Cloud-Native Confidential Computing: Trusted Execution Environments and Confidential Containers in Modern Public Clouds

(Sun Yining)[1]

Nanjing University of Information Science and Technology, Nanjing, China
1728041450@qq.com

Abstract. Confidential computing protects data in use through hardware-based Trusted Execution Environments (TEEs), addressing a longstanding vulnerability in cloud security. As of December 2025, the Confidential Computing Consortium's IDC study reveals accelerating adoption, with confidential computing emerging as a strategic imperative for secure AI, multi-party data collaboration, and compliance with regulations like EU DORA. This paper provides an in-depth survey of the latest advancements in cloud-native confidential computing, including confidential GPUs and integrations in public clouds. We examine TEE technologies, system architectures bridging hardware isolation with Kubernetes orchestration via the CNCF Confidential Containers project, and real-world use cases such as privacy-preserving analytics, confidential AI, and secure key management. Performance evaluations and third-party benchmarks demonstrate minimal overheads in modern implementations. Finally, we analyze persistent challenges like side-channel risks, attestation complexity, and interoperability, while outlining future directions including confidential edge computing and enhanced standardization.

Keywords: Cloud computing · Confidential computing · Trusted execution environments · Kubernetes · Confidential containers · Confidential GPUs

1 Introduction

The explosive growth of cloud computing has revolutionized data storage and processing, enabling unprecedented scalability, flexibility, and innovation across industries ranging from finance and healthcare to manufacturing and entertainment. Organizations can now deploy complex applications globally with minimal upfront investment, leveraging pay-as-you-go models and elastic resources. However, this convenience comes with significant security challenges. Traditional encryption mechanisms effectively protect data at rest (stored on disks) and data in transit (moving over networks via TLS), but they fall short when data is actively processed in memory—commonly referred to as "data in use." In multi-tenant public cloud environments, data loaded into CPU registers or RAM is typically

in plaintext, making it accessible to privileged software layers such as the hypervisor or host operating system. This creates vulnerabilities to advanced threats, including hypervisor-level exploits, malicious insiders with administrative access, co-tenant side-channel attacks exploiting shared hardware resources, and even scenarios where the cloud provider itself could theoretically access customer data under legal compulsion. Confidential computing emerged to address this exact gap. It leverages hardware-based Trusted Execution Environments (TEEs) to create isolated processing regions where code and data are protected from all external access, including from the host OS, hypervisor, and cloud provider. Key guarantees include confidentiality (encrypted memory), integrity (tamper detection), and remote attestation (cryptographic proof of the environment's state). As of December 28, 2025, confidential computing has transitioned from a niche research topic to a mainstream production technology. A comprehensive IDC study commissioned by the Confidential Computing Consortium (CCC) and released in December 2025 surveyed over 600 global enterprises and highlighted strong adoption momentum for AI use cases, while identifying attestation complexity and skills shortages (75% of organizations) as primary barriers. Adoption is accelerating due to regulatory pressures—such as the EU Digital Operational Resilience Act (DORA), fully effective in 2025, which mandates protection of data in use for financial institutions—and the explosive growth of sensitive AI applications. Major public cloud providers have invested heavily: Amazon Web Services (AWS) has achieved global availability for Nitro Enclaves; Microsoft Azure offers production confidential VMs with AMD SEV-SNP, previews of Intel TDX, virtual TPM support, and confidential GPUs (e.g., NCCads H100 v5 series); Google Cloud provides Confidential Space with integrated GPU capabilities; and NVIDIA has matured confidential computing mode on Hopper (H100) and Blackwell GPUs, enabling secure AI acceleration with near-native performance. In the cloud-native ecosystem, the Cloud Native Computing Foundation (CNCF) Confidential Containers (CoCo) project—maintaining Sandbox maturity as of December 2025—represents a critical bridge, allowing Kubernetes pods to run inside TEE-protected lightweight virtual machines while preserving standard orchestration semantics. This report offers a thorough exploration of cloud-native confidential computing as one of the most significant emerging trends in cloud technology for 2025. The contributions are fourfold: (1) a detailed, up-to-date survey of TEE hardware and public cloud integrations; (2) comprehensive system architecture descriptions, including comparisons and integration details; (3) in-depth analysis of practical use cases with performance considerations; and (4) a balanced examination of current limitations, challenges, and promising future research directions. The remainder of this paper is organized as follows: Section 2 presents foundational background and related work; Section 3 delves into system architectures; Section 4 explores representative use cases with evaluation insights; Section 5 discusses limitations and ongoing challenges; Section 6 concludes and proposes future directions.

2 Background and Related Work

2.1 Cloud Security Threat Model and the Imperative for Data-in-Use Protection

Public cloud platforms operate under a well-established shared responsibility model: the provider secures the physical infrastructure, virtualization layer, and network, while customers are responsible for securing their applications, data, and access controls. Mature protections include server-side encryption for data at rest (using standards like AES-256) and transport-layer security for data in transit. However, when data is decrypted for computation, it resides in plaintext within processor caches and main memory, creating a significant exposure window. Advanced persistent threats can exploit this window through various vectors: hypervisor compromises (e.g., via zero-day vulnerabilities), insider threats from cloud personnel with privileged access, side-channel attacks (e.g., cache timing or power analysis exploiting shared hardware), and supply-chain attacks on firmware or management tools. Real-world incidents, such as historical virtualization escapes and side-channel demonstrations (e.g., Spectre/Meltdown variants), underscore these risks. Regulatory evolution has heightened the urgency. The EU DORA regulation, fully effective in 2025, explicitly requires financial entities to implement controls protecting data in use. Similar requirements appear in updated frameworks like GDPR, HIPAA, and PCI-DSS interpretations. These drivers, combined with the rise of sensitive AI workloads processing proprietary models or personal data, have made data-in-use protection a board-level priority. Confidential computing directly addresses these concerns by moving trust from software stacks to hardware roots-of-trust embedded in modern processors.

2.2 Trusted Execution Environments: Hardware Foundations

Modern TEEs are built directly into processor silicon, providing isolated execution contexts with dedicated encrypted memory regions. The leading implementations available in late 2025 are:

- Intel Trust Domain Extensions (TDX): Extends earlier SGX enclave technology to full virtual machine isolation, supporting seamless migration of existing workloads. Production deployments are available with Intel Xeon processors.
- AMD Secure Encrypted Virtualization with Secure Nested Paging (SEV-SNP): Provides memory encryption, integrity protection, and defenses against certain physical attacks. It is widely deployed in public clouds, powering confidential VMs on Azure and Google Cloud.
- ARM Confidential Compute Architecture (CCA): Introduces realms as isolated worlds, optimized for cloud, mobile, and edge scenarios.
- NVIDIA Confidential GPUs: A groundbreaking extension of TEE principles to GPU accelerators on Hopper (H100) and Blackwell architectures. GPU memory and compute units are protected, enabling secure training and inference of large language models with near-native performance.

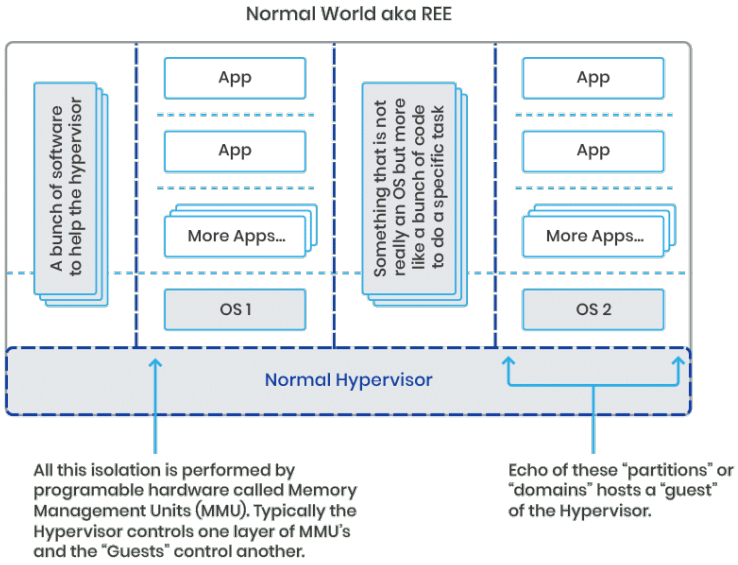

Fig. 1: Generic Trusted Execution Environment (TEE) architecture illustrating strict isolation from the host OS and hypervisor (source: Trustonic).

A universal and critical feature across all TEEs is remote attestation. During boot or instance launch, the hardware generates a cryptographically signed report (quote) containing measurements of firmware, hypervisor, OS, and application code. This quote can be verified against expected values by a remote relying party, establishing trust before any sensitive data is released.

2.3 Confidential Computing Platforms and Ecosystem

The Confidential Computing Consortium (CCC), hosted by the Linux Foundation, serves as the primary industry body for standardization, open-source collaboration, and advocacy. Its December 2025 IDC report provides the most current market insights, highlighting strong adoption momentum for AI use cases while identifying attestation complexity and skills shortages as primary barriers. Public cloud providers have reached production-grade maturity:

- AWS Nitro Enclaves: Dedicated isolated compute environments attached to EC2 instances, leveraging the Nitro hypervisor for strong boundaries.
- Microsoft Azure Confidential VMs: Comprehensive feature set including SEV-SNP, TDX support, virtual TPM for disk encryption keys, and confidential GPU instances (H100-based).
- Google Cloud Confidential VMs and Space: Integrated tooling for attested workloads and GPU-accelerated confidential computing.

– NVIDIA Confidential Computing Mode: Hardware-level protection for GPU memory pages and compute, integrated across major clouds on Hopper and Blackwell.

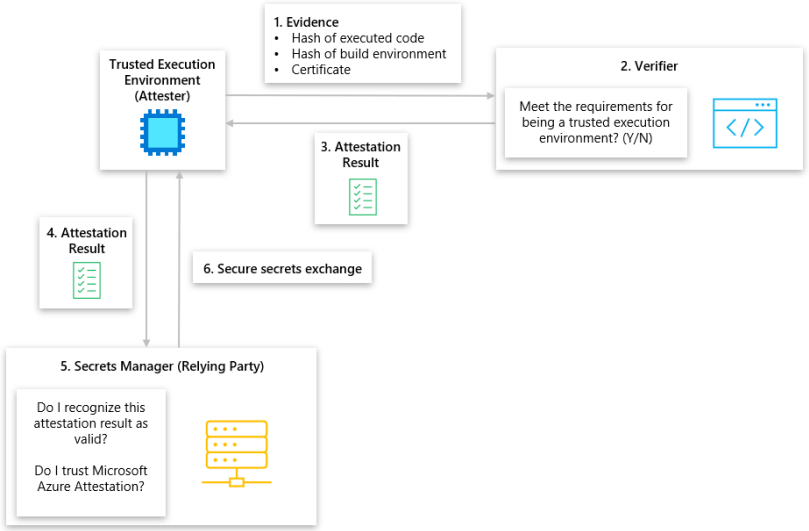

Fig. 2: High-level confidential computing workflow showing remote attestation and secure key provisioning (source: Red Hat).

2.4 Cloud-Native Integration: The Confidential Containers Project

Traditional confidential computing focused on VMs or enclaves, but modern applications are increasingly containerized and orchestrated with Kubernetes. The CNCF Confidential Containers (CoCo) project addresses this by enabling pod-level TEE protection. As of December 2025, CoCo remains at Sandbox maturity but enjoys broad industry support from Intel, AMD, ARM, NVIDIA, Red Hat, and Microsoft. CoCo works by replacing standard container runtimes with lightweight VM-based runtimes (primarily Kata Containers) that launch pods inside hardware TEEs. Attestation is performed automatically during pod creation, and secrets are injected only after successful verification.

3 System Architecture

3.1 Generic Confidential Computing Architecture and Attestation Flow

A complete confidential computing deployment consists of several tightly integrated components:

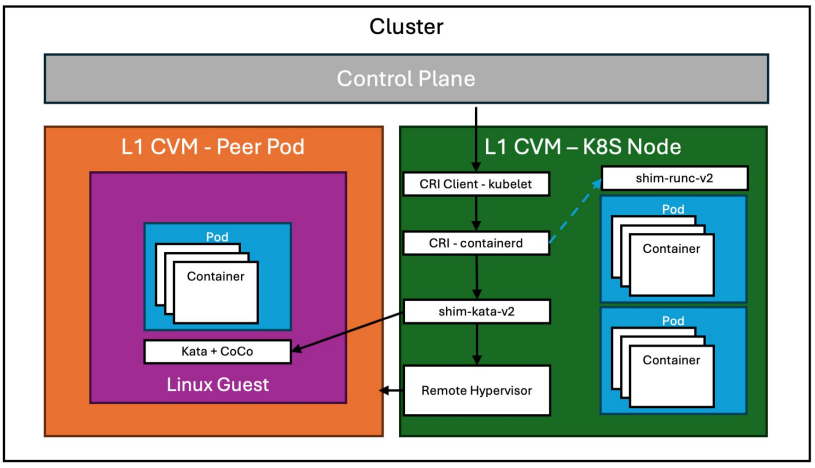

Fig. 3: Confidential Containers (CoCo) architecture within a Kubernetes cluster (source: Thomas Van Laere).

1. Hardware Root-of-Trust: Processor silicon and firmware providing encrypted memory and measurement capabilities.
2. Isolated Workload: Running as a confidential VM, enclave, or CoCo pod.
3. Attestation Service: Independent verifier (often cloud-provided or third-party like Intel Trust Authority) that checks quotes against policies.
4. Key Management Integration: Cloud KMS or external HSM that releases secrets only upon successful attestation.
5. Relying Party (Client): The external application or service that initiates attestation and provisions data.

The typical flow is: (1) workload launches in TEE and generates quote; (2) quote sent to attestation service; (3) verification succeeds; (4) KMS releases secrets into the protected environment.

3.2 Comparison of Confidential Computing Approaches

Approach	Granularity	Typical Overhead	Key Features	Majo
Confidential VMs	Full VM	5–15%	Lift-and-shift, full OS support, vTPM	Azure, GC
Nitro Enclaves	Process/module	¡5%	Fine-grained, local channel to parent	
Confidential Containers (CoCo)	Pod	10–20% (variable)	Cloud-native, Kubernetes-native	CNC
Confidential GPUs	Accelerator	5–10%	Secure AI training/inference	NVIDIA +

Table 1: Comparison of major confidential computing approaches as of December 2025.

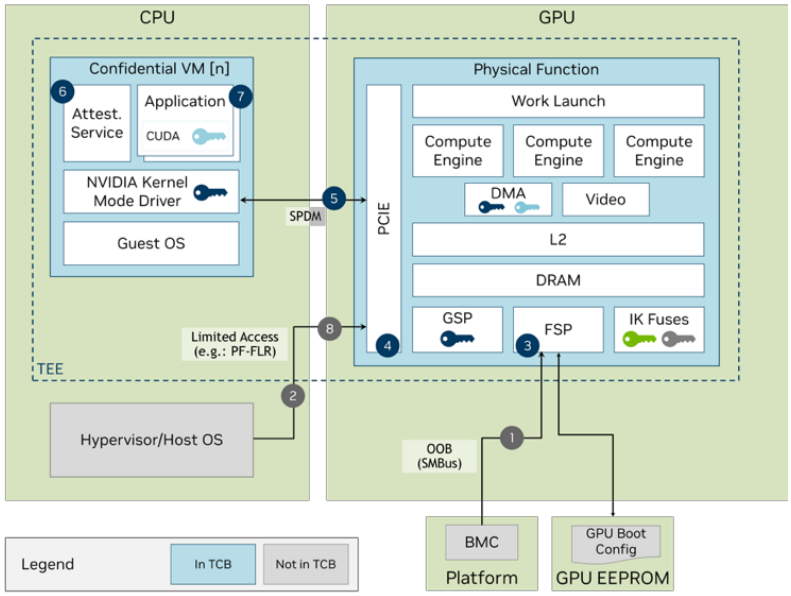

Fig. 4: NVIDIA Confidential GPU secure initialization process (source: NVIDIA Developer Blog).

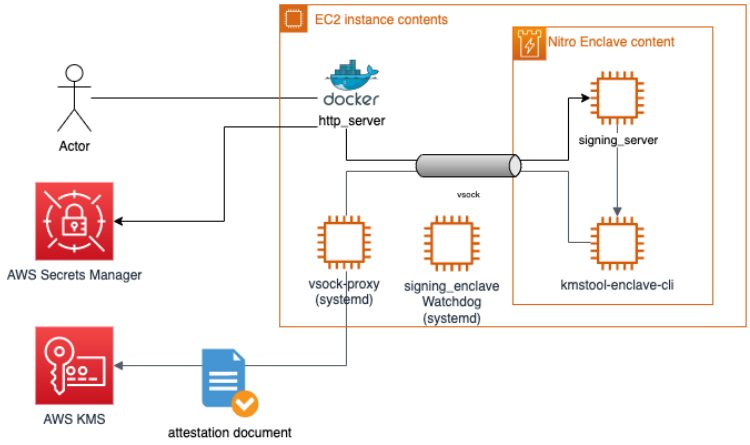

Fig. 5: AWS Nitro Enclaves architecture showing isolation from parent EC2 instance (source: AWS).

3.3 Integration Details: Confidential Containers and Confidential GPUs

CoCo extends Kubernetes with custom runtime classes and admission controllers. Pods annotated for confidentiality are scheduled only on nodes with TEE-capable hardware. The CoCo shim runtime launches a lightweight VM (Kata) inside the TEE, performs hardware attestation, and signals success to the Kubernetes control plane. NVIDIA confidential GPUs can be requested as resources in CoCo pods, enabling attested AI workloads where both CPU and GPU memory remain protected throughout execution.

4 Use Cases

4.1 Privacy-Preserving Analytics and Data Clean Rooms

Organizations increasingly need to perform joint analytics on sensitive datasets (e.g., advertising measurement, healthcare research) without exposing raw data. Confidential clean rooms execute queries inside TEEs, releasing only aggregated results. Azure Confidential Computing and partners like Snowflake provide managed solutions built on confidential VMs and attestation.

4.2 Confidential Machine Learning and Generative AI

The rise of large language models has created acute needs to protect proprietary model weights and sensitive user inputs (e.g., medical records, financial transactions). Confidential GPUs enable end-to-end protection: prompts are decrypted only inside the TEE, inference occurs on protected GPU memory, and outputs are encrypted before leaving. This satisfies regulatory requirements while preserving model IP.

4.3 Secure Key Management and Cryptographic Operations

Confidential environments serve as software-based hardware security modules (HSMs). Keys can be generated inside the TEE and never exported in plaintext, supporting use cases like blockchain wallet management, certificate authorities, and payment processing.

4.4 Performance Evaluation and Benchmark Insights

Independent benchmarks conducted throughout 2025 (e.g., on Azure confidential series, NVIDIA testing) consistently show that modern TEE implementations have dramatically reduced overhead compared to early generations:

- CPU-bound workloads: 2–5% overhead on latest AMD EPYC and Intel Xeon with SEV-SNP/TDX.
- Memory-intensive workloads: 5–8% due to on-the-fly encryption.

Workload Type	Standard Performance	Confidential Performance (Relative)
CPU-intensive (SPEC)	100%	95–98%
Memory bandwidth	100%	92–97%
GPU Inference (H100 LLM)	100%	90–95%
Mixed AI training	100%	88–94%

Table 2: Representative performance overheads aggregated from 2025 public benchmarks.

– GPU inference (LLM serving): 5–10% on H100/Blackwell in confidential mode.

Overheads continue to decrease with each hardware generation as memory encryption engines become more efficient.

5 Limitations and Challenges

Despite impressive progress, several challenges remain that limit universal adoption:

– Security Limitations: Side-channel attacks (e.g., cache-based timing, voltage glitch) persist as an active research area. While newer TEEs (SEV-SNP, TDX) include mitigations, no hardware guarantee is absolute against nation-state adversaries.
– Operational Complexity: Designing, managing, and troubleshooting attestation policies requires specialized expertise. Debugging inside isolated environments is restricted—no live memory dumps or traditional profilers.
– Performance and Cost Trade-offs: Confidential instances typically carry a 20–50% price premium due to dedicated hardware SKUs. Certain I/O-heavy workloads may experience higher latency.
– Ecosystem and Standardization: The CoCo project is still Sandbox-level, meaning APIs may evolve. Full cross-vendor attestation interoperability remains a work in progress within the CCC.
– Developer Skills Gap: The December 2025 IDC study reports that 75% of organizations cite lack of in-house expertise as a major barrier.

Addressing these will require continued collaboration between hardware vendors, cloud providers, and the open-source community.

6 Conclusion and Future Directions

As of December 28, 2025, confidential computing stands as one of the most transformative advancements in cloud security over the past decade. By shifting trust to hardware roots-of-trust and enabling verifiable isolation, it has made

previously impossible workloads—secure multi-party AI, compliant data clean rooms, protected generative inference—feasible at scale. Public cloud providers have delivered production-ready, globally available services, while cloud-native projects like Confidential Containers ensure compatibility with modern DevOps practices. Looking ahead, several exciting directions promise even broader impact:

- Standardized, vendor-agnostic attestation formats and verification services.
- Extension of confidential computing to edge devices, IoT, and 5G infrastructure.
- Tight integration with emerging agentic AI and autonomous systems requiring verifiable computation.
- Next-generation processor designs approaching zero measurable overhead.
- Open-source tooling to democratize attestation policy authoring and debugging.

These developments will likely make confidential computing a default expectation rather than an optional enhancement in the coming years.

References

1. Confidential Computing Consortium. New Study: Confidential Computing Emerging as a Strategic Imperative for Secure AI and Data Collaboration. December 2025.
2. IDC. Unlocking the Future of Data Security: Confidential Computing as a Strategic Imperative. Sponsored by Confidential Computing Consortium, December 2025.
3. NVIDIA. Confidential Computing on Hopper and Blackwell GPUs Whitepaper. 2025.
4. Microsoft Azure. Confidential Computing Documentation. 2025.
5. Amazon Web Services. AWS Nitro Enclaves Documentation. 2025.
6. CNCF. Confidential Containers Project Status Report. 2025.
7. Red Hat. Attestation in Confidential Computing. 2025.