST has been developed to safeguard data and model assets throughout their lifecycle.

In computing security, the Kunpeng platform uses the virtCCA/CCA technology to implement confidential computing in the ARM architecture and uses mainstream encryption algorithms and hardware roots of trust (RoTs) to meet stringent security compliance requirements. The Ascend NPU uses NPU TEE to ensure that model weights, user data, and intermediate results remain within the confidential domain during inference, training, and fine-tuning, preventing potential threats from malicious users and administrators.

In terms of storage security, HCIST introduces a confidential storage solution that is designed to establish a trusted foundation at the hardware level. Utilizing different technologies including hardware identity, link encryption, bidirectional authentication, and data passthrough, HCIST has established an E2E data security system. Its architecture—featuring deep integration of storage-compute decoupling and intelligent storage—minimizes security risks across nodes and networks, while maintaining high levels of performance and protection.

In cloud security, Huawei Cloud establishes mechanisms such as physical isolation, secure boot, and hardware identity authentication based on the Qingtian architecture to defend against threats within the cloud platform. Qingtian Enclave technology provides tenants with highly isolated execution environments, enabling the secure operation of sensitive services like encryption modules and confidential AI. This establishes a trusted foundation for deploying critical services in multi-tenant cloud environments.

In terms of AI platform security, HCIST introduces the Ascend+Kunpeng heterogeneous confidential computing acceleration platform, which enables deep integration between CPU trusted execution environments (TEEs) and NPU TEEs. This forms a comprehensive architecture characterized by dual hardware RoTs, E2E runtime isolation, and task-level zero-trust verification. The confidential container technology integrates AI runtime with TEE security protection, ensuring the security and performance of LLMs.

In terms of communication security, HCIST builds device-cloud protocol security and Unisec secure communication system. It uses trusted group key isolation, line-speed encrypted transmission, PHYSec encryption, and other technologies in different computing networks (including scale out and scale up) to ensure secure scheduling of computing resources and secure data transmission across nodes and clusters.

HCIST supports both full-stack integrated deployment and modular, layered, and decoupled deployment. It can be deployed in different hardware platforms and

application scenarios, including device-cloud synergy-based LLM inference, zero-loss data security protection in financial services, confidential storage and secure inference of proprietary models, and cloud-native cryptographic applications. Looking ahead, HCIST will continue to evolve toward post-quantum security, cluster-level confidential computing, distributed RoTs, and comprehensive AI lifecycle protection.

1 Challenges in the AI Era

1.1 LLM-Driven Transformation

AI technologies are rapidly reshaping the paradigms of human society. Among these, LLMs serve as a core driving force, fundamentally redefining how we live, learn, and work. Generative AI, represented by ChatGPT, has sparked an industrial revolution of LLMs. At the 2024 World Artificial Intelligence Conference (WAIC), it was explicitly emphasized that the deep integration of LLMs into industrial scenarios has become a strategic frontier in global technological competition. The 2025 WAIC published Global AI Governance Action Plan, emphasizing the need for AI governance. This includes establishing a widely accepted security governance framework, enhancing standards for data security and personal information protection, exploring traceable management systems for AI services, and promoting international collaboration on AI governance. Despite these strides, the LLM-driven transformation presents two structural challenges:

Imbalance Between AI Computing Supply and Demand

The widespread deployment of LLMs is severely constrained by computing power bottlenecks. As LLMs scale from hundreds of billions to trillions of parameters, their inference processes demand vast compute resources due to the intensive matrix operations required. This imbalance is especially acute in edge computing scenarios, where edge devices are inherently limited by power consumption and physical size, making it difficult to support real-time inference for models with tens of billions of parameters. While cloud data centers offer substantial computing power, traditional virtualization architectures often fall short of meeting low-latency requirements. Compounding the issue is the mismatch of computing resources with security protection. Large-scale AI computing environments frequently lack efficient security protection and trusted execution assurance, making it difficult to deploy AI systems in scenarios with stringent security and privacy requirements. This further widens the gap between AI computing supply and demand.

Security Gap in Data Flows

As industrial intelligence accelerates, the demand for high-quality data flows has become increasingly urgent. Yet, current infrastructure remains inadequate in ensuring the trusted, secure flows of data assets. In sensitive sectors such as finance and healthcare, different organizations have their own data silos. For instance, banking risk control models require access to cross-institutional transaction data, but are constrained by financial compliance regulations. Medical AI development depends on

multi-center clinical data, yet faces significant hurdles in protecting sensitive patient information. Traditional solutions such as federated learning and homomorphic encryption are used to address the dilemma, but they struggle to balance trade-offs among model performance, communication overheads, and E2E controllability. These limitations underscore a fundamental issue: the absence of trust mechanisms in cross-entity data exchanges. Data providers fear losing control over data sovereignty and the value of data assets. Application and model developers are concerned about intellectual property rights of core algorithms. The lack of mutual trust makes it difficult to streamline the data value chain and impedes the development of domain-specific LLMs and data value creation across different organizations.

At the core of today's challenges lies a structural mismatch between traditional computing architectures and the emerging AI paradigm. On one end, edge devices are physically constrained, making them ill-equipped to support complex models. On the other, centralized processing on the cloud introduces privacy risks. The evolution of computing has progressed through two distinct paradigms: The first generation was represented by device-side, single-node computing, and the second generation featured distributed computing on the cloud. To meet the demands of LLMs, the industry is now advancing toward a third-generation computing paradigm—device-cloud synergy. This architecture integrates device-level security with high-security computing infrastructure on the cloud to unleash AI potential while ensuring data sovereignty. This paradigm shift is not merely a technological upgrade, but also a key breakthrough for promoting large-scale industrial intelligence. The key technological foundations of the third-generation computing paradigm are high-security computing and storage environments and computing infrastructure with high-performance link security.

1.2 Cloud AI Privacy Protection Dilemma

While traditional cloud AI services offer vast computing power, they face severe privacy challenges due to architecture limitations. In AI inference scenarios, cloud services must access unencrypted user request data to perform complex model computation. This makes it increasingly difficult for conventional cloud architectures to meet increasingly strict privacy protection requirements. The core dilemma manifests in three interrelated dimensions: lack of verifiable commitments, insufficient runtime transparency, and uncontrollable risks of privileged access.

Traditional cloud services lack reliable technical mechanisms to verify the implementation of privacy commitments. Although cloud service providers may pledge not to record specific user data, there is no effective technical mechanism for security researchers to independently verify whether these commitments are consistently upheld. As a result, users have to trust the cloud service providers while lacking technical assurance.

Another privacy risk arises from the opacity of runtime environments. Cloud AI services often do not offer transparent or verifiable runtime environments, nor do they support remote attestation. Without these safeguards, it is difficult for users to determine whether the cloud software stack is operating in a trusted environment or promptly verify whether the stack has been maliciously altered. For instance, a model inference engine could be tampered with to include a malicious version that collects user data.

The third challenge is the systemic risks stemming from privileged access mechanisms. The operations and maintenance (O&M) of cloud AI services inevitably requires administrator operations. These high-privilege roles typically rely on remote

interfaces—such as SSH—for fault diagnosis and system maintenance. Although access control policies are in place, it remains difficult to enforce effective controls under real-world O&M pressures. The privileged interfaces often become primary targets for ill-intentioned actors, with ransomware groups frequently attempting to steal administrator credentials to gain access to sensitive data. Service providers are unable to establish verifiable technical safeguards to eliminate the risk resulting from privilege abuse while maintaining necessary O&M flexibility.

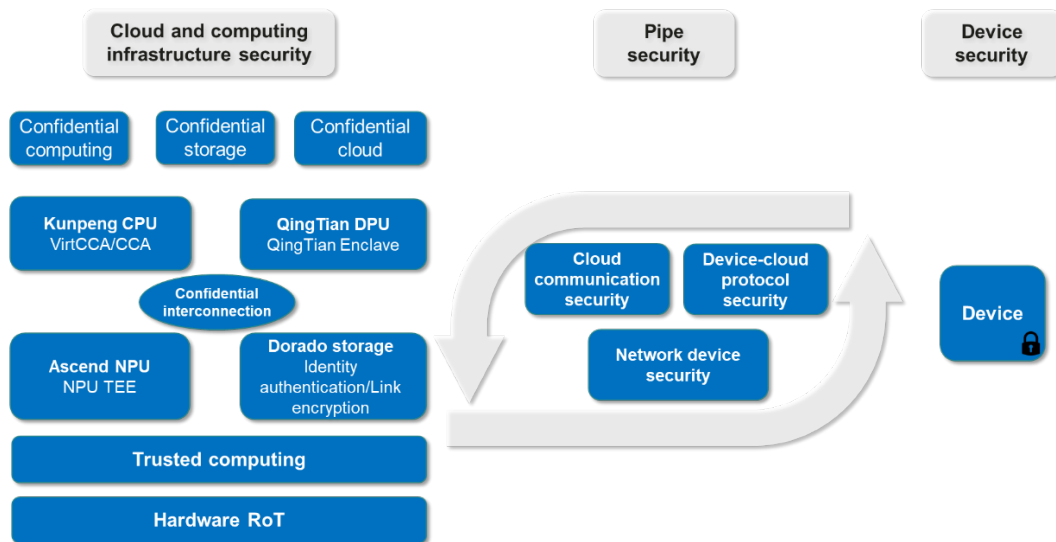
These dilemmas expose the inherent limitations of traditional computing infrastructure. When handling computing-intensive tasks—such as LLM inference, fine-tuning, and training—conventional privacy-preserving computation cannot efficiently process ciphertexts. As a result, unencrypted user data has to be directly accessed and processed. In the absence of hardware RoT, TEEs, comprehensive remote attestation mechanisms, and other hardware-based features provided by computing infrastructure, it becomes technically infeasible to uphold privacy protection commitments. Once user data leaves local devices or enters the cloud from local devices, it moves beyond the user's control and into a computing environment that lacks auditability, verifiability, and clearly defined permission boundaries. To address the structural weakness, the industry is seeking a fundamental solution—extension of device-level security models to computing centers by constructing a robust computing infrastructure security technology system. This is to implement new computing infrastructure with privacy protection and ensure secure and trusted data flows throughout the data lifecycle, enabling efficient utilization of computing infrastructure.

2 HCIST Architecture and Key Technologies

Huawei Computing Infrastructure Security Technical White Paper (HCIST) systematically introduces the key technologies developed by Huawei's business units (BUs) and research departments to address computing infrastructure security. It presents a hierarchical framework of E2E data protection, with core technologies encompassing computing, cloud, storage, communication, and other service scenarios. Huawei develops computing infrastructure security capabilities for different service scenarios, provides hardware-level trust assurance for LLMs, data privacy, and other scenarios, and establishes robust computing infrastructure to address major concerns of data processing in the AI era.

HCIST implements device-pipe-cloud synergy and provides security protection throughout the data lifecycle from data generation and transmission to processing, thereby building an in-depth defense system. HCIST supports integrated deployment and layered decoupling (for on-demand combination). This flexible design enables comprehensive protection across diverse scenarios, allows for single-point technology integration into solutions of other vendors, and supports security solutions tailored to specific industries and application needs. The device-pipe-cloud synergy mechanism enables cross-layer technical integration and is designed to align with open ecosystem policies. It supports flexible computing infrastructure architectures, enabling efficient interworking with devices from different vendors to meet different customer security requirements. The synergy architecture enhances privacy protection in cross-domain scenarios and provides dynamic, verifiable, and elastically scalable secure execution environments for AI computing infrastructure, becoming a key foundation for trusted running of large-scale AI services. Figure 2-1 shows the HCIST architecture.

Figure 2-1 HCIST architecture—Decoupling of user interface modules in devices, pipes, clouds, software, hardware, chips and ecosystem compatibility



On the device side, HCIST enhances native trust mechanisms at the hardware level. By establishing TEEs, it ensures that sensitive data—such as biometric information and payment credentials—is always processed within physically isolated TEEs, effectively preventing unauthorized access and data leakage. Additionally, through a distributed device authentication mechanism, HCIST moves beyond the traditional default trust model by enabling dynamic trust negotiation between devices, significantly boosting anti-attack capabilities in multi-device synergy scenarios.

For data transmission, HCIST adopts a zero-trust architecture to reinforce the communication pipeline in an E2E manner. Its mutual anonymity communication mechanism isolates directly identifiable links between devices and the cloud, eliminating attack vectors such as traffic analysis and reverse IP lookup.

On the computing platform side, HCIST builds security protection across chips, firmware, software, single nodes, multi-node systems, clusters, and cloud environments, delivering E2E data protection. It strictly adheres to the principle of "data available but invisible" to safeguard user data and model assets throughout their lifecycle while supporting high-performance AI inference, fine-tuning, and training for future AI scenarios such as agents.

HCIST establishes full-stack intrinsic security capabilities, spanning from chips to systems, and develops a dual protection system based on hardware confidential computing and trusted computing technologies. At the hardware level, the next-generation trusted computing architecture is adopted. Through built-in or external hardware RoT, HCIST enables secure boot and trusted boot, ensuring that only authorized and cryptographically verified code can be executed. At the runtime level, the heterogeneous confidential computing environment on both host and device sides is enabled. This setup not only blocks privileged access on the host side but also prevents unauthorized access to device data by host administrators or ill-intentioned users. Additionally, HCIST fully leverages confidential storage to safeguard user data and intermediate inference data, enhancing AI service efficiency, protecting data privacy, and preventing risks associated with intermediate data leakage. By extending the confidential computing environment from CPUs to heterogeneous computing resources, HCIST significantly strengthens E2E security protection across the

computing infrastructure—marking a key feature of its architecture. HCIST is designed to be scalable and composable, offering flexibility to interwork with diverse security products and technologies from various vendors, series, and domains—including computing, storage, communication, and devices. This adaptability allows it to meet specific security requirements and objectives, while effectively supporting the computing infrastructure ecosystem and establishing a unified, flexible, and secure computing foundation.

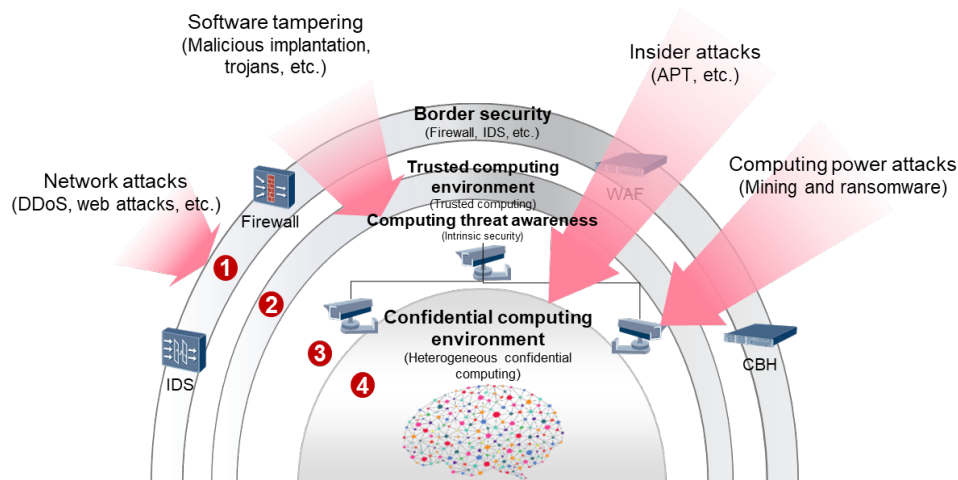
3

Computing Infrastructure Security in Computing Scenarios

This chapter focuses on computing infrastructure security in computing scenarios, primarily the security capabilities of Kunpeng CPU and Ascend NPU computing infrastructure. These capabilities can either be a part of the HCIST unified technical architecture, or be flexibly decoupled and combined as required by diverse computing environments to implement security solutions for heterogeneous platforms. These security solutions and their combinations allow Huawei computing infrastructure security system to defend against both traditional network risks and emerging AI security threats in computing scenarios.

Figure 3-1 illustrates the architecture of Huawei computing infrastructure security for computing scenarios. It includes four layers for comprehensive protection of computing security in the AI era. The first layer is traditional border security defense, which consists of products like firewalls and IDSs to block common network attacks such as DDoS and web attacks. The second layer is trusted computing. Both Kunpeng CPUs and Ascend NPUs support integrity measurement to prevent software tampering, malicious implantation, and trojans. The third layer is hardware-based computing threat awareness, which detects emerging threats to computing infrastructure and data, such as mining and ransomware. The fourth layer is confidential computing. It focuses on hardware security capabilities such as chip confidential computing and heterogeneous secure interconnection, paired with cryptographic computing technologies like federated learning and homomorphic/multi-party secure computation, to create a highly secure and trusted computing environment for E2E data protection.

Figure 3-1 Computing infrastructure security architecture



The next sections will explain the key security capabilities of Kunpeng CPUs and Ascend NPUs, along with the heterogeneous security solution that implements secure interconnection between them.

3.1 Kunpeng Computing Infrastructure Security

CPU Confidential Computing

Kunpeng servers powered by Arm CPUs are widely used in data centers, cloud computing, and edge environments. Arm is widely recognized for its high performance and low power consumption. Kunpeng, a server-level processor platform, has been deeply optimized in terms of performance and energy efficiency, making it suitable for large-scale computing deployment.

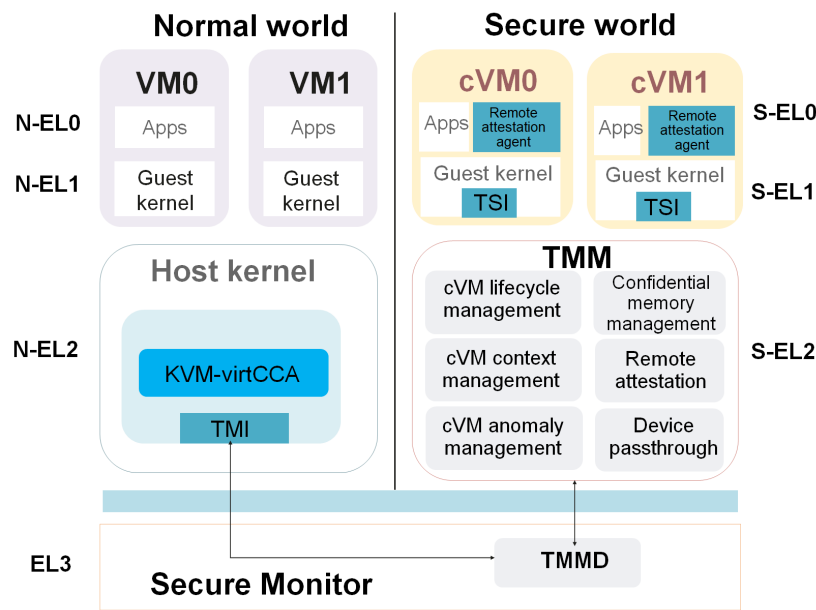
As data security becomes a core challenge in major computing scenarios, especially in multi-tenant cloud environments, traditional OSs and virtualization technologies are inadequate in ensuring data privacy. To address this issue, Arm introduced the Confidential Compute Architecture (CCA), a hardware-based trusted execution model aiming to create isolated execution spaces called Realms for sensitive data and computing workloads.

Arm CCA is designed to provide execution environments independent of the OS, hypervisor, and firmware, ensuring the integrity and confidentiality of workloads running in Realms. The CCA defines three logical worlds: normal world, secure world, and Realm world. The Realm world has independent resource access paths and encrypted memory. It implements strong security boundary isolation by using hardware-based access control and encryption/decryption engines, making it suitable for hosting high-value or privacy-constrained applications.

However, existing Arm server platforms in the market lack full CCA hardware support. In this context, the virtCCA solution is developed to implement the main CCA security features on current Arm platforms while maintaining CCA-compatible interfaces. virtCCA allows users to experience CCA-based confidential computing capabilities before deploying new hardware. Figure 3-2 shows the virtCCA framework.

virtCCA relies on Arm's TrustZone technology and fully utilizes the Secure EL2 (S-EL2) feature available in Armv8.4 and later releases. virtCCA creates a lightweight security monitoring component named TrustZone Management Monitor (TMM) to manage the lifecycle, memory isolation, and interface calls of confidential virtual machines (cVMs). Hypervisors and common OSs are considered potentially untrustworthy and prohibited from accessing or managing cVM status and resources. Performance evaluations show that virtCCA maintains much lower overhead for both I/O-intensive and compute-intensive workloads than traditional TEE solutions in virtualized environments. As cVMs are compatible with common OS images, enterprises can migrate critical workloads to cVMs without modifying applications.

Figure 3-2 virtCCA framework



virtCCA, as a pioneer of Arm CCA, lays the groundwork for building secure and trustworthy cloud infrastructure over the Kunpeng platform. As Arm's CCA hardware support matures, Kunpeng will launch native CCA hardware design supporting Realms. Users can choose between virtCCA or CCA as required by underlying hardware to enable E2E confidential computing from chips to clouds.

3.2 Ascend Computing Infrastructure Security

Ascend NPUs meet AI computing demands with high-bandwidth memory, high-throughput AI Core design, and full software stack support. They are extensively used in cloud and data center environments. As data privacy protection and intellectual property rights become increasingly important for AI applications, building trusted AI execution environments has become a priority in NPU design. In particular, in multi-tenant and mixed-trust cloud environments, ensuring model and data security when using AI accelerators is crucial for establishing trustworthy AI infrastructure.

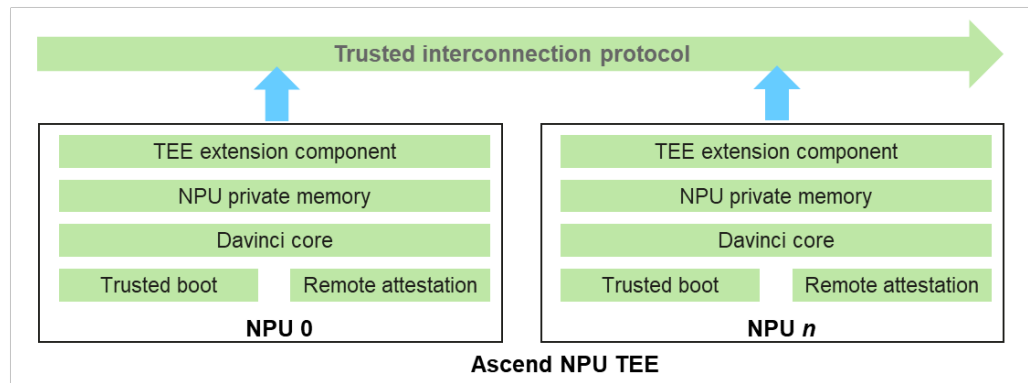
NPU Trusted Computing

Ascend NPUs and BMC chips include built-in hardware RoTs to enable security capabilities such as secure boot and DICE. These RoTs establish system trust by ensuring trusted execution of key operations (such as firmware verification and key generation) during system boot and running. They protect against both physical attacks (such as chip tampering) and logical threats (such as malicious code injection) to create a trusted base for upper-layer security measures. BSBC boot firmware, BaseBIOS, HSM runtime firmware, and iBMC software have passed the CC EAL4/5+ security certification, ensuring solid computing security for customers. Moreover, Ascend servers are compatible with standard TPMs, providing TPM-based trusted boot for customer services. Ascend also supports NPU application measurement and remote attestation to continuously measure key files, mandatory access control policies, and NPU running status. This prevents unauthorized modifications to key NPU software and ensures NPU software stack integrity and application security.

NPU Confidential Computing

To create TEEs for real-world applications, Ascend introduces NPU TEE, a full-stack confidential computing system using NPUs as the root and spanning from hardware to runtime. This system sets NPUs as the trust boundaries. Host systems in the untrusted domain, including the host OS, drivers, AI runtime framework, and peripheral interfaces, cannot access sensitive user data in the NPU TEE. This ensures data isolation and integrity verification are completed on the NPUs. Figure 3-3 shows the Ascend confidential computing architecture.

Figure 3-3 Ascend NPU TEE architecture



Ascend employs a TEE extension component to establish the NPU trust boundaries. This component connects NPUs to the trusted interconnection bus and isolates NPU memory. This prevents malicious programs and even administrators in the REE from reading the model and user inference data in the NPU TEE private memory. During inference, the NPU private memory stores the user inputs, intermediate computation results, model parameters, KV caches, and operator workspace for each task, and the TEE extension component blocks REE access to such data. After completing a user's computing task, the NPU clears and releases its memory space before switching to another user, preventing data residues or leaks when users share computing resources. This is crucial for LLM inference in mitigating memory residue risks, particularly when processing massive data with user features, such as KV caches, attention weights, and intermediate representations. In addition, Ascend NPU TEE uses the NPU RoT for trusted boot and remote attestation of the systems and software

running on the NPU, and generates a measurement report containing the NPU identity and security status. This prevents malicious tampering on the software stack within the trusted domain, ensuring computing infrastructure integrity and platform authenticity.

These protection mechanisms of the Ascend platform are compatible with mainstream AI frameworks (such as PyTorch and vLLM), without the need to modify the model structure or service logic. This facilitates trusted AI deployment, allowing model developers and data providers to integrate confidential computing into existing processes without participating in underlying security designs or calls.

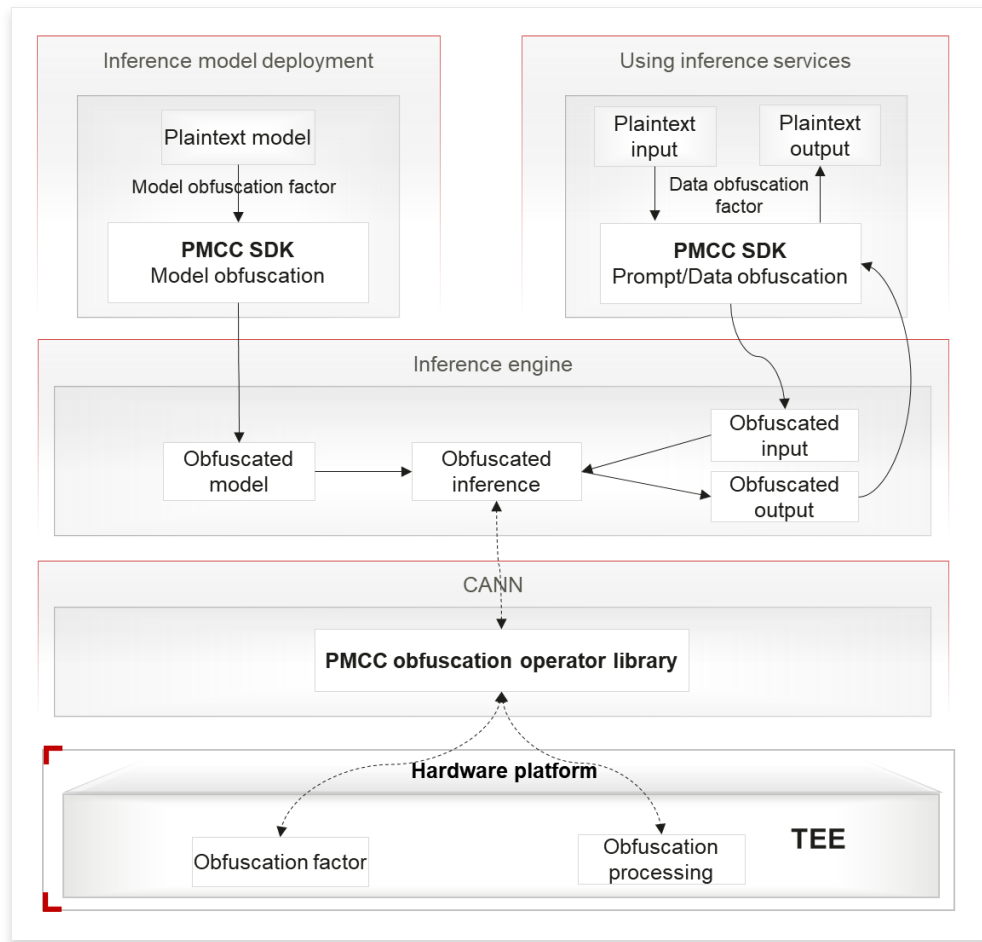
In conclusion, Ascend creates a confidential computing execution environment for NPUs through a comprehensive system of the TEE extension component, trusted boot, and remote attestation. This system redefines the security boundaries for AI inference and sets up a trusted hardware base for emerging applications like generative AI and AI as a service (AlaaS). Ascend confidential computing will be a critical support for the secure and reliable operation of LLMs as they integrate into general infrastructure.

Privacy and Model Confidential Computing (PMCC)

Ascend offers Privacy and Model Confidential Computing (PMCC) to further enhance runtime privacy and model asset protection. As illustrated in Figure 3-4, PMCC integrates the obfuscation technology into the E2E inference process. The technology sets model weights, inference inputs, and outputs to the obfuscated state, making them available but invisible over computing paths. This offers fine-grained, cost-effective, and easy-to-integrate protection for sensitive data without changing the service logic and model structure. PMCC is always effective throughout the execution process. Even with high-level system permission, attackers can only access the transformed intermediate data and weight representations but cannot restore the plaintext.

PMCC is implemented based on obfuscation operators, obfuscation factors, and hardware binding. Lightweight obfuscation operators are inserted into inference graphs and securely executed in the Ascend NPU confidential computing environment. Obfuscation factors are generated, stored, and called in this secure area and bound to the specified NPU. The obfuscated model can be restored to the executable format only on the hardware with the consistent factors, ensuring strong protection by binding the model to the hardware. During operation, inference inputs are obfuscated before entering the execution path, and outputs are de-obfuscated before leaving. Obfuscation applies across CPU memory, NPU memory, card interconnections, and bus transmission layers, minimizing plaintext exposure. Measurement and audit mechanisms are available to verify and trace the obfuscation process and execution environment status for compliance purposes.

Figure 3-4 Ascend PMCC technical architecture



PMCC supports both client obfuscation and server obfuscation modes to meet various boundary and deployment requirements. In client obfuscation mode, clients obfuscate requests before submission and de-obfuscate results upon reception. Servers have no access to plaintext data throughout the process, ensuring high security. In server obfuscation mode, servers obfuscate and de-obfuscate data, minimizing changes to existing clients for quick implementation. Model weights remain in the obfuscated state during operation in both modes. In server mode, PMCC can collaborate with the hosts' trustworthy and confidential features to further reduce exposure in the running environments.

PMCC protects security and maintainability for both models and data. For models, PMCC implements continuous obfuscation and device binding to prevent weight leaks and asset generalization. For data, it obfuscates inference requests and results to hide plaintext of sensitive content from execution channels. Obfuscation factors are manageable throughout their lifecycle (registration, rotation, and revocation). Policies can be delivered and validated by task, session, or tenant, and combined with minimal audit records to establish controllable, verifiable, and revocable runtime protection.

The collaborative optimization of lightweight obfuscation operators and acceleration paths helps minimize PMCC's performance overhead: protecting only inference data causes less than 1% E2E performance loss, while protecting only model weights or both models and data results in around 5% E2E loss. For model providers, PMCC

implements continuous weight obfuscation and device binding; for users, PMCC protects inference requests and results from plaintext exposure during transmission and execution. Overall, PMCC works with NPU confidential computing to extend the protection boundary to each inference interaction at runtime, securing both model assets and user privacy. This provides solid support for Ascend to be stably implemented in more sensitive and diverse service scenarios.

NPU Cluster Access Control

The next-generation Ascend platform provides a mandatory access control policy to effectively isolate cluster management processes and ports for enhanced security. This policy restricts cluster management processes to accessing only predefined objects and limits management ports to sole use for communication between cluster management processes and external systems. Access control is achieved using a white list, which blocks malicious programs from accessing models, data, and operators and prevents internal processes from leaking information through management ports. In addition, to prevent user-defined operator extensions from causing NPU function misuse and damage to other operators, the platform enhances the mandatory access control on the user-defined operator processes and driver interfaces opened by NPU character devices to the operators. This prohibits the user-defined operator processes from accessing risky driver interfaces that are prone to DDoS attacks or data leaks.

3.3 Heterogeneous Device Confidential Passthrough with PCIPC

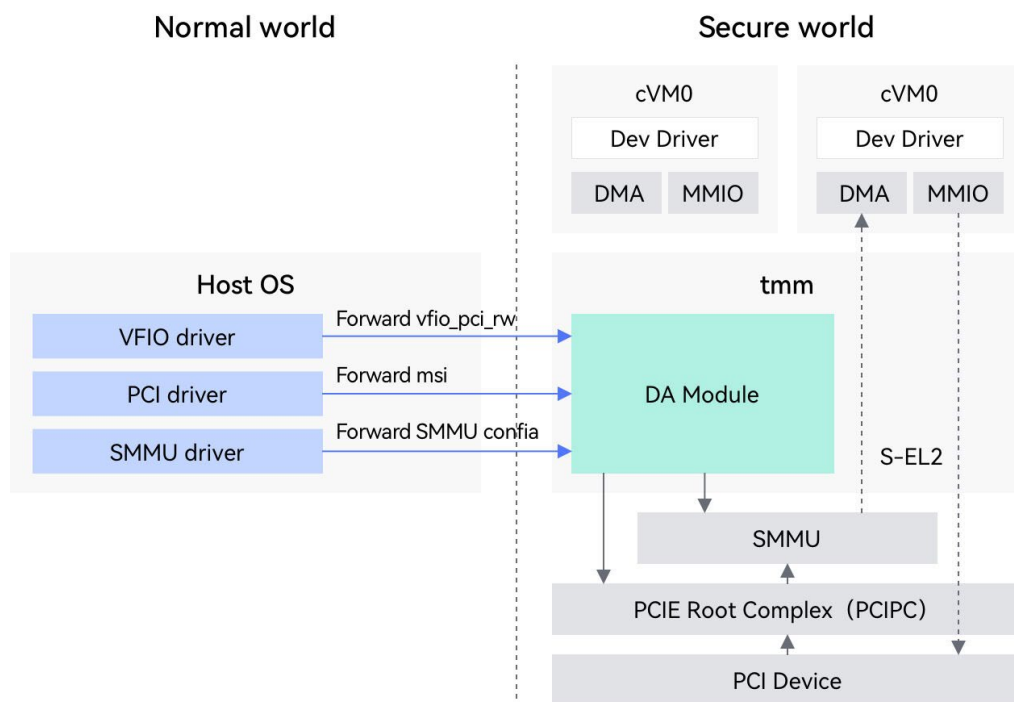
Traditional general-purpose confidential computing systems are CPU-centric, making them hard to reach complex heterogeneous device links. This poses a major challenge to building heterogeneous TEEs: how to directly connect devices like NPUs and GPUs to cVMs to gain near-local performance while ensuring robust security. Major industry players have proposed heterogeneous confidential computing solutions to address this pain point. For example, the TEE Device Interface Security Protocol (TDISP) uses the PCIe bus as a confidential communication channel and adds security bits to the bus for device-level identification and isolation. Such solutions generally rely on the collaborative upgrades in next-generation chips or protocol stacks. Currently, their large-scale adoption is limited by hardware time sequence issues, slow updates to bus standards, and immature ecosystem.

To ensure feasibility, Huawei introduces the Confidential Device Assignment (CoDA) framework based on existing chip capabilities. CoDA enables heterogeneous device passthrough to cVMs even if the device does not support confidential computing. This solution leverages Arm's native System Memory Management Unit (SMMU) and Huawei-developed PCI Protection Control (PCIPC) module to establish a controllable and secure device communication link on the CPU side, allowing heterogeneous device access in confidential computing semantics. Figure 3-5 illustrates the confidential device passthrough architecture implemented based on the PCIPC module.

CoDA aims to encrypt and isolate peripheral communication paths from the CPU side without modifying devices or introducing new bus standards. In specific implementation, the Kunpeng platform uses the chip's embedded PCIPC component to perform joint software and hardware control for the data flows with heterogeneous devices over the PCIe bus. CoDA collaborates with the SMMU to create an independent device address

mapping table in the Secure world. This design allows only cVMs to access device I/O resources and blocks host administrators or nVMs from snooping or interfering. Unlike TDISP, CoDA can be deployed on existing hardware and does not need heterogeneous devices to support security authentication or channel encryption/decryption, greatly simplifying deployment.

Figure 3-5 Confidential device passthrough based on the PCIPC module



In terms of software stack, CoDA fully reuses the mature VFIO, SMMU, and PCIe driver capabilities of the Linux community. A device can access cVMs as long as it supports the VFIO paravirtualization mechanism. This means that PCIe devices such as GPUs, NPUs, NVMe SSDs, and SmartNICs can be transparently and securely passed through to cVMs without changing the driver or framework code. Applications like the inference framework, data service, and edge push can seamlessly migrate their services to confidential environments. CoDA has completed performance testing and verification on multiple types of devices. Tests with NVMe storage show that cVMs in passthrough mode deliver the same sequential and random access performance as common VMs, and 2x higher performance than paravirtualization solutions like VirtIO.

The early PCIPC design performs isolation by root port, which cannot distinguish access boundaries between virtual functions (VFs) and physical functions (PFs) on the same device. Considering that SR-IOV devices need to support cVMs in real-world applications to meet multi-tenant and high-performance cloud computing requirements, the updated CoDA solution uses the "driver-based forwarding" mechanism to enable the SR-IOV feature. When an SR-IOV device moves to the Secure world, the device driver on the host outside the Secure world can still configure VFs through PFs. PF DMA uses the non-secure page table of the SMMU, while VF DMA uses the secure page table of the SMMU, achieving isolation between the VF and PF data planes.

In summary, CoDA has successfully established a complete heterogeneous confidential computing link, and for the first time, implemented device isolation by SR-IOV VF on Arm platforms, driving confidential computing evolution from static, closed architectures to elastic, dynamic solutions. CoDA builds a secure, universal, and high-performance device access channel. This is achieved by combining SMMU and PCIPC control capabilities, leveraging the device management engine in TMM/RMM, and maintaining compatibility with VFIO and existing drivers. It realizes heterogeneous confidential computing on Kunpeng platforms and provides a solid computing base for services with stringent security demands, such as LLMs, image inference, and privacy-preserving computation. As hardware and software evolve, CoDA will continue to expand its adaptability and establish itself as a key technology for scheduling heterogeneous resources and enforcing data isolation within Kunpeng trusted computing platforms.

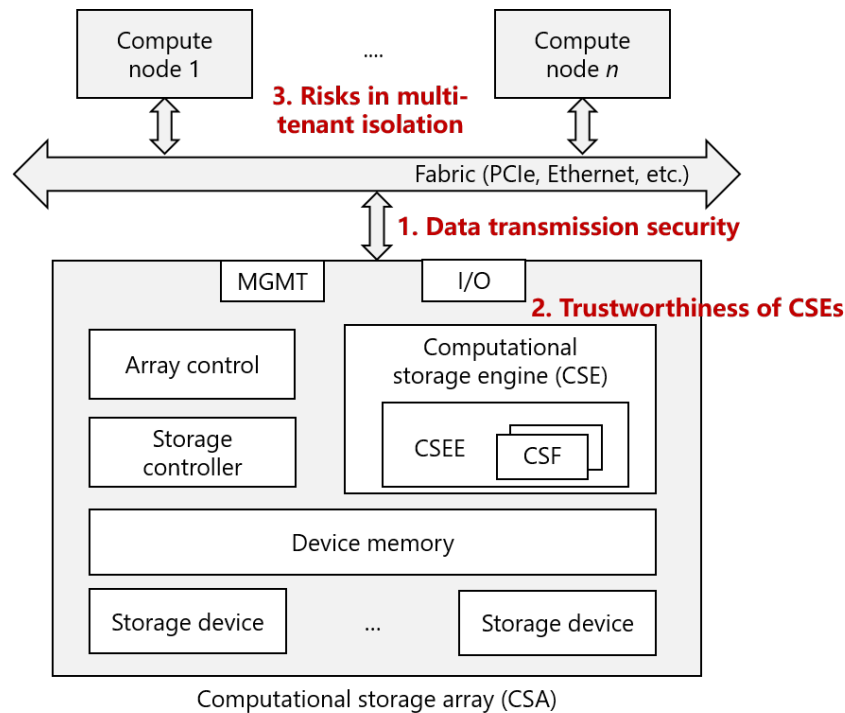
4 Computing Infrastructure Security in Storage Scenarios

4.1 Security Challenges of Storage Architecture

As the service complexity increases sharply, the data scale far exceeds the capacity limit of a single system, and the memory specifications cannot meet the growing processing requirements. Against this backdrop, service systems inevitably evolve from the traditional "storage-compute integration" architecture to "storage-compute decoupled". Security capabilities must adapt to this new computation architecture and support modular combination to meet differentiated requirements of services for compute resource binding and isolation policies. In addition, the computing power gradually extends to storage. Besides storing data, storage systems also have certain data processing and analysis capabilities and start to change from "passive storage" to "intelligent storage". This reduces frequent data migration between compute and storage devices and improves system processing efficiency.

The storage architecture keeps evolving as services develop from databases to large-scale AI applications. According to the industry consensus, storage is evolving towards an in-depth integration of "storage-compute decoupled" and "intelligent storage". This new architecture not only features high scalability and flexibility, but also implements efficient computing near the data source, laying a solid foundation for coping with the explosive growth of data-driven services.

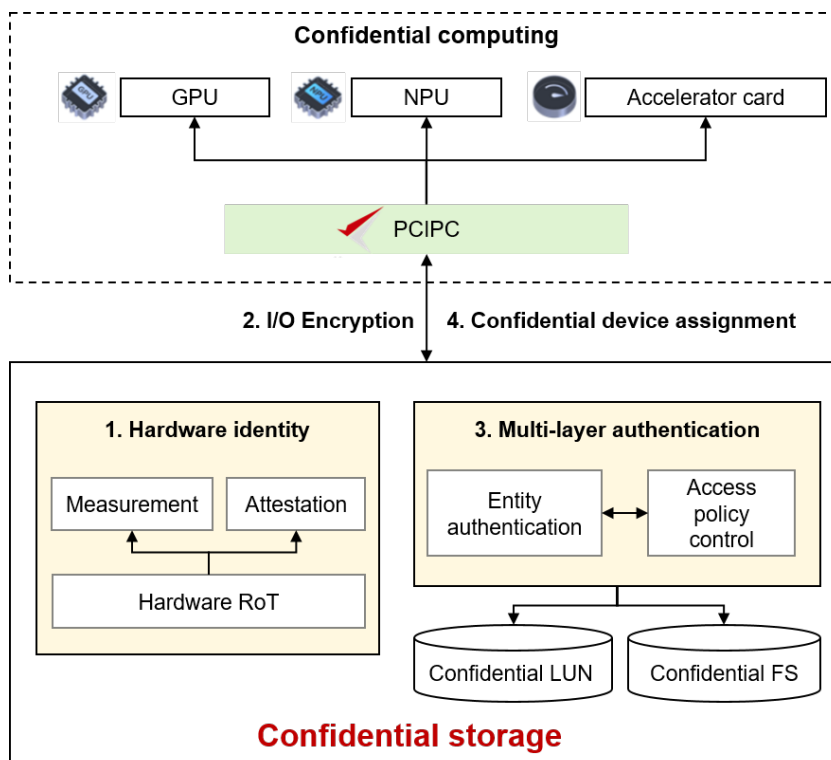
Figure 4-1 Security challenges of "storage-compute decoupled" and "intelligent storage"



Unlike traditional storage where data is processed within storage devices, storage-compute decoupled and intelligent storage face frequent data flows between compute nodes, storage nodes, and networks. This significantly expands potential attack surfaces, rendering traditional computing-centered security models obsolete. As shown in Figure 4-1, potential attack surfaces lie in the following aspects:

- **Data transmission security:** In the storage-compute decoupled architecture, data needs to be frequently transmitted across nodes and networks. Therefore, transmission links become potential attack channels. Eavesdropping, tampering, or man-in-the-middle attacks on transmission links threaten data integrity and confidentiality.
- **Trustworthiness of computational storage engines (CSEs):** As storage nodes also provide computing functions, CSEs may become the target of attacks. If there is no effective trusted execution and verification mechanism, the computing process transparency and the reliability of computing results cannot be ensured, facing data tampering and forgery risks.
- **Risks in multi-tenant isolation:** In cloud and shared resource environments, storage and computing resources of different tenants may not be well isolated. If the identity authentication and authorization mechanisms are weak, data leakage, unauthorized access, and even horizontal attacks may occur to tenants, directly breaking the system security boundary.

Figure 4-2 Combined confidential domain of computing and storage



To cope with the security challenges brought by storage architecture evolution, HCIST proposes the concept of confidential storage as well as corresponding solutions. Confidential storage ensures trustworthiness of data movement across trusted domains by enhancing security at the hardware layer. Specifically, a storage-compute confidential domain covering multiple storage protocols is constructed using technologies such as hardware identity, link encryption, dual-layer authentication, and data passthrough. Figure 4-2 shows the framework of this confidential domain. The collaborative protection system of compute nodes, storage nodes, and networks can protect the security of stored, in-transmission, and in-use data. It also improves the transparency and trustworthiness of the entire system. Besides independently implementing high-security storage management, confidential storage can work with confidential computing and trusted computing technologies on the compute side. This way, flexible combination solutions can be formed based on different security objectives and risk considerations, achieving differentiated security hardening paths.

4.2 Storage Node Identity

Similar to confidential computing, confidential storage takes the hardware RoT as the cornerstone. Storage nodes use the hardware RoT (such as TPM) to ensure the uniqueness and non-forgery of a storage node identity. In addition, measured boot, remote attestation, and mandatory access control can be used to ensure the trustworthiness of storage nodes throughout the entire lifecycle.

As the anchor of the entire trust chain, the hardware RoT performs integrity measurement on the system firmware, OS, driver modules, applications, and security policies during the storage node startup. The hardware key signs the measurement

results to generate an undeniable measurement report. Tenants can perform remote attestation based on the report to check whether the identity and status of a storage node meet the expectation. The expected status includes: the storage node loads the expected software components, the CSE performs computing tasks based on the expected logic, and mandatory access control restricts application behavior based on the specified policy.

In the running phase, storage nodes ensure permission separation and multi-tenant isolation based on multiple security mechanisms, including mandatory access control. Isolation measures cover multiple dimensions, such as management isolation, network isolation, and service isolation. The execution policies of the preceding security mechanisms are continuously measured and verified by the hardware RoT to build a trust system with software-hardware synergy and implement security assurance for the entire lifecycle of storage nodes.

4.3 Storage Link Encryption

In a storage-compute decoupled architecture with multiple protocols, link encryption protects data transmission security across nodes and networks and therefore is an important security mechanism of confidential storage. The core objective of link encryption is to ensure confidentiality, integrity, and tamper resistance during transmission. Confidential storage relies on Transport Layer Security (TLS) and Internet Protocol Security (IPsec) to achieve this objective.

- TLS: provides encryption and authentication at the transport layer. The core process includes negotiating the encryption algorithms supported by two ends, performing key exchange and identity authentication, and using symmetric encryption algorithms to encrypt data and authenticate messages. Block, file, and object storage can work with TLS to effectively defend against link eavesdropping, tampering, and replay attacks, ensuring secure communication between storage nodes and tenants.
- IPsec: provides encryption and authentication at the network layer. It uses the Internet Key Exchange (IKE) protocol to complete key exchange and security policy negotiation, and establishes security associations (SAs) between nodes. IPsec can provide secure channels for multiple storage protocols and implement consistent protection across protocols.

By supporting TLS and IPsec on storage nodes, confidential storage not only provides consistent link security assurance for multi-protocol environments (see Table 4-1), but also reduces intrusive reconstruction of upper-layer applications. In addition, confidential storage significantly improves the performance of link encryption through instruction set optimization and hardware offloading.

Table 4-1 Storage protocols with link encryption enabled

Storage Type	Storage Protocol	Storage Protocol with TLS Enabled	Storage Protocol with IPsec Enabled
Block storage	NVMe-oF	NVMe over TLS	NVMe over IPsec
	iSCSI	iSCSI over TLS	iSCSI over IPsec
File storage	NFS	NFS over TLS	NFS over IPsec
Object storage	S3	S3 over HTTPS	S3 over IPsec

4.4 Dual-Layer Authentication

Due to the increasing popularity of the storage-compute decoupled architecture with multi-tenancy, security threats are becoming more and more complex. Most early attacks forge identities to gain unauthorized access and can be defended against by adopting reliable identity authentication mechanisms between communication parties. However, deeper risks lie in the hijacking of the operating environment and system status. Even if a node has a valid certificate, its firmware, OS, or security policy may have been tampered with, resulting in the risk of "trustworthy identity but untrustworthy environment."

To cope with such threats, confidential storage proposes a progressive dual-layer authentication mechanism.

- The first layer focuses on identity trustworthiness and answers the question of "who are the communication parties".
- The second layer extends to environment trustworthiness and answers the question of "where are the communication parties and in what states are they running".

The first layer leverages the certificate authentication mechanism of TLS and IPsec. Communication parties use digital certificates to ensure the uniqueness and unforgeability of the storage node and tenant identities, effectively defending against spoofing attacks and unauthorized access.

The second layer relies on the remote attestation mechanism. In addition to identity verification, the operating environment and system status of the communication parties need to be measured and verified. For example, communication parties can exchange measurement reports generated by their respective hardware RoTs to check that they are all in the trusted environment and running on the expected firmware, OS, and security policies.

In addition, the authentication system supports unidirectional and bidirectional authentication modes. In scenarios with low security concerns, tenants verify the identities of storage nodes. In high-security scenarios, bidirectional authentication must be enabled to ensure that both storage nodes and tenants are in trusted states, thereby defending against high-level attacks.

4.5 Storage-Computing Data Passthrough

Data needs to be efficiently transmitted between computing nodes and storage nodes. However, in traditional virtualization, I/Os are forwarded by the host kernel and intermediate layer. This not only brings performance loss due to frequent copy and context switching, but also increases the risk of data leakage and tampering because the intermediate layer may be tampered with.

To solve this problem, confidential storage is combined with the compute CoDA framework, which is described in Section 3.3, to implement high-performance data transmission in the storage-compute confidential domain. This way, the network throughput and latency close to those in the bare metal environment can be achieved without compromising security. Specifically, CoDA uses the PF driver forwarding mechanism to support SR-IOV VF-based device passthrough, so that VF-level device interfaces can be invoked in VMs. CoDA features excellent performance, robust security, and high usability.

In addition, the confidential domain also applies between storage nodes. Confidential storage uses protocol encryption, array encryption, replication link encryption, as well as data passthrough and hardware offloading to implement secure backup between storage nodes and efficient access to backup data.

5 Computing Infrastructure Security in Huawei Cloud

Leveraging Kunpeng, Ascend, and large-scale AI clusters, Huawei Cloud provides powerful AI cloud services compatible with mainstream open-source LLM (including openPangu, DeepSeek, Llama, and Qwen) in multiple geographic regions. These AI services have been used in many sectors, such as city governance, smart finance, healthcare, and weather forecasting. Huawei Cloud integrates ecosystem capabilities and works with partners to develop scenario-specific solutions that meet specific customer requirements for security, compliance, and performance. The development and evolution of the Huawei Cloud computing infrastructure security technology is critical to implementing these solutions.

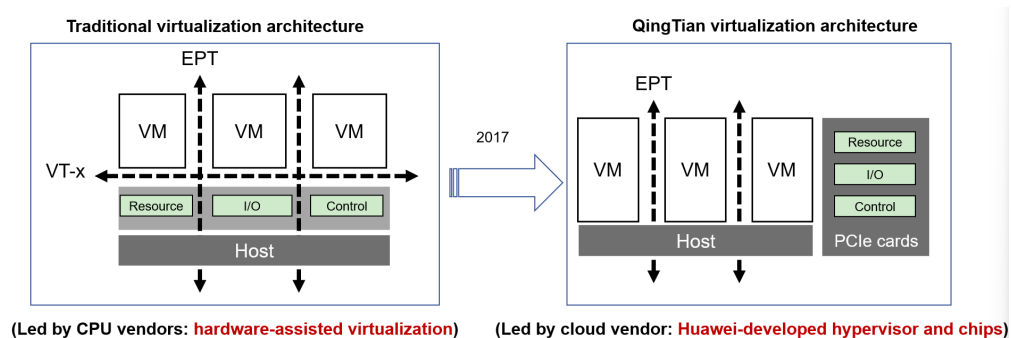
5.1 Huawei Cloud QingTian Architecture

Virtualization allows one physical server to run multiple OSs simultaneously, greatly improving resource utilization. In the traditional architecture, virtualization provides an independent hardware abstraction layer for VMs through a hypervisor and uses instruction translation and device simulation for VM isolation and security. Common computer instructions are directly executed on physical CPUs. Sensitive operations, such as operations on privilege control registers, are intercepted and simulated by the hypervisor to ensure system stability and isolation. In addition, device models are required by a virtualization system to simulate resources such as networks, storage devices, and input peripherals.

Although hardware-assisted virtualization technologies represented by KVM are mature, the traditional architecture still cannot meet the requirements of large-scale cloud platforms for zero resource wastes, low latency, and deterministic computing power because of resource reservation, computing power loss, service jitter, and other issues. To address these challenges, Huawei Cloud launched a next-generation QingTian virtualization system in 2017. This system reshaped the underlying virtualization architecture of Elastic Cloud Server (ECS) based on Huawei-developed dedicated hardware cards, hypervisor, and security controller. Hardware-software synergy helps achieve zero resource reservation, zero computing power loss, zero service jitter, and enhanced security isolation. After multiple rounds of iteration, the QingTian system has been fully applied to next-generation ECS instances. It provides computing services with high security, strong isolation, and high performance. QingTian has become the core platform for Huawei Cloud computing infrastructure.

QingTian is not only a virtualization solution, but also a key form of HCIST at the cloud virtualization layer. HCIST emphasizes that data is available but not visible. QingTian adheres to this principle through hardware-level isolation, encrypted execution, and zero trust. By deeply integrating with Kunpeng and Ascend processors, the QingTian architecture extends full-stack intrinsic security to the virtualization layer to provide a trusted computing environment for AI model training and inference, financial-grade applications, and cross-tenant data exchange. This architecture not only meets the requirements of cloud computing for high performance and low costs, but also lays a solid foundation for data security and compute performance in the AI era. Figure 5-1 shows the differences between the QingTian virtualization architecture and the traditional virtualization architecture.

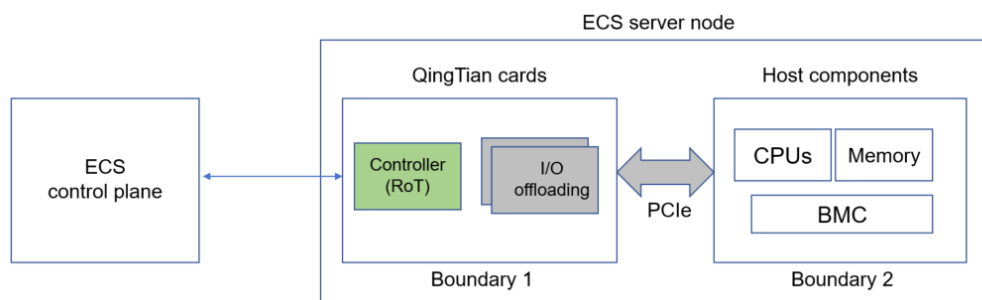
Figure 5-1 Differences between the QingTian virtualization architecture and the traditional virtualization architecture



5.2 Components of the Huawei QingTian Architecture

Huawei QingTian architecture consists of Huawei-developed QingTian cards, QingTian controllers, and QingTian hypervisor. They work together to create a high-performance, high-security, and strong-isolation cloud infrastructure. In the QingTian system, ECSs provide services relying on hosts and QingTian cards. Hosts process computing workloads, including ECS and Bare Metal Server (BMS) workloads, based on the QingTian hypervisor. QingTian cards operate independently from hosts. A QingTian card integrates control, acceleration, and security capabilities to provide solid hardware support and resource offloading for the entire architecture.

Figure 5-2 Huawei Cloud ECS server architecture



QingTian Card

QingTian cards are core hardware of the QingTian architecture. They are dedicated hardware independent from hosts. QingTian cards are connected to host CPUs via standard PCIe interfaces and provide control and I/O virtualization capabilities. They not only provide a hardware RoT for the entire system, but also integrate network, storage, and AI acceleration capabilities, and provide I/O interfaces of Virtual Private Cloud (VPC) networks, Elastic Volume Service (EVS) block storage, local disk storage, and other functions. QingTian cards have all control interfaces required by the ECS service. These interfaces are used for unified pre-configuration and management of CPU, memory, and storage resources of hosts. This way, storage and network capabilities can be offloaded from hosts. For management and control, the central node of a cluster only needs to interact with QingTian cards and does not need to interact with hosts. All VM lifecycle management commands are delivered to QingTian cards, which unidirectionally control hosts. This simplifies resource allocation and completely isolates O&M personnel from customer service data, greatly improving security. In addition, a driver simulates local and networked resources as local resources of hosts. With QingTian cards, users do not need to perform complex configuration. This further strengthens security isolation between cloud infrastructure and customer applications. A QingTian card is physically connected to a mainboard via PCIe and is isolated from host hardware. A card has Huawei-developed SPU chips, runs a simplified OS, and allows for independent live upgrade of the OS and virtualization components on the card, with no impacts on customer services. Dedicated ASIC hardware is used for virtualization of storage, network, and other capabilities, achieving higher performance at a low cost. Logically, there is one primary card (controller) and multiple secondary cards (used for I/O offloading). Figure 5-2 shows the architecture of an ECS instance with primary and secondary cards.

QingTian Controller

A QingTian controller is a QingTian card used for control. It is the hardware RoT of a server node. A controller manages all other components and loaded firmware of the system and protects firmware from being tampered with. After a system is powered on, the controller starts from the boot ROM, verifies and measures the local firmware, and performs composite authentication and integrity verification on the host based on the Device Identifier Composition Engine (DICE) mechanism, chip's embedded key, firmware measurement metrics, and key derivation algorithm. After the authentication is successful, the hypervisor image is decrypted and started. The chain of trust is extended from the controller to the entire host system. System firmware is encrypted and stored in the local solid-state drive (SSD) of a QingTian card. The encryption key is jointly protected by secure boot and the TPM device. The decryption key is only released in a trusted environment where the startup measurement value meets the expectation. The controller completes host authentication and trusted management through the challenge-response mechanism. Then, the host can receive and run customer workloads.

A QingTian controller is the only security gateway between physical hosts and the cloud control plane. It abstracts the cloud control plane as the ECS control plane for interaction. All traffic in and out of physical hosts must be forwarded through an encrypted API channel. A controller provides a hardware-based bidirectional authentication link for E2E encryption. API access control is based on the minimum permissions. All API operation logs are recorded to detect exceptions in real time. The

control plane is isolated from the data plane through a dedicated network to ensure system security.

I/O Offloading

For I/O offloading, QingTian provides hardware acceleration and encryption for VPC networks, EVS block storage, and local NVMe storage through secondary QingTian cards. These secondary cards share SoC and basic firmware design with the controller and are equipped with dedicated hardware to improve performance. The hardware offloading engine and secure key storage integrated in the SoC enable efficient encryption for network and storage data. Standard cryptographic algorithms can be used for hardware acceleration. Cryptographic keys are only stored in the protected volatile memory of QingTian cards in plaintext. Huawei Cloud O&M personnel and any customer code running in hosts cannot access the keys. So, the keys are highly secure. Key materials are distributed by multiple management components independently to avoid single point of failure in key distribution. Hourly key rotation can be used to adapt to the cloud computing SDN architecture and reduce performance overheads for key negotiation. This is helpful for large-scale encrypted communication.

If the QingTian hypervisor is used, QingTian cards virtualize I/O devices into multiple virtual functions (VFs) using SR-IOV and directly connect them to VMs. Service data (such as processing, storage, or hosting data) is directly transmitted between VMs and virtualized I/O devices, bypassing the hypervisor layer. Hardware-level data passthrough is achieved. This minimizes software and hardware dependencies in the I/O path, significantly improves security and I/O efficiency, and delivers performance close to that of physical servers with lower costs of computing power per unit.

QingTian Hypervisor

QingTian hypervisor is a lightweight VM management tool. It adopts a completely different design from the traditional type-2 architecture. QingTian hypervisor achieves ultra-high performance and ultra-low overheads through full-stack simplification and reconstruction. With QingTian hypervisor, VMs have excellent performance and resource isolation close to physical servers.

QingTian hypervisor is innovative in four aspects. First, it can offload all capabilities of the management plane and I/O data plane to QingTian cards. Only basic virtualization capabilities are retained. This way, server resources can be fully utilized by users. Second, NanoOS, a lightweight, stateless virtualization OS is used. NanoOS has only the necessary components. All kernel modules and software packages irrelevant to virtualization are removed from it to greatly reduce the trusted computing base (TCB). Third, Huawei-developed VRAM memory management system with an innovative pageless architecture is used to reduce the management overheads by dozens of times while maintaining the memory compatibility of VMs. Fourth, the secure, low-noise, fast, and modular runtime base properly meets the requirements of scenarios that are highly sensitive to startup speed and resource usage, such as serverless computing and AI agents.

In terms of security, QingTian hypervisor sets up a multi-layer defense system. The attack surface is greatly reduced by the extremely small TCB and network-free and storage-free design. Core affinity is used to allocate workloads of VMs to specific CPUs to avoid CPU scheduling and context switching overheads and defend against side-channel attacks. Hardware-assisted virtualization and Huawei-developed VRAM memory management are used for strong isolation of memory and I/Os between VMs. All logs and monitoring data are periodically uploaded to the cloud through APIs to prevent local data tampering. The runtime in-memory file system is configured to be

read-only and trusted audits are enabled to effectively prevent VM escape and external tampering. Software package upgrades undergo multiple layers of verification, including Cyclic Redundancy Check (CRC) and certificate validation, to ensure the transmission process is not tampered with. In addition, QingTian hypervisor is closely associated with QingTian controllers to ensure the entire process from startup, loading, to operating are trustworthy and secure.

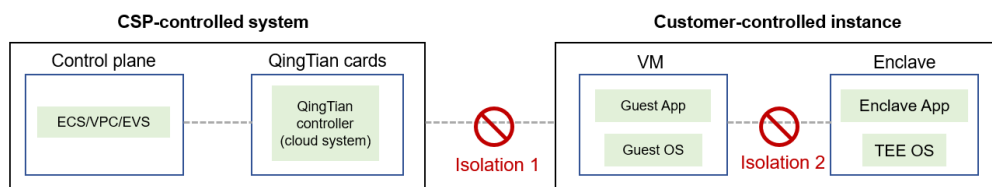
5.3 Huawei QingTian Confidential Computing

The QingTian confidential computing solution is developed based on the QingTian architecture. It aims to use QingTian cards and the QingTian hypervisor to protect customers' application data and code from external access during data processing. The protection includes:

1. **Confidentiality of customer data and code:** Ensure that customer data and code are not accessed by cloud service provider (CSP)'s internal personnel or cloud systems, and are not accessed by customers' internal personnel or VM administrators.
2. **Integrity of customer data and code:** Ensure that customer data and code are not tampered with by CSPs' internal personnel or cloud systems, and are not tampered with by customers' internal personnel or VM administrators.

To achieve these objectives, QingTian confidential computing provides security isolation in two dimensions (as shown in Figure 5-3).

Figure 5-3 QingTian isolation solution

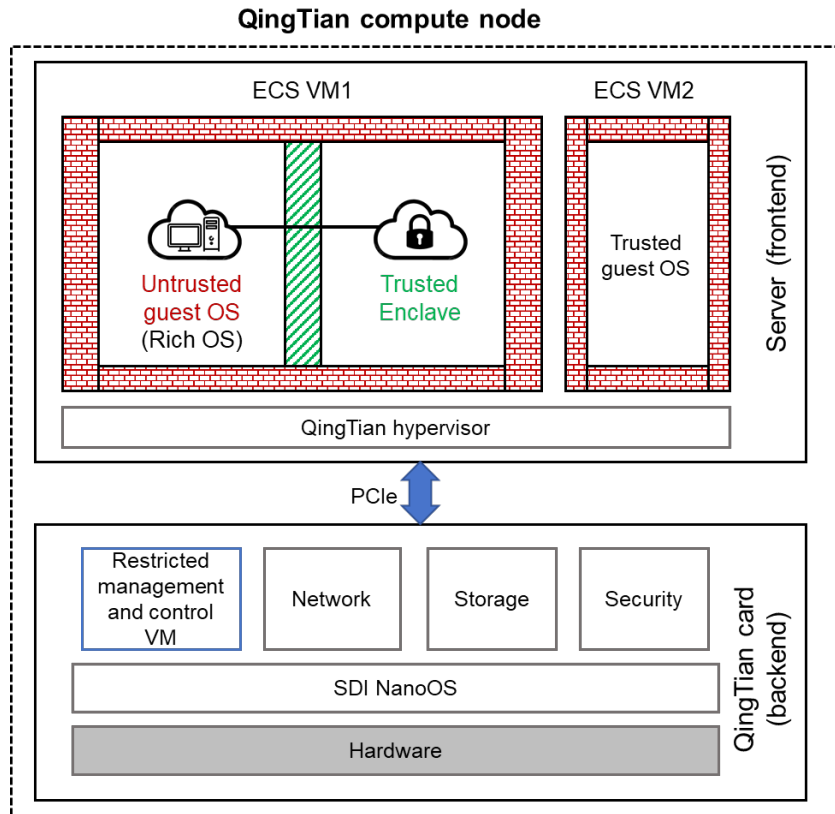


- **Isolation dimension 1:** Isolate customer data and code from CSPs' internal personnel and cloud system software to defend against attacks from CSPs' internal personnel.
- **Isolation dimension 2:** Isolate customer data and code from customers' internal personnel and untrusted guest OSs to defend against attacks from customers' internal personnel.

Isolation Dimension 1: Intra-CSP Isolation

This dimension implements Intra-CSP Isolation. Its security objective is to shield customer code and data within Elastic Cloud Server (ECS) instances from potential threats originating from the CSP's internal personnel and underlying cloud infrastructure (as shown in Figure 5-4).

Figure 5-4 QingTian compute node defending against attacks from CSPs' internal personnel



The QingTian system uses the following methods for isolation in this dimension:

- **Escape prevention:** The QingTian system is a frontend and backend separated VMM architecture, where the frontend and backend are physically isolated based on the PCIe bus. The frontend hypervisor isolates customer instances' CPU and memory based on hardware virtualization, while the backend SDI card uses SR-IOV passthrough to access VM instances (without using management software). The QingTian hypervisor has a code volume less than 1% of traditional virtualization management systems, significantly lowering the VM escape risk.
- **System tampering prevention:** The QingTian system uses forcible secure boot. Customers can enable UEFI Secure Boot and QingTian TPM when creating ECS instances to implement industry-standard secure boot, trusted measurement, and integrity verification.
- **Defense against physical attacks:** A QingTian card has an independent hardware identity. This identity is used to establish a trusted connection with the ECS management and control system to prevent node identity spoofing caused by software credential leaks. We enable volume encryption and VPC encryption on QingTian cards using hardware-protected keys. Customers have full control over the usage of data keys. QingTian cards encrypt data when it leaves compute nodes and decrypt data when it enters.
- **Zero-privilege O&M:** To ensure strong isolation between customer instances and cloud infrastructure, the QingTian hypervisor does not provide any remote login capabilities. Huawei Cloud internal SRE personnel can only use O&M APIs for remote diagnosis. In exceptional emergency scenarios, only a few authorized SRE

personnel can log in to the management and control VM via a bastion host after obtaining temporary authorization. The management and control VM only supports white list-controlled restricted O&M operations. It cannot be used to access the frontend hypervisor or the memory data of customer instances on the frontend server. We adhere to the principle of zero-privilege access to the production environment in designing the O&M system and continually update our best practices to mitigate potential risks in extreme situations.

Based on the security design methods outlined, ECS QingTian instances can meet the isolation requirements in this dimension. For QingTian bare-metal instances, there is no QingTian hypervisor running on the hosts. Customers can exclusively access the underlying mainboard systems and use related hardware features to meet their strict isolation requirements. For QingTian VM instances, customers can enable features such as secure boot, trusted boot, and remote attestation. We will soon introduce Kunpeng-based secure VM instances (such as Arm CCA) to support memory encryption to meet more customers' security compliance requirements.

Isolation Dimension 2: Intra-Customer Isolation

This dimension addresses internal threats within the customer's boundary. It is designed to isolate critical workloads from potential threats originating from the customer's internal personnel and untrusted guest OSs.

In addition to the security design in Isolation Dimension 1, QingTian Enclave is provided to isolate the customer code and data from the customers' internal personnel and untrusted guest OSs. QingTian Enclave is an isolated runtime environment created from an ECS instance. It connects to the ECS instance through a dedicated vsock secure channel. QingTian Enclave and ECS instances are isolated through hardware virtualization. QingTian Enclave not only inherits the same security protection capabilities as ECS instances, but also provides a highly isolated computing environment using the following methods:

- **Minimal TCB of QingTian Enclave:** Generally, a guest OS (such as a rich OS) has a large TCB, which usually leads to a large security attack surface. QingTian Enclave excludes the rich OS from its trust boundary. As a result, security threats to the rich OS will not affect the security of applications and data in QingTian Enclave. To reduce the attack surface, QingTian Enclave does not provide network functions, network interface attachments, persistent storage, or SSH interactive access. By default, a QingTian Enclave OS is a secure OS tailored by Huawei Cloud. Customers can also customize their own QingTian Enclave OSes.
- **Defense against guest OS attacks:** QingTian Enclave is isolated from the primary ECS instance through hardware virtualization. It does not share physical memory and CPU cores with the primary ECS instance. They can communicate only through a dedicated vsock channel protected by the hypervisor. Even if the guest OS of the primary ECS instance has security vulnerabilities or the super administrator is attacked, the attacker who controls the guest OS of the primary ECS instance cannot access the code and data in the QingTian Enclave environment. The primary ECS instance usually needs to forward vsock-based requests from QingTian Enclave applications to external networks (for example, when QingTian Enclave needs to access the external KMS service). The most severe attacks from the primary ECS instance are denial of service (DoS) attacks.
- **Trusted boot and attestation of QingTian Enclave:** When the QingTian Enclave is started, the QingTian hypervisor verifies the digital signature of the QingTian Enclave image, measures the QingTian Enclave image file and digital signature public key certificate, and stores the measurement results in the QingTian Security

Module (QTSM). QTSM provides TPM-like trusted measurement and remote attestation. The difference is that QTSM redefines the trusted measurement attributes and attestation security protocols based on ECS scenarios.

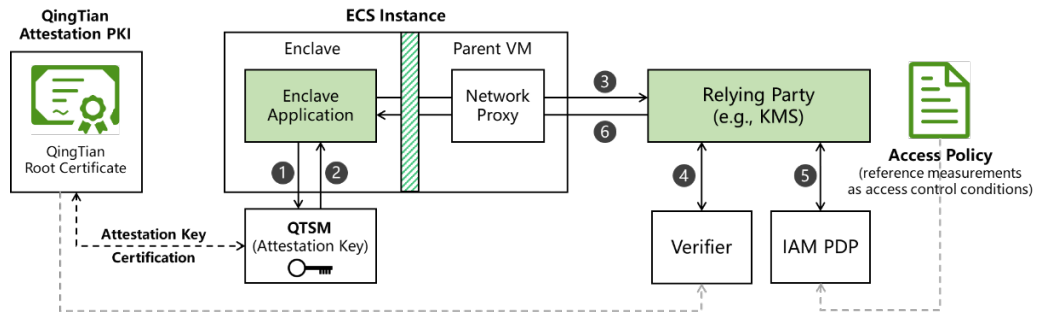
- **High usability and compatibility:** QingTian Enclave is developer-friendly. Developers can easily develop QingTian Enclave applications without CPU microarchitecture expertise and advanced cryptography. QingTian Enclave supports both x86 and Arm architectures. Developers can use their familiar language frameworks to build QingTian Enclave images using container images.
- **Cloud service integration:** The QingTian system supports cryptographic attestation for QingTian Enclave identities and trusted measurement results. QingTian Enclave applications use the attestation protocol to prove their QingTian Enclave identities and establish trust with external services. Huawei Cloud Key Management Service (KMS) and Identity and Access Management (IAM) inherently support QingTian Enclave attestation. QingTian Enclave application developers can use the open-source Enclave SDK to access KMS APIs. These APIs allow them to obtain data encryption/decryption keys or secure random numbers and ensures E2E security. Customer administrators can use preset IAM authorization policies or guardrail policies to enforce attestation-based conditional access control on KMS APIs.

QingTian Enclave enables customers to create a reinforced, highly isolated computing environment in the ECS VM environment and divide their system components into functions with different trust levels. Customers have built production applications based on QingTian Enclave, such as vHSM, vault credential management, MPC wallets, and AI confidential inference. For the cloud native confidential container solution, Huawei Cloud allows customers to configure QingTian Enclave device add-ons in Kubernetes, so that customer pods and containers can access the QingTian Enclave device driver.

Measured Boot and Remote Attestation

QingTian Enclave provides complete measured boot and remote attestation capabilities. It obtains hash values through standard trusted measurement operations and stores the hash values in the Platform Configuration Registers (PCRs) of the QTSM. This allows applications to obtain the attestation document of the current environment and execute security protocols, such as key agreement and E2E encryption, with external entities. The remote attestation Public Key Infrastructure (PKI) is designed based on the multi-level Certificate Authority (CA) hierarchy and complies with the security principle. It supports hourly certificate rotation. Each deployment in Huawei Cloud has an independent QingTian Attestation CA to ensure isolation and security. Huawei Cloud KMS and IAM services support this attestation protocol. Tenants can set IAM condition policies to implement fine-grained control (for example, only allowing the specific QingTian Enclave to call specific KMS APIs). In a typical application workflow (as shown in Figure 5-5), the QingTian Enclave generates an RSA key pair and obtains an attestation document containing the public key. Then, the QingTian Enclave submits the document and ciphertext to KMS for decryption. After verifying the document validity and checking the IAM policies, KMS encrypts the decryption result using the public key and returns the result. Finally, the QingTian Enclave application locally decrypts the result to obtain the plaintext.

Figure 5-5 Secure decryption workflow between the Enclave application and KMS service



Each ECS instance has an instance identity document generated by the service. The document provides metadata information about the instance and is updated when the instance is started, stopped, or restarted. Applications can obtain the document and its digital signature through the instance metadata service (IMDS) to verify instance attributes for remote dependencies. This provides anti-replay attack capabilities. This default birth certificate provides initial identity authentication for instances. Application access credentials can be securely obtained based on the principle of transferable trust, avoiding hardcoding static credentials in code or configurations. Users can also configure an IAM agency when creating an ECS instance. An agency is a virtual identity created by the IAM administrator and does not require static credentials. This effectively reduces the risk of long-term credential leakage. Applications obtain STS-issued temporary security tokens through IMDS. The tokens are used to access authorized cloud service resources on behalf of the agency. The IAM agency serves as the machine identity of the instance. Applications can securely call cloud service APIs without hardcoding static credentials such as AK/SK.

5.4 Prospect of Huawei Cloud QingTian Heterogeneous Architecture

Driven by the requirements of AI confidential computing, the growth of compute requirements is driving the rapid evolution of composable TEEs. A QingTian Enclave running on a CPU is expanding to a powerful secure computing cluster that can be interconnected with multiple accelerator TEEs. Huawei Cloud proposes xEnclave to organically combine CPU compute with accelerator compute such as NPU compute to support confidential AI services.

To reduce the potential risks of a single RoT in terms of security and reliability, Huawei Cloud QingTian will build a distributed RoT to enhance architecture security. The TEE running on the CPU or NPU accelerator needs to initiate access authentication to the distributed RoT before deploying or executing confidential AI services. The RoT verifies the hardware identity and trusted measurement value of the device, and stores the related results in the measurement storage of the xEnclave in ExtendPCR mode. In this way, customers can perform unified remote attestation on the entire secure cluster through the distributed RoT. Multiple measurement values and xEnclave's identity private keys are stored in different RoTs in distributed mode. This ensures that even if some RoTs have security or reliability problems, the overall cluster will not be affected.

In addition, QingTian xEnclave provides finer-grained network isolation capabilities. Through instance-level isolation, communication is strictly confined to the CPU

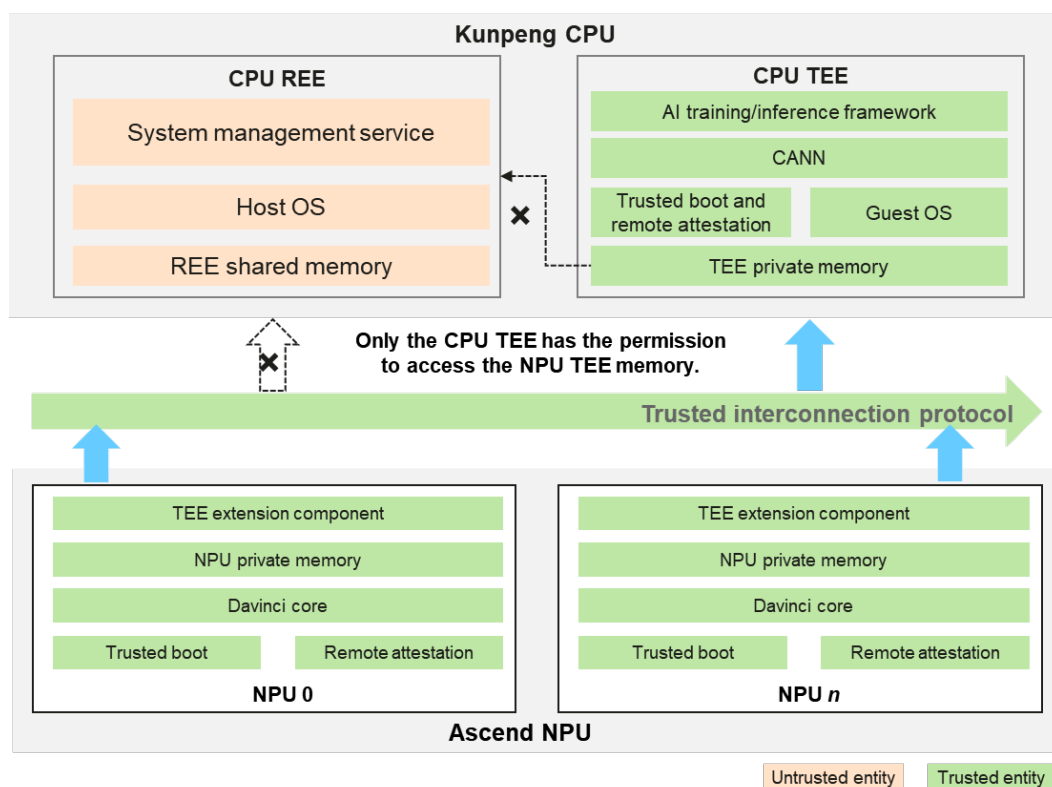
Enclave in the xEnclave and the secure accelerator. This design significantly reduces the attack surface and improves the overall security. To improve the efficiency of cross-component communication, QingTian xEnclave uses the secure passthrough technology. This mechanism enables direct communication between the QingTian Enclave and the secure accelerator without detouring through the primary instance to complete interaction through vsock. This reduces the overhead of data copy and path detours, achieves higher communication bandwidth and lower latency, and provides efficient support for confidential AI services.

6 Key Technologies of AI Platform Security

6.1 A+K Heterogeneous Confidential Computing Acceleration Platform

In AI-oriented computing power security systems, ensuring trustworthiness at a single compute node is insufficient for meeting the demands of increasingly complex AI scenarios. The growing demand for collaborative execution across heterogeneous computing units such as CPUs and NPUs—particularly in areas such as LLM inference, high-performance training, and multi-device collaboration—has introduced significant security challenges stemming from expanded computational boundaries and cross-domain data flows. To address this, HCIST provides a new security architecture called Ascend and Kunpeng (A+K) confidential computing acceleration platform that delivers heterogeneous collaboration, on the basis of the CPU-based general confidential computing platform and the NPU-based confidential computing platform, as shown in Figure 6-1. The platform establishes a cross-device trust transfer protocol and a unified memory security mapping solution through an extended security protocol – based interconnection mechanism. Using a security architecture that integrates dual hardware RoTs, E2E runtime isolation, and task-level zero-trust verification, it ensures data confidentiality while fully unleashing the collaborative efficiency of computing resources.

Figure 6-1 Huawei A+K heterogeneous confidential computing framework



Built on the Kunpeng CPU trust domain, Ascend NPU trust domain (NPU TEE), and trusted interconnection between the CPU and NPU, this platform ensures runtime data protection for heterogeneous LLMs. Guided by the design principle of "confidential data not leaving the trust domain," it ensures "data is available but invisible", safeguarding data within the A+K computing infrastructure against theft by malware or unauthorized access.

- **CPU trust domain:** Based on the Kunpeng virtCCA/CCA feature, user services are deployed in the trust domain within the hardware-based or hardened OS on the CPU. By employing technologies like memory isolation and permission control, it ensures strict permission separation between the trust domain and common domains, protecting sensitive data from unauthorized access by malicious programs or administrators.
- **Trusted interconnection:** The security interconnection protocol is extended over the bus between the CPU and NPU to implement trusted interconnection between the CPU TEE and NPU TEE. This blocks data exchange between the trust domain and common domains, preventing malicious programs and administrators from accessing data transmitted between the NPU and CPU trust domains.
- **NPU trust domain:** User services use the Ascend NPU to accelerate AI computing. Leveraging the NPU TEE confidential computing technology deployed on the NPU, the A+K confidential computing acceleration platform connects NPU devices to the trusted interconnection bus through the TEE extension component. In this way, the memory on the NPU is effectively isolated, blocking malicious programs and administrators from reading models and user inference data in the NPU trust domain.

- **Trusted boot and remote attestation:** The systems and software running on the CPU and NPU undergo rigorous trusted measurement and remote attestation to ensure that the software stack running in trust domains is not tampered with. This ensures computing infrastructure integrity and platform authenticity.

When users deploy LLM services, their critical assets—such as models, training data, inference requests, and inference results—are calculated only in heterogeneous trust domains, ensuring that user data in the computing system is available but invisible. The A+K heterogeneous confidential computing acceleration platform seamlessly integrates the computational and security strengths of the CPU and NPU, overcoming traditional bottlenecks in heterogeneous device collaboration while paving the way for next-generation secure and trustworthy computing infrastructures.

6.2 Confidential Container

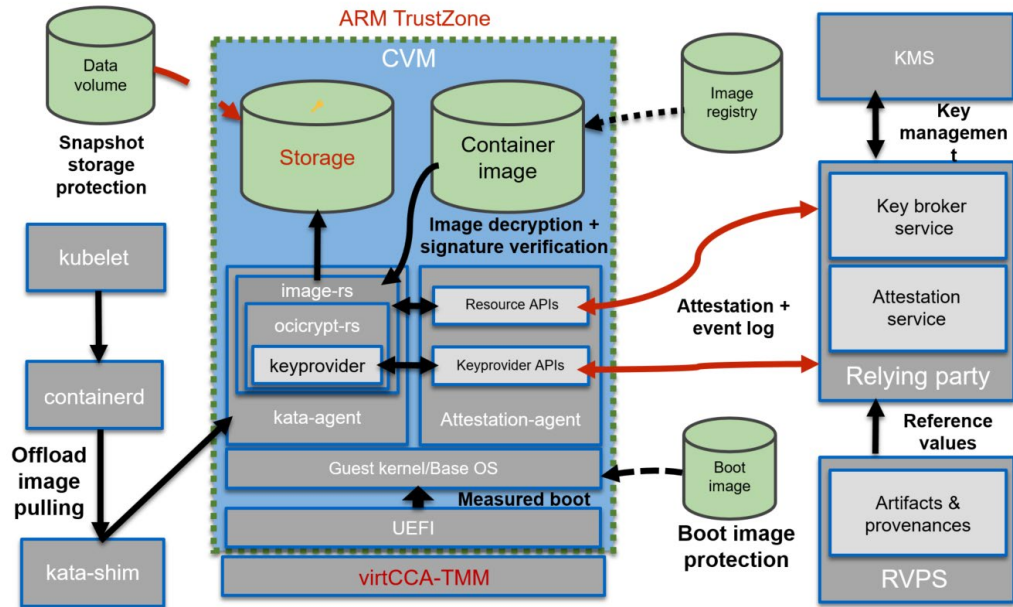
The rapid iteration and expansion of container technology has driven its growing adoption across enterprise IT architectures. The popularity of orchestration platforms such as Kubernetes has further boosted its large-scale application. However, its dynamic nature—characterized by a short lifecycle (activation/deactivation within seconds), cross-node deployment, and massive instance deployment caused by microservice splitting—poses new requirements on security protection. The security solution must ensure the integrity of static configurations and images, provide real-time protection in dynamic scheduling scenarios, and ensure the lightweight design, portability, and automated O&M of containers.

Traditional security solutions fall short in meeting these requirements. Host-based security tools often rely on privileged processes of the host machine, making it difficult to isolate sensitive operations in containers. Network layer defense struggles to handle frequent inter-container communication and monitor the dynamically changing IP addresses and ports. While application-layer encryption protects specific data, its key management and runtime decryption remain vulnerable. More importantly, most of these solutions fail to address the core security pain points in the runtime environment of containers. When containers are executed in public or hybrid cloud environments, unencrypted in-memory data becomes susceptible to theft—whether by malicious programs in the host OS kernel or hypervisor layer, or via unauthorized access by privileged cloud administrators using memory dumps or debugging tools. This poses significant risks to data protection.

To address the underlying security challenges in cloud native environments, confidential containers emerge as an innovative security enhancement technology. As a trailblazer in supporting confidential containers, Kunpeng virtCCA utilizes open-source solutions Kata Containers and CoCo for effective deployment. Kata Containers is a secure container implementation method built on the lightweight VM technology. The core idea is to run each container instance in an independent lightweight VM, leveraging hardware-assisted virtualization of the VM to build an isolated environment. In this solution, the runtime environment of a container is completely encapsulated in a VM, achieving VM-level isolation from the host and other containers. Kata Containers are compatible with the Open Container Initiative (OCI) standard and Kubernetes orchestration platform, enabling rapid scheduling, deployment, and management akin to conventional containers. They maintain the lightweight and cloud native features of containers while ensuring robust isolation. The CoCo solution aims to offer a standardized confidential container solution. Based on Kata Containers, this solution integrates the TEE capability into container lifecycle management to implement hardware-level isolation for container runtime data. It prevents unauthorized access by

the host's kernel, processes, and other entities to the VM's resources such as the memory and file system, effectively preventing attacks from the bottom layer.

Figure 6-2 Kata+CoCo confidential container framework based on Kunpeng virtCCA



Leveraging the isolation capability of virtCCA cVMs, the VM running each Kata Container can be mapped to an independent virtCCA cVM, enabling hardware-level isolation for both instruction execution and memory access while effectively blocking potential attacks from the host, hypervisor, and other containers. Taking memory management as an example, the virtCCA's TMM module can partition the physical memory by attributes in the Kata container VM to ensure that memory data is accessed only by authorized processes in the container. Even if the host kernel or other malicious entities attempt to spy, they cannot break through the security boundary of the Kata container VM. This provides a more robust hardware foundation for Kata Containers to run securely in multi-tenant environments. Moreover, the customized container runtime, image service, and key management mechanism of Kata Containers can use the virtCCA's memory isolation feature for higher security of container images during transmission, storage, and loading. Figure 6-2 shows the implementation framework of Kunpeng virtCCA - based confidential containers.

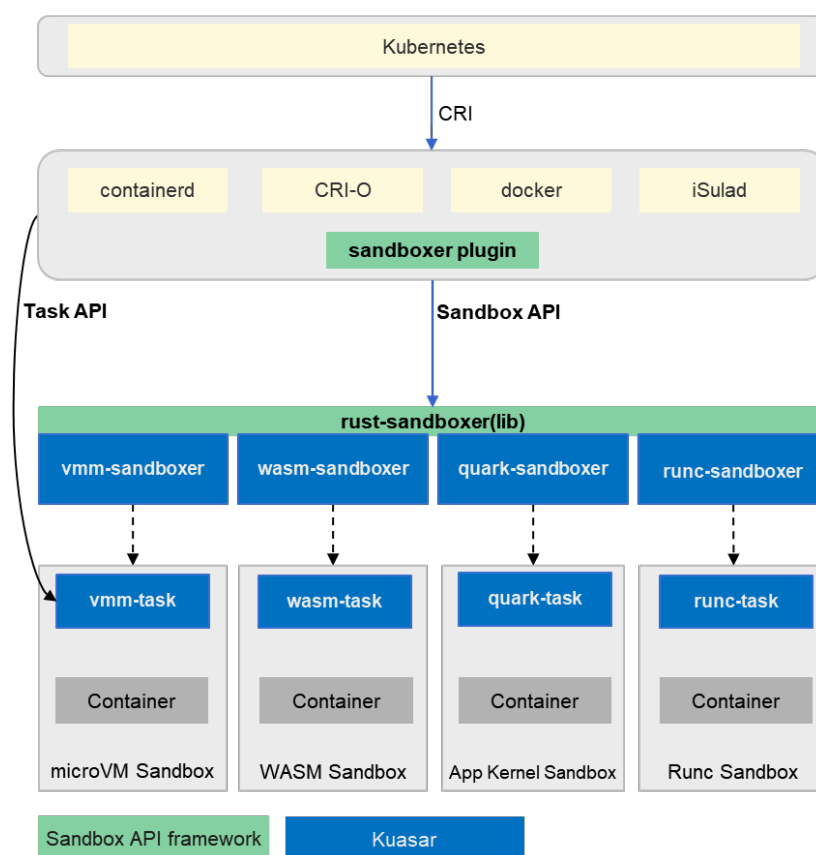
Additionally, CoCo integrates with the virtCCA's remote attestation via standardized interfaces. Starting a confidential container managed by CoCo will trigger the remote attestation process. When a container is running, it collects its environment information, such as the hash value of the container image, runtime configuration, and security policy, and generates a complete attestation report based on the hardware-level attestation information generated by virtCCA TMM. The verifier examines the attestation report to assess whether the container's running environment meets security standards, including whether the container image has been tampered with and whether the container has been attacked during running. The verifier can only transmit sensitive data to or interact with the container upon successful attestation, ensuring a trusted communication channel across nodes and platforms in a cloud-native environment.

In addition to Kata+CoCo, Huawei has launched the Kuasar+secGear confidential container solution for cloud users, which supports multiple sandbox isolation technologies. This solution can flexibly run multiple types of sandbox containers on a single node, providing powerful support for diversified application scenarios.

The following describes the Huawei Kuasar container solution. The Kuasar runtime consists of two modules, as shown in Figure 6-3:

- **Kuasar-Sandboxer**: implements the Sandbox API and manages the sandbox lifecycle and resource allocation. Sandboxer interacts with containerd as a plugin.
- **Kuasar-Task**: implements the Task API and manages the container lifecycle and resource allocation.

Figure 6-3 Kuasar secure container framework

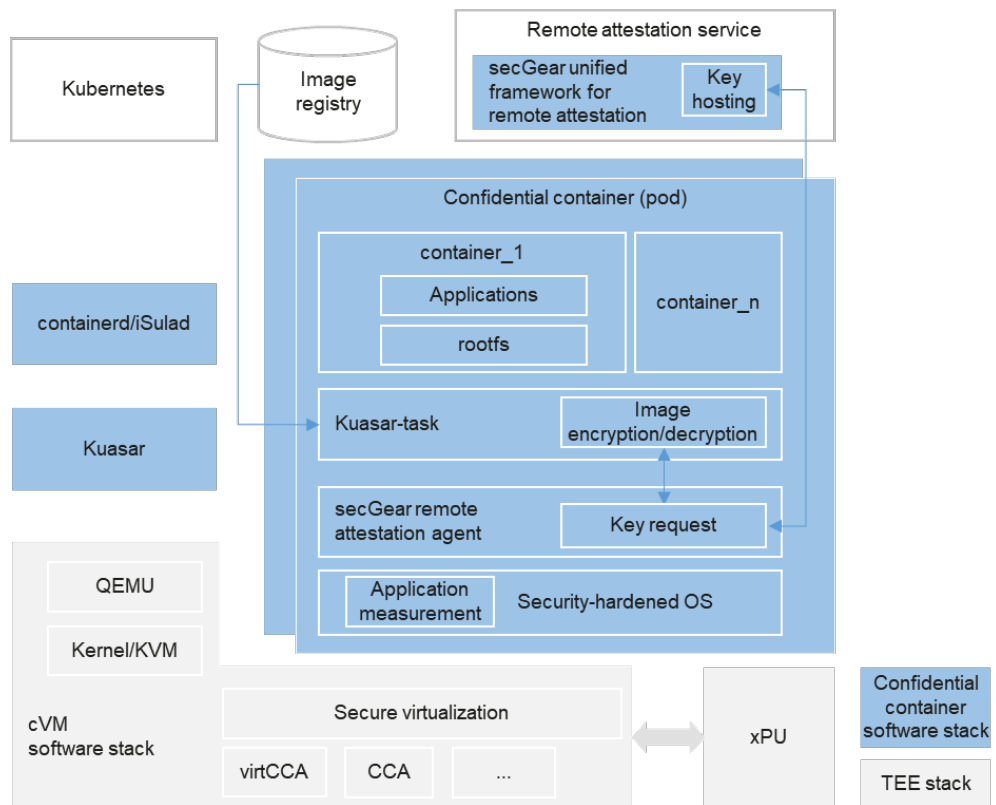


In the container runtime's shim v2 model, each time containerd creates a pod, a shim process must be created to manage the pod. Then, the shim process creates VMs and containers. In this scenario, the number of shim processes on the management plane is the same as that of pods. However, in Kuasar, only one Kuasar-Sandboxer process is required. containerd manages pods by calling external Sandboxer APIs. Therefore, containerd does not need to start a management process for each pod. Accordingly, the ratio of Sandboxer processes to pods on the management plane is 1:N. This model significantly minimizes both resident processes and management plane noise, resulting in a clearer and more streamlined architecture.

Kuasar confidential containers combine Kuasar secure containers with TEEs to safeguard the confidentiality and integrity of container images, workloads, and VM environments, preventing them from unauthorized access. In addition, the solution is compatible with the cloud-native application ecosystem, enhancing the usability of confidential computing.

Figure 6-4 shows the framework of the Kuasar+secGear confidential container solution. In this framework, Kuasar connects to QEMU in order to manage the cVM lifecycle and create containers across various platforms, such as Kunpeng virtCCA/CCA. iSulad assigns the task of pulling container images to Kuasar-task which pulls and decrypts container images within confidential containers, protecting the confidentiality and integrity of the container images. secGear abstracts away the differences between TEE hardware and provides a unified process of obtaining remote attestation and image encryption/decryption keys.

Figure 6-4 Kuasar+secGear confidential container framework



7 Pipe Security

As computing power breaks through the boundaries of individual devices and evolves towards the device-pipe-cloud synergy architecture, HCIST also needs to consider new challenges brought by large-scale AI deployment and cross-domain collaboration. To be specific, data flows frequently across different terminals, networks, and compute clusters, leading to increasingly complex security risks. This chapter focuses on pipe security in the HCIST system, covering all-round protection systems involving device-cloud synergy, network devices, cluster interconnection, and physical links. Key technologies used to ensure device-cloud protocol security, intrinsic security of network devices, and network communication security are leveraged, providing a unified and reliable network communication security foundation for large-scale AI service deployment and cross-domain computing collaboration. In addition to supporting layered decoupling and on-demand enabling, this multi-layer pipe security solution can work with the security solutions of different vendors to form different security combinations (from lightweight to E2E), meeting the security requirements of different computing infrastructures.

7.1 Device-Cloud Protocol Security

In the device-pipe-cloud synergy architecture, device-cloud protocol security is the first line of defense for secure flow of user data and computing power across domains. As smart terminals and cloud services are deeply integrated, user privacy, sensitive data, and model interaction data need to be frequently transmitted between devices and the cloud. Based on the data minimization principle and E2E verifiability, the device-cloud protocol security system built by HCIST integrates hardware RoT, dynamic key agreement, and encrypted transmission mechanisms to ensure that data is always within the trust boundary when flowing across devices, networks, or computing domains. Device-cloud protocols are designed from the following four security dimensions: who is accessing, how data is transmitted, how keys are used, and how verification is performed. Specifically:

- **Identity and access control:** Multi-level identity authentication and continuous trust evaluation are performed to implement two-way trusted verification between the users and cloud as well as between the devices and cloud. In inter-device interaction, a security key agreement protocol based on shared secrets is used to prevent any plaintext password from being exposed on the network. In addition, the risks of access behaviors are dynamically evaluated based on the zero-trust principle, so that session restriction or key invalidation is automatically triggered whenever necessary.

- **Privacy protection and anonymous access:** For mobile terminal users, the private relay and anonymous credential technologies are introduced to hide the real identities of users and the characteristics of devices. This prevents traffic profiling and cross-domain tracking. Zero-knowledge proof decouples identity validity verification from identity association, implementing a privacy protection mechanism that verifies identity but does not track identity information. This mechanism is especially suitable for cross-service collaboration in large-scale AI deployment scenarios.
- **Data encryption and permission binding:** Access permissions are directly bound to keys and encryption policies based on E2E data encryption. For example, dynamic hierarchical encryption and policy enforcement mechanisms are used for data of different sensitivity levels to ensure that high-sensitivity data can be accessed, forwarded, or edited only within the authorized scope, achieving permission and encryption integration.
- **Key management and post-quantum security:** The protocol-centric key management system ensures that the lifecycle of keys is controlled, with session keys being "one-time use" and discarded immediately after use, whereas long-term keys can be changed or even revoked according to predefined policies.
- **Verifiable computing and transparent audit:** The protocols collaborate with the cloud TEE to support remote attestation across sessions and regions, ensuring that user data is processed only in a trusted environment. Minimum recording and verifiable audit are implemented for key operation events based on the transparent log mechanism, ensuring compliance with regulatory requirements while preventing user privacy breach.

With the preceding capabilities, device-cloud protocol security moves the "data available but invisible" principle to the very beginning of communication, ensuring user privacy, compliance, and cross-domain trustworthy collaboration while also meeting requirements for low latency and high performance. This provides a solid communication protocol foundation for the large-scale deployment of AI services and multi-terminal intelligent applications.

7.2 Intrinsic Security of Network Devices

In the device-pipe-cloud synergy architecture, the security synergy between computing power and data requires not only protocol-layer protection but also the intrinsic trustworthiness of network devices. With the growth of LLM applications and cross-domain computing collaboration, network devices have become the core hub supporting data flows between clusters or regions. The security of network devices directly affects the reliability of the entire pipe system. To cope with unknown threats and enhance the resilience of computing infrastructure, HCIST introduces the "intrinsic security" concept of network devices. With security mechanisms systematically embedded into the entire lifecycle of device design, manufacturing, and operation, devices develop intrinsic self-defense, self-detection, and self-repair capabilities, helping establish a trustworthy network foundation from the source.

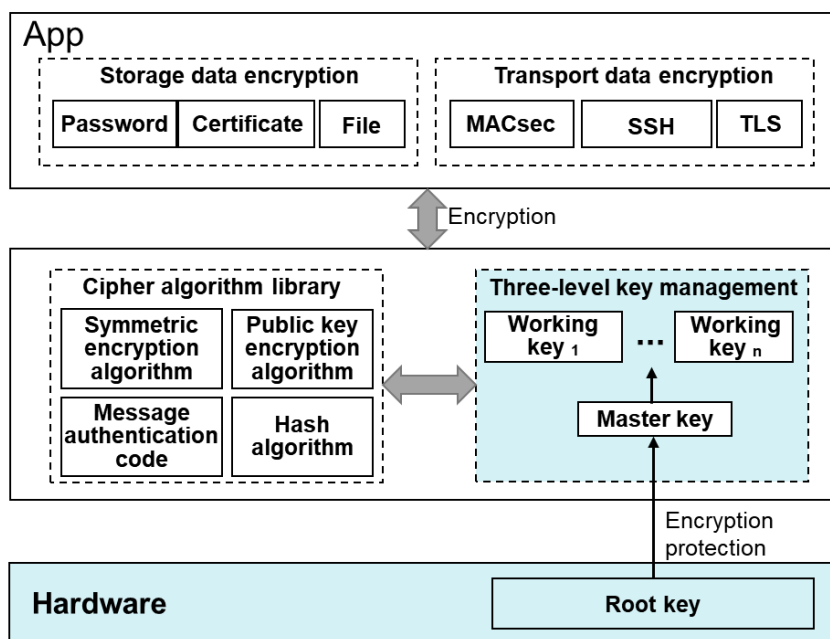
Currently, network devices face increasingly severe security challenges. Specifically, zero-day vulnerabilities continue to emerge, third-party component and supply chain risks keep rising, and malicious firmware implantation and system-level attacks become more covert. This makes it difficult to cope with continuously evolving new threats through traditional add-on defense measures represented by patch installation, firewall deployment, and virus protection. HCIST integrates security capabilities into hardware, firmware, and system architecture, ensuring that devices are in a verifiable

and trustworthy state from the design phase. In addition, the trusted boot chain implemented based on hardware RoT ensures the verifiable integrity and source reliability of each network device during deployment.

During the boot phase, the system relies on hardware RoT to perform hierarchical digital signature verification from the BIOS to the OS and then to the upper-layer software. If any verification fails, the boot process is terminated, and the corresponding log is generated. In terms of network onboarding security, RFC 8572-compliant SZTP is adopted. Leveraging two-way authentication between device and customer certificates as well as deployment file transfer through secure protocols, it protects the onboarding process against hijacking and tampering.

During the running phase, HCIST provides network devices with continuous intrusion detection and proactive protection capabilities. Devices can monitor attack behaviors such as abnormal configurations, malicious code injection, and rootkit attacks in real time, and can also work with the control system to implement automatic response. This intrinsic protection mechanism drives network devices to transform from passive defense to active immunity, deeply integrating security capabilities into the system. Meanwhile, a three-level key hierarchy (root key, master key, and working key) is implemented, with the root key fixed in the CPU to achieve hardware-level theft prevention. Data is encrypted throughout the storage and transmission processes, covering critical protocols such as SSH, BGP, OSPF, MACsec, and PHYSec. In addition, various encryption algorithms are supported, providing comprehensive protection for communication between the management plane and forwarding plane. Figure 7-1 shows the three-level key management component and overall framework.

Figure 7-1 Three-level key management component



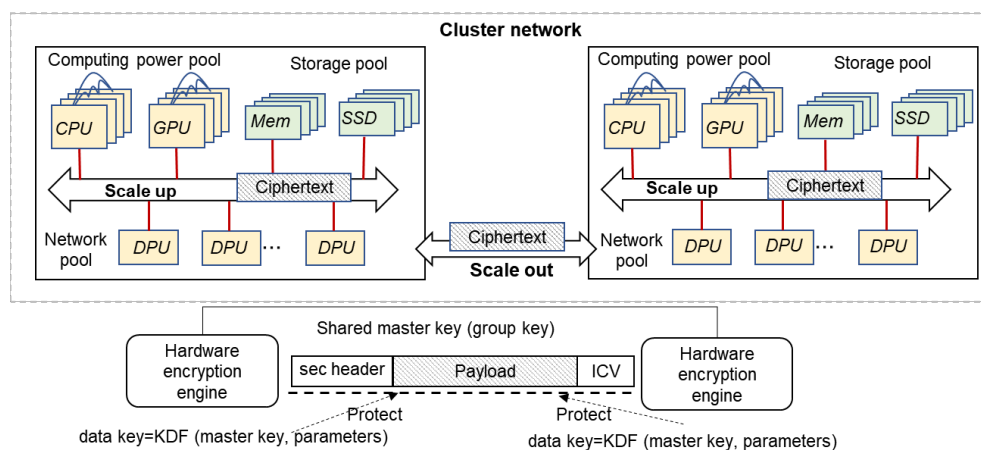
HCIST constructs a future-oriented independent and trustworthy network device system by integrating verifiable trust mechanisms throughout the device lifecycle. This system enables devices to maintain the trustworthy state in a complex, heterogeneous, and cross-domain computing environment, providing a solid hardware foundation for pipe security and critical guarantee for high-performance communication security and cross-domain computing collaboration.

7.3 Communication Security in Scale Up and Scale Out Scenarios

As LLM, generative AI, and multi-party data collaboration develop rapidly, computing demands increase exponentially, and the capability of a single compute node cannot meet the performance requirements of complex tasks. As a result, to achieve higher throughput and lower-latency data transmission, more and more computing tasks rely on scale up (multi-chip collaboration on a single node) and scale out (cross-node and cross-cluster collaboration). However, this highly interconnected computing architecture implies that sensitive information such as model weights, enterprise data, RAG indexes, and user privacy will flow frequently between different computing chips, nodes, clusters, and even data centers, presenting more complex security challenges than traditional networks.

In operations within a compute cluster or in inter-cluster collaboration, security threats exhibit multi-dimensional characteristics: In scale up scenarios, when multiple tenants share hardware resources, malicious programs can launch attacks through firmware vulnerabilities, covert channels, or configuration tampering; in scale out scenarios, high-bandwidth transmissions across different devices and regions are more susceptible to threats such as line sniffing, data tampering, and source address spoofing. Moreover, LLM training and inference have extremely high performance requirements, so that nanosecond-level latency and Tbps-level bandwidth are required in some scenarios. However, traditional security communication protocols that implement software-based encryption cannot strike a balance between performance and security. How to ensure security while maintaining high performance becomes one of the key challenges in the HCIST pipe security system.

Figure 7-2 Unisec cluster network-specific encryption/decryption communication technology



To address the preceding challenge, HCIST proposes the Unisec high-performance secure communication system (as shown in Figure 7-2). Unisec integrates trusted groups, hardware offloading, and dynamic key management to build a communication framework that balances security and performance. First, it dynamically partitions trusted groups based on task context, allowing only compute nodes and chips that have passed trusted measurement to join the groups. This prevents malicious node access. By requiring a master key to be shared within each group and data keys to be independently derived for different sessions, Unisec achieves key isolation in multi-

task parallel processing and reduces risk spreading. Second, Unisec employs hardware-accelerated stateless encryption and decryption technologies to avoid storing large amounts of security context in chips, improving key update efficiency. It also implements cross-layer unified encryption to process data at the IP, transport, and application layers at a time, reducing the latency and hardware overhead caused by repeated encryption and decryption. Finally, Unisec leverages the multi-master-key derivation technology to minimize single-key leakage risks and supports high-speed data key rotation in high-bandwidth scenarios. While maintaining high security, Unisec offers nanosecond-level response capabilities and E2E encrypted communication throughput ranging from hundreds of Gbps to Tbps, meeting the ultimate performance requirements in LLM scenarios.

Thanks to this communication system, HCIST achieves the unification of high performance and robust security in large-scale computing collaboration. Communication security in scale up and scale out scenarios, device-cloud protocol security, and intrinsic security of network devices complement each other to build a trustworthy communication foundation covering multiple terminals, networks, and computing domains, assuring AI model training, cross-region computing collaboration, and secure data flows among multiple parties.

7.4 Physical Layer Communication Security

In LLM training and cross-region computing collaboration scenarios, AIDCs need to transmit large amounts of sensitive data, including enterprises' core model parameters, training samples, and user privacy information, in high-speed network environments. During transmission, if such data is eavesdropped or tampered with, or if the associated service characteristics are inferred through traffic analysis, severe economic and security risks will arise. However, because AIDCs have high requirements on network bandwidth, transmission latency, and energy efficiency, traditional encryption solutions relying on upper-layer protocols often fail to strike a balance between security and performance.

Although MACsec is widely used in the industry for its capability to offer encryption protection at the Ethernet link layer, it has obvious limitations in AIDC and other scenarios that require high bandwidth and low latency. For example, the frame-by-frame encryption mode of MACsec introduces high processing latency and extra byte overhead, reducing the bandwidth utilization. In addition, MACsec cannot conceal traffic characteristics such as the frame length and sending frequency. As a result, encrypted traffic may still be analyzed and inferred, limiting its applicability in large-scale high-speed environments.

To address these limitations, HCIST proposes the physical layer security (PHYSec) architecture, which moves the encryption function to the physical layer so that Ethernet bit streams are directly encrypted. PHYSec leverages native OAM channels at the physical layer to transmit security parameters in advance. The receiver can pre-calculate the information required for decryption, thereby minimizing the latency caused by encryption operations and significantly reducing power consumption and chip resource usage. In addition, this mechanism encrypts all bit information (e.g., the header, payload, and length information of data frames) to effectively hide traffic characteristics and prevent traffic pattern-based analysis and side channel attacks. PHYSec can be flexibly deployed in optical modules or physical-layer chips of devices, offering compatibility with existing infrastructure and facilitating incremental upgrades.

Figure 7-3 Technical architecture of PHYSec

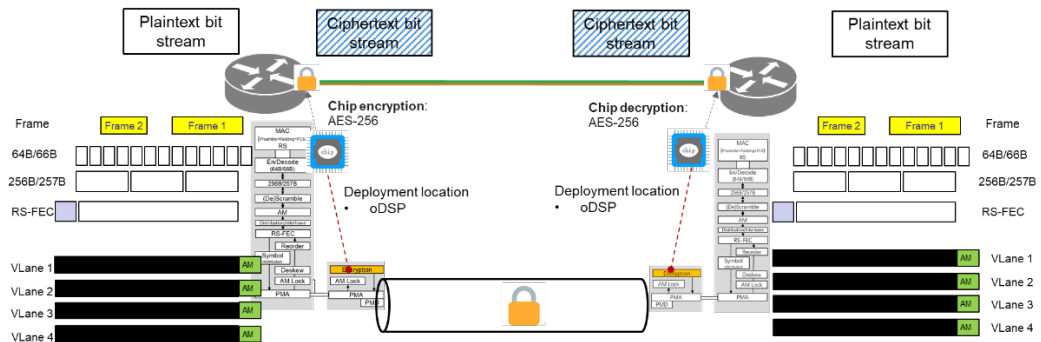


Figure 7-3 shows the general architecture of PHYSec used in the DSP of high-speed direct detection optical modules in a 200G/400G high-speed scenario. The encryption and decryption functions are mainly implemented at the physical media attachment (PMA) layer of the Ethernet physical layer. In terms of encryption, all bit streams are encrypted by virtual lane. Compared with a traditional upper-layer security mechanism, PHYSec protects all user frames and upper-layer protocols, protects Ethernet frame headers, and masks traffic characteristics such as the frame length and frame sending frequency, effectively providing defense against traffic analysis attacks.

PHYSec can be implemented in not only the DSP of optical modules (high-speed direct detection and high-speed coherent optical modules) but also PHY interfaces through flexible deployment. For PHYSec implemented in the DSPs of optical modules, security features can be incrementally deployed to upgrade the security capability of links without the need to change the hardware of existing devices. Additionally, to meet the secure encryption requirements of new Ethernet devices, PHYSec can be deployed in the PHY interfaces of new devices to protect the entire device's ports and links and remain compatible with existing optical modules.

PHYSec provides confidentiality protection and integrity check for physical-layer bit streams. High-speed direct detection links are used for short-distance interconnection and can support only confidentiality protection. High-speed coherent links are used for long-distance interconnection and can preferentially support confidentiality protection and optionally support integrity check. Compared with MACsec, PHYSec significantly reduces the overall power consumption and implementation cost. In addition, leveraging the native mechanism of the physical layer to carry security parameters, PHYSec uses one frame to transmit the security parameters required for decryption to the decryption side, so that the decryption side can calculate the strings for decryption in advance to mask the decryption latency. Compared with the traditional MACsec solution, PHYSec introduces a latency of about 20ns for 400G, which is about one order of magnitude lower than that of MACsec.

PHYSec provides low-latency, low-overhead, and high-confidentiality underlying security capabilities for high-speed computing networks by implementing the encryption protection mechanism at the physical layer. PHYSec, together with device-cloud protocol security, intrinsic security of network devices, and cross-dimensional communication security measures, forms a critical layer of the HCIST pipe security system, providing all-round, efficient, and reliable data transmission assurance for scenarios such as AI model training and cross-domain computing collaboration.

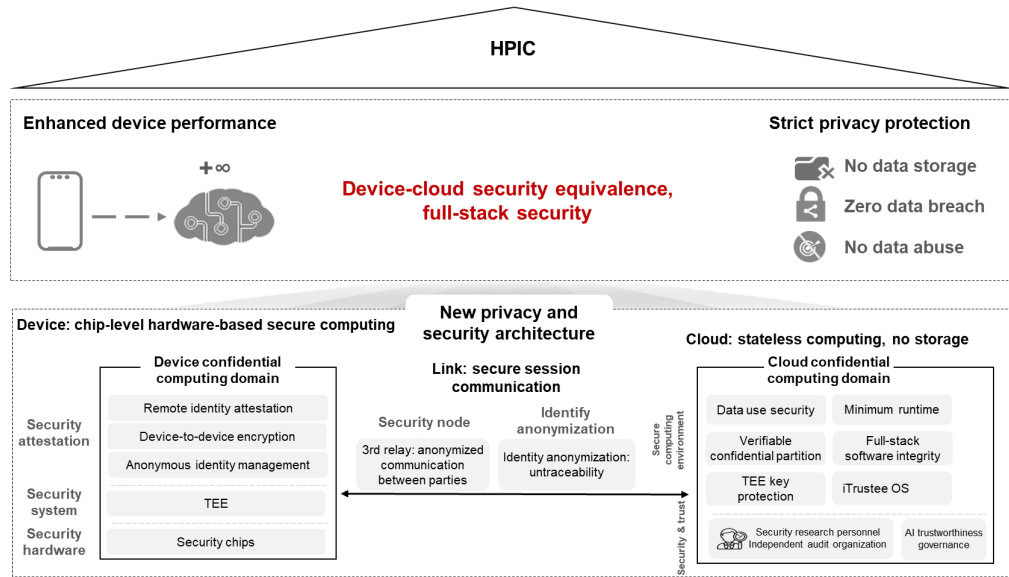
8 Typical Applications

HCIST integrates a privacy protection architecture featuring device-pipe-cloud synergy with defense-in-depth capabilities to provide secure and reliable computing power for sensitive data processing across industries. This architecture can be deployed in full-stack mode to fully harness the security strengths of computing infrastructure through device-pipe-cloud synergy. It can also interwork with computing infrastructures from different vendors to implement flexible security defense policies and adapt to diverse service scenarios as well as hardware and software facilities. This chapter describes the significant advantages HCIST offers in typical high-value scenarios in terms of security features such as device-cloud synergy, data security, LLM data security, and cloud-native security. While the cases use Huawei full-stack devices as computing infrastructure, HCIST is designed to seamlessly interwork with computing devices and security features from different vendors to enable flexible solution deployment in real-world applications.

8.1 On-Device and Cloud Collaborative LLM Inference

With the security capabilities of HCIST computing infrastructure, HarmonyOS-powered Celia—a personal privacy assistant—builds a user-centric privacy protection system (as shown in Figure 8-1) based on the core concepts of device-cloud synergy, device-cloud security equivalence, and full-stack security. Its defining feature goes beyond simply encrypting data on the device or isolating it in the cloud. Instead, it creates a dynamic, verifiable collaboration mechanism based on a chain of trust between the device and cloud. This ensures that user data remains under control throughout the entire process. Regardless of whether data is generated locally on devices or used for AI computing on the cloud, all transmission, processing, and storage paths are protected in a fine-grained manner through tightly coupled encryption and policy management between the device and cloud. Under this device-cloud-synergy-based security framework, the potential of secure cloud computing power is fully unlocked, ensuring service experience.

Figure 8-1 LLM privacy-preserving inference used in Huawei Celia based on device-cloud synergy



Specifically, when a user's intelligent processing request cannot be handled on the local device, a temporary device-level key is first generated within the device's secure environment. An encrypted channel is then established with a trusted cloud node using a privacy-preserving, 0-RTT, E2E key agreement and encryption mechanism. The cloud node can receive data only after completing TPM-based remote attestation, and the request data enters the privacy-preserving inference framework in the cloud solely as ciphertext encrypted with a one-time key. Throughout the process, untrusted cloud nodes can neither access plaintext data nor infer the user's identity—identity authentication uses a blind signature token mechanism, which is only used for session legitimacy validation, not for user tracking. Computing tasks are executed within attested secure containers and are accompanied by zero-knowledge proofs, ensuring transparency and non-repudiation of model execution.

To prevent link-level data leakage, Huawei Celia employs a device-cloud communication path designed to combine privacy-enhanced relay, i.e. OHTTP, with encrypted routes. Even in the presence of traffic-based side-channel attacks, it remains extremely difficult to retrieve original request content or trace sources. Each request channel generates a one-time key pair bound to the session lifecycle, ensuring that historical data cannot be decrypted in the future—even if the encryption algorithm is compromised. In addition, a dynamic risk-sensing mechanism is deployed. When a user engages in high-risk activities such as connecting to unknown Wi-Fi or encountering brute-force attacks, the system adjusts the device-cloud synergy policy and clears sensitive keys to minimize the attack surface of the trusted channel between the device and cloud.

During LLM inference, data such as voice and images are first encrypted and feature-extracted on the device side, with only encrypted vectors sent to the cloud. Models run within the Ascend NPU confidential computing domain, where model weights and intermediate results are isolated from the host system and administrator privileges to prevent internal and external personnel from accessing data.

Celia's true strength in privacy protection lies not in the independent capabilities of the device or cloud, but in real-time, dynamic, and verifiable collaboration between them.

Through real-time negotiation, transparent authorization, and verifiable execution, it ensures that every privacy-related operation is user-controlled, verifiable, and tamper-resistant. Following the principles of "data always under control, processing always transparent, and traces never persistent", Celia delivers intelligent privacy protection that is available, secure, and verifiable. With Celia, users can benefit from powerful cloud-based AI capabilities while retaining data subjects' rights and control over their personal data, enabling intelligent privacy protection through device-cloud synergy.

8.2 Comprehensive Data Protection

For sectors such as finance and government, where data security and service continuity are critical, Huawei has launched the industry's first "zero-loss" data security solution, providing solid protection for core industry data. Underpinned by the HCIST computing infrastructure, this solution integrates trusted computing, confidential computing, confidential storage, and high-reliability network technologies to deliver comprehensive protection from data generation and transmission to storage and ensure stable operations of critical services.

At the heart of the solution is a three-layer defense system that builds multiple trust safeguards from the data source to final storage: Layer 1 (data generation & access): By combining GaussStore's two-way authentication with chained hash verification, this solution ensures that the data sources are verifiable and undefiable, preventing forgery and data tampering. Layer 2 (data processing): Leveraging the confidential computing environment of high-performance Kunpeng servers, core service processing logic runs inside a hardware-isolated trusted execution domain. Even if the host OS or virtualization layer is compromised, service data cannot be stolen or altered. Layer 3 (data storage): With Dorado confidential storage technology, combined with write-path integrity verification and tamper-proofing mechanisms, the solution ensures immutability and auditability of core data throughout the data lifecycle.

In terms of service continuity, this solution achieves zero RPO through forced database synchronization, ensuring all committed data is replicated in real time between active and standby instances. This ensures zero loss of critical data even in extreme circumstances such as system failures, network outages, and natural disasters, enabling rapid service recovery. Combined with high-availability cluster scheduling and fast failover mechanisms, this solution enables enterprises to restore services within seconds, minimizing service interruption for users and customers.

This solution represents not only a joint technical innovation but also a strategic milestone in Huawei's data security strategy. It signals a new stage of deep integration between confidential computing and core financial systems, while also providing a replicable security blueprint for other highly sensitive sectors, such as government and energy. In today's era of accelerated digital intelligence, the "zero-loss" data security solution offers industries a long-term, trustworthy safeguard for their most valuable assets, laying a solid foundation for the development of the digital economy.

8.3 LLM Lifecycle Protection

As AI has been increasingly adopted across industries, AI models emerge as core enterprise assets. Unlike traditional software or services, these models are massive in size (often tens or even hundreds of gigabytes), expensive to train, and highly

knowledge-intensive. If leaked or maliciously exploited, it could cause immeasurable losses to an enterprise's intellectual property, competitive edges, and service security.

AI infrastructure is rapidly evolving. As model sizes and data volumes continue to grow, the storage-compute decoupling architecture has become dominant. Model files are centrally stored in a distributed storage system, while inference nodes retrieve them on demand to execute inference tasks, enabling elastic resource scheduling, cross-cloud and cross-domain deployment, and high-concurrency scaling. However, the openness and flexibility of this architecture also bring new security challenges. Centrally stored models may be vulnerable to unauthorized access or tampering. Models in transit may be intercepted or hijacked. Once encryption protection is unavailable for these models, they may be exposed to theft by host systems or malicious processes.

To address these challenges, a storage-compute decoupling protection framework has been developed using HCIST to deliver security protection across storage, transmission, and application all in confidential domains.

- Confidentiality and integrity protection in storage: Using confidential storage technologies, each storage device is equipped with a unique hardware identity derived from a hardware RoT during model persistence, and the integrity of the storage software stack is measured to ensure that the environment is uncompromised. With the compute-storage coupling mechanism, models can be accessed only by hardware-based, attested confidential compute nodes, effectively preventing models from being extracted by unauthorized nodes.
- Dual authentication and link encryption: During model transmission, a dual authentication mechanism involving compute nodes and storage devices is introduced. Both parties verify each other's identity and security states through hardware-issued measurement reports—eliminating device spoofing and credential theft risks. Hardware-accelerated link encryption, enabled by single root I/O virtualization (SR-IOV) and remote direct memory access (RDMA), ensures secure transmission even in high-performance scenarios.
- Confidential computing protection at runtime: During inference execution, models are loaded into a confidential computing environment (such as Kunpeng virtCCA/CCA), where model weights and intermediate inference results remain strictly within the confidential domain. This prevents unauthorized access by the host OS, administrators, and malicious processes. Confidential compute nodes also provide hardware-based attestations to validate the integrity of both inference task execution environments and loaded model versions.

Together, these mechanisms protect AI models throughout their lifecycle—from static storage and link transmission to runtime use—forming a verifiable, tamper-resistant, and E2E security system. In the financial industry, this solution addresses critical needs such as protection against model theft, reverse engineering, and unauthorized invocation while ensuring compliance and auditability through trusted attestation and full-lifecycle encryption. In doing so, it enables the secure deployment of AI services in cross-domain, cloud-device synergy, and multi-party data exchange scenarios.

8.4 Cloud-Native Cryptographic Applications

With the rapid adoption of cloud computing, industries are accelerating digital transformation and increasingly relying on cloud services. However, the cloud environment faces new security challenges, such as blurring of traditional security boundaries and the need to protect data in virtualized environments. Traditional

cryptographic applications typically depend on dedicated hardware cryptographic devices. However, these devices often have problems such as limited reliability, restricted availability, complex deployment, and high costs—making them unable to meet the core requirements for elastic scalability and on-demand use in a cloud-native environment. To address this, Huawei Cloud has launched a cryptographic application solution built on confidential computing in HCIST to further strengthen cloud data security. Its core advantages include:

- **Hardware-based security:** All cryptographic operations are executed within the TEE of Kunpeng servers, with key materials stored only in encrypted memory. Cloud administrators and external attackers with root privileges cannot steal them through memory dumps or other means.
- **Elastic scalability:** By virtualizing TA components (e.g., dynamically loading multiple TA instances), this solution overcomes the physical limitations of traditional hardware security modules by enabling cryptographic capacity expansion as needed.
- **High performance and low latency:** With Kunpeng chips' hardware acceleration instructions, cryptographic operations in the TEE can achieve a performance close to that in the REE. Cryptographic services and service applications are deployed on the same node, eliminating the latency caused by cross-network invoking of traditional hardware security modules.
- **Lower costs and easier O&M:** There is no need to purchase dedicated cryptographic hardware devices. Instead, the TEE capabilities of cloud servers provide equivalent levels of security, reducing both procurement and O&M costs.

This solution integrates confidential computing with cryptography, overcoming the elasticity, cost, and O&M limitations of traditional hardware security modules, while providing equivalent levels of security assurance. It addresses data security challenges on the cloud, and redefines the delivery model of cryptographic services by shifting from dedicated hardware to elastic services tailored for cloud-native technologies. Looking ahead, as confidential computing gains wider adoption, cryptographic applications built on this technology will become a critical component of cloud security infrastructure.

9 Future Outlook

In the future, HCIST will adopt a holistic architectural design to address both technological evolution and service needs. This approach will enhance collaboration efficiency and expand the application scope while ensuring security. First, as quantum computing continues to develop, traditional cryptographic algorithms will gradually become ineffective in providing long-term protection. Post-quantum cryptography will therefore become essential for securing the infrastructure. This requires deep integration of quantum-resistant security capabilities into the hardware and software—integrating quantum-resistant algorithms directly into processors and accelerators and working with system-level key management mechanisms to ensure stable encryption and authentication capabilities in quantum threat environments. This evolution is not only a defensive upgrade, but also lays the technical foundation for trusted computing over the next decade.

With the evolution of cluster architectures, HCIST will further integrate with Huawei's latest Unified Bus (UB) architecture to build a high-performance, secure, and trusted interconnection mechanism for clusters. Developed by Huawei, the UB is a high-bandwidth, low-latency intra-cluster interconnection bus that enables shared memory access and unified data communication paths among nodes, facilitating the deployment of high-performance computing clusters. By integrating hardware-based security capabilities—such as link-layer encryption, port access control, and device identity authentication—directly into the UB and related devices, HCIST enables compute nodes to achieve high-performance interconnection while maintaining strict data and task isolation.

On this basis, confidential computing will evolve from single-node and cross-node deployments to cluster and superpod deployments. Through the UB's intrinsic security mechanism, encrypted communication protocols, and hardware-based remote attestation, heterogeneous compute units can collaborate within a unified, cross-node, and cross-cluster confidential domain. This enables sensitive data to be processed in a distributed manner without leaving security boundaries. Such capabilities not only support distributed training, fine tuning, and inference of super-large models, but also sustain efficient scheduling and trusted isolation of computing resources in cross-regional, multi-party service scenarios—allowing confidential computing to achieve true cloud-native elasticity and global deployment.

At the same time, the trust architecture supporting cross-cluster synergy will evolve from the single-point RoT to the distributed RoT. Trust anchors are dispersed across hardware nodes and platforms, leveraging technologies such as key fragmentation to mitigate the risks of single-point failures. Combined with cryptographic measurement and distributed storage and attestation, the security state of both intra- and inter-clusters can be transparently validated. This architecture will significantly strengthen

overall resilience in complex environments—such as large-scale distributed deployment, edge computing, or multi-cloud synergy—ensuring that trusted computing remains stable even beyond physical and organizational boundaries.

Through these advancements, HCIST will evolve into a comprehensive secure computing infrastructure framework encompassing quantum security, cluster collaboration, distributed trust, and AI protection throughout the lifecycle. It will provide long-term, trustworthy, and efficient technical support for the intelligent ecosystem of the future.