Running AI Agents Inside TEEs Without Losing Your Mind

Sonali Mishra

Principal Product Manager, Cloud Native & AI

Confidential Computing Summit 2026

NUTANIX

AI Agents now

Make decisions

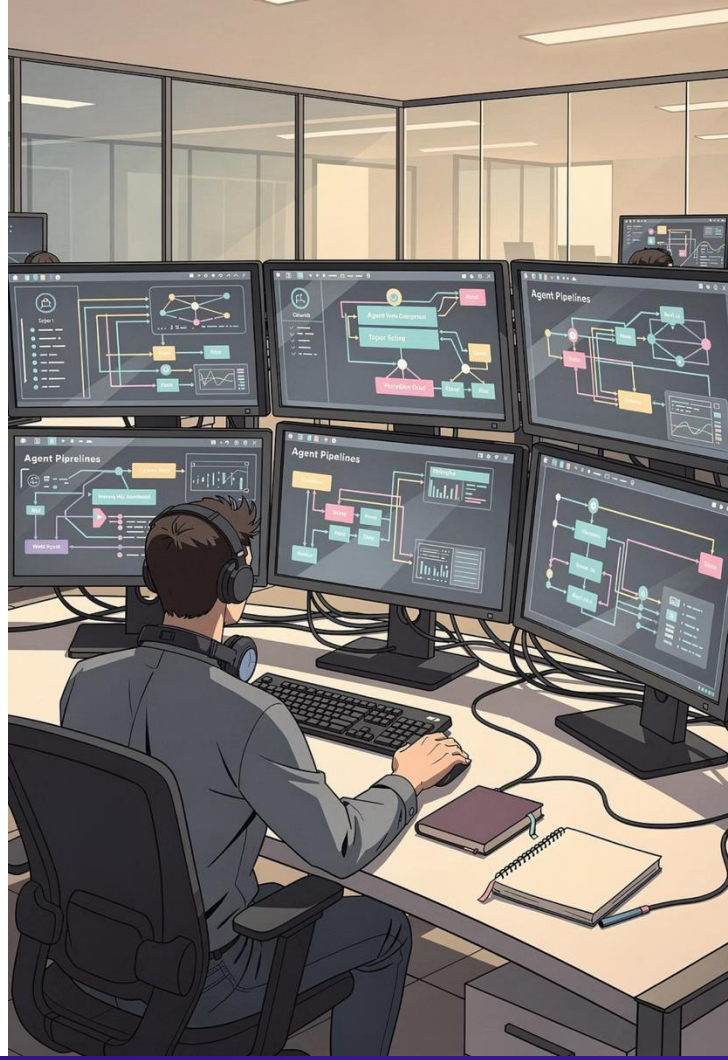
Acting autonomously on behalf of users with real data in real time

Call tools

Talking to APIs, other agents, and external systems at runtime

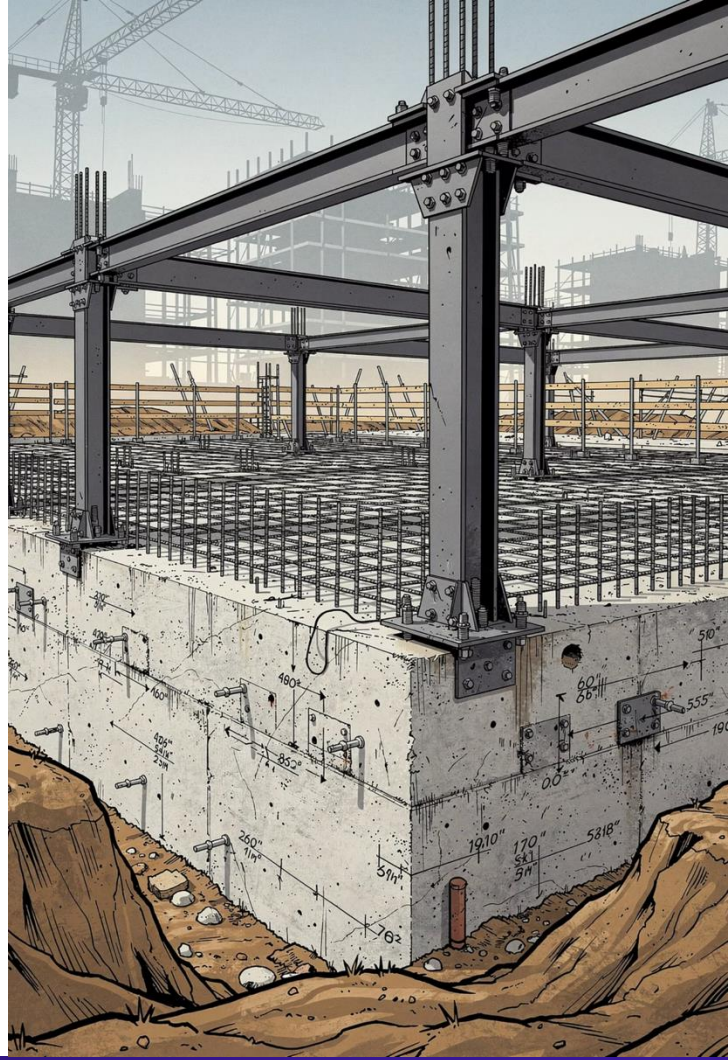
See sensitive data

Often plaintext - data their operators cannot legally access



The cloud provider's hypervisor has the technical capability to read your VM's memory

- Strong controls. No evidence of routine abuse. Excellent compliance posture.
- But capability ≠ no capability. That's the question regulators are starting to ask.



What a TEE gives you

Secrecy

Memory is encrypted by the CPU

Integrity

Code can’t be tampered with

Provability

You can prove what’s running, remotely

TCB: What you have to Trust

Regular Cloud VM TCB

- Your code + libs
- OS + drivers
- Hypervisor
- Firmware
- Cloud operators
- Datacenter staff
- Supply chain

TEE TCB

- Your code + libs
- OS + drivers
- Small firmware
- CPU vendor

The Three Primitives

Measured Boot

Fingerprint of what loaded

Remote Attestation

Signed proof, sent to a verifier

Sealing

Encrypt secrets to your own identity

Vendor Landscape

Pick what fits your cloud and your performance profile.

Intel SGX

Process-level, broken many times. Don't put new workloads on classic SGX.

Intel TDX

VM-level, current default. First public audit Feb 2026: 5 CVEs.

AMD SEV-SNP

VM-level, AMD's answer. Steady stream of attacks.

ARM CCA

Newer, cleaner, less battle-tested.

AWS Nitro

Software TEE on AWS hypervisor. Tight KMS integration. AWS is in your TCB.

Azure / GCP

Support all of the above.

Why Agents are Harder

Agent to Agent Communication

Many TEEs must verify each other - solved with RA-TLS, but adds significant complexity.

Dynamic Tools

Runtime calls to external APIs and tools weren't in the boot measurement - the TCB shifts at runtime.

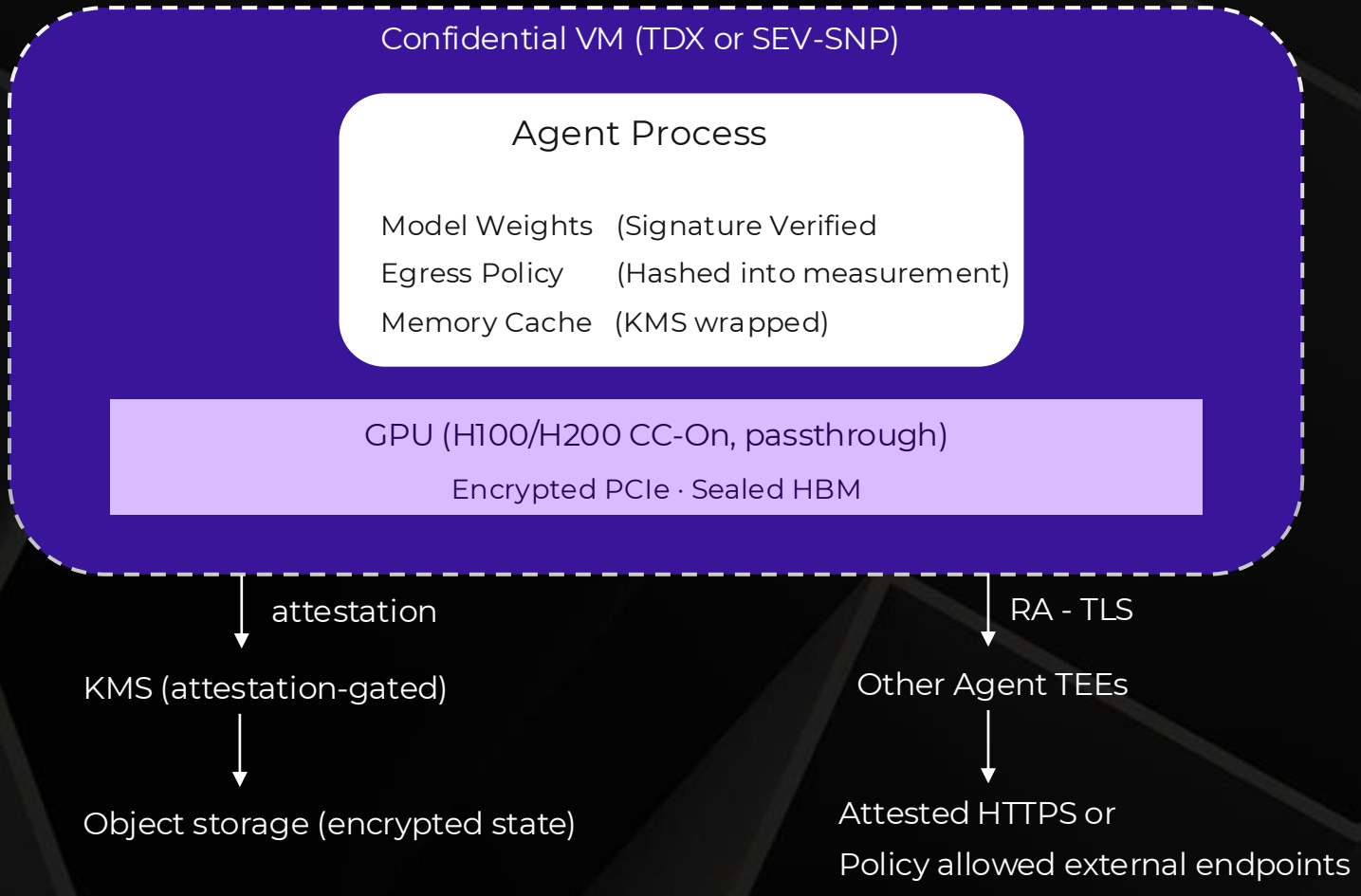
Long-Lived Memory

Chat history and vector stores outlive the TEE - persistent state must be sealed and re-attested.

Trust Chain Seams

Orchestrator, model weights, prompt injection, and CPU <-> GPU binding each introduce a gap attackers can exploit.

Reference Architecture

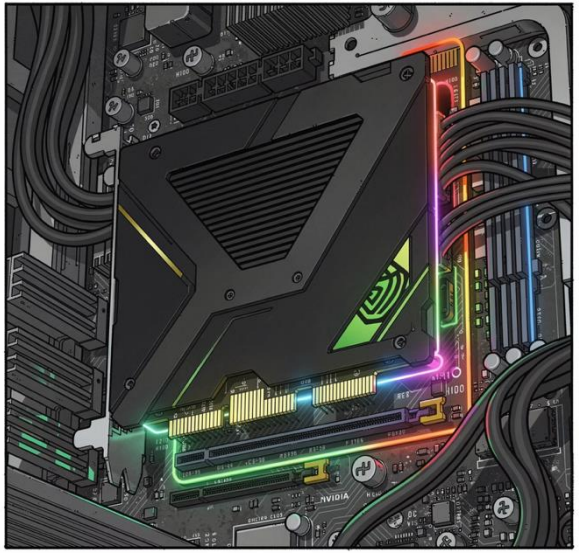


GPUs are a different problem

Why GPUs Are Harder

CPU memory lives inside the chip package. GPU memory lives **across a PCIe bus** - a wire on the motherboard that is physically much easier to probe than memory inside a CPU package.
NVIDIA's answer on H100/H200 has three parts working together to close that attack surface.

- GPU memory hardware-firewalled
  Only the GPU itself can read its own HBM.
- PCIe traffic encrypted
  Keys established at attestation time between CPU and GPU.
- GPU has its own attestation key
  Burned in at the factory, verified via NVIDIA Remote Attestation Service (NRAS).



GPU Confidential Computing - What's Protected and What Isn't

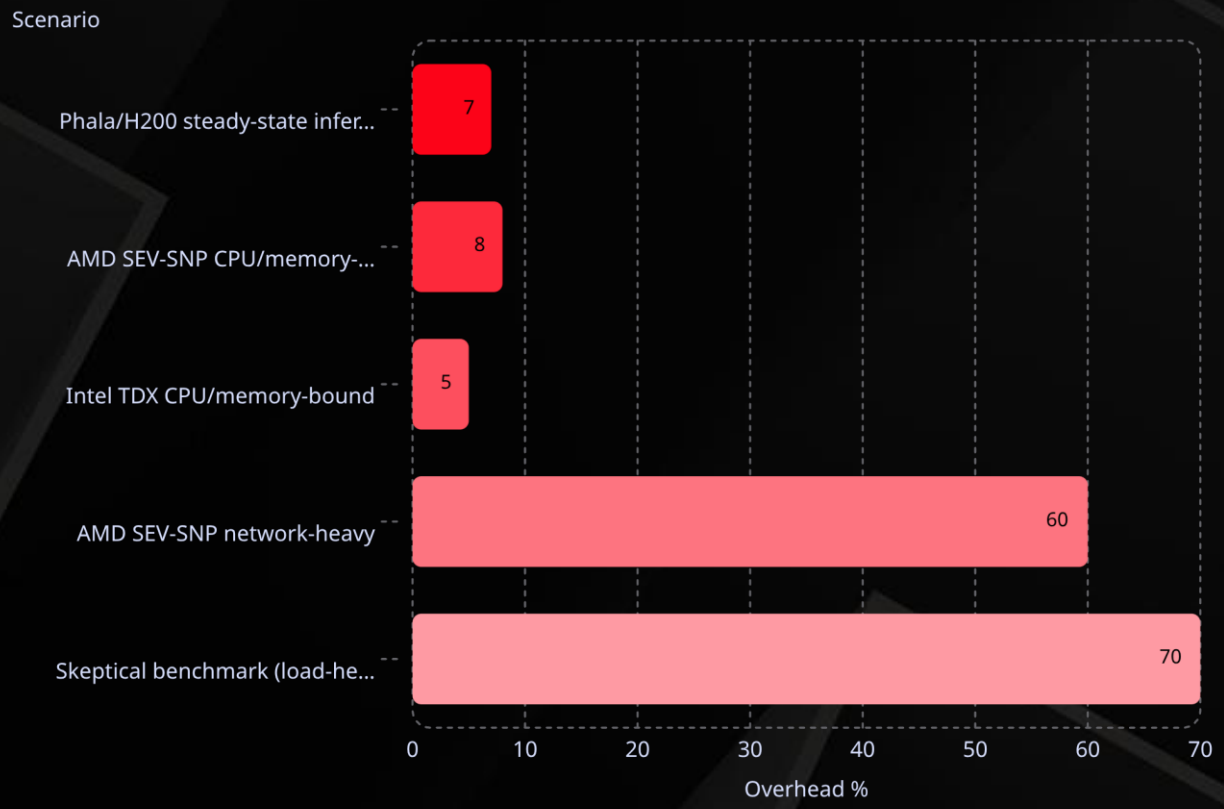
Protected on H100/H200

- GPU memory
- PCIe traffic CPU⇿GPU
- GPU firmware

Not Protected on Hopper

- **NVLink between GPUs** - huge for multi-GPU LLMs
- MIG partitioning (CC is single-tenant only)

Performance: Vendor vs. Independent



0–7% for steady-state large-model inference; 45–70% for load-heavy workloads. Measure yours.

CPU CVM Performance

AMD SEV-SNP

2–8% overhead for CPU/memory-bound workloads. Up to **60%** for network-heavy.

Intel TDX

~5% for CPU/memory-bound per Intel's own figures. Same I/O penalty as SEV-SNP on network-heavy paths.

Intel SGX (classic)

Highly variable. **2–3× slower** for memory-heavy workloads due to the small enclave memory budget forcing expensive paging.

AWS Nitro Enclaves

Minimal raw CPU overhead. Significant **engineering cost** - no network stack inside the enclave, only a private vsock channel to the parent VM.

What Others Have Built (1 of 2)

Apple Private Cloud Compute

Anthropic Confidential Inference

What Others Have Built (2 of 2)

Azure Confidential Inferencing

OHTTP + HPKE · SEV-SNP CVM + H100 CC · keys released only on attestation matching ledger-registered policy · signed SBOMs

Self-Hosted Open Source

Edgeless Contrast

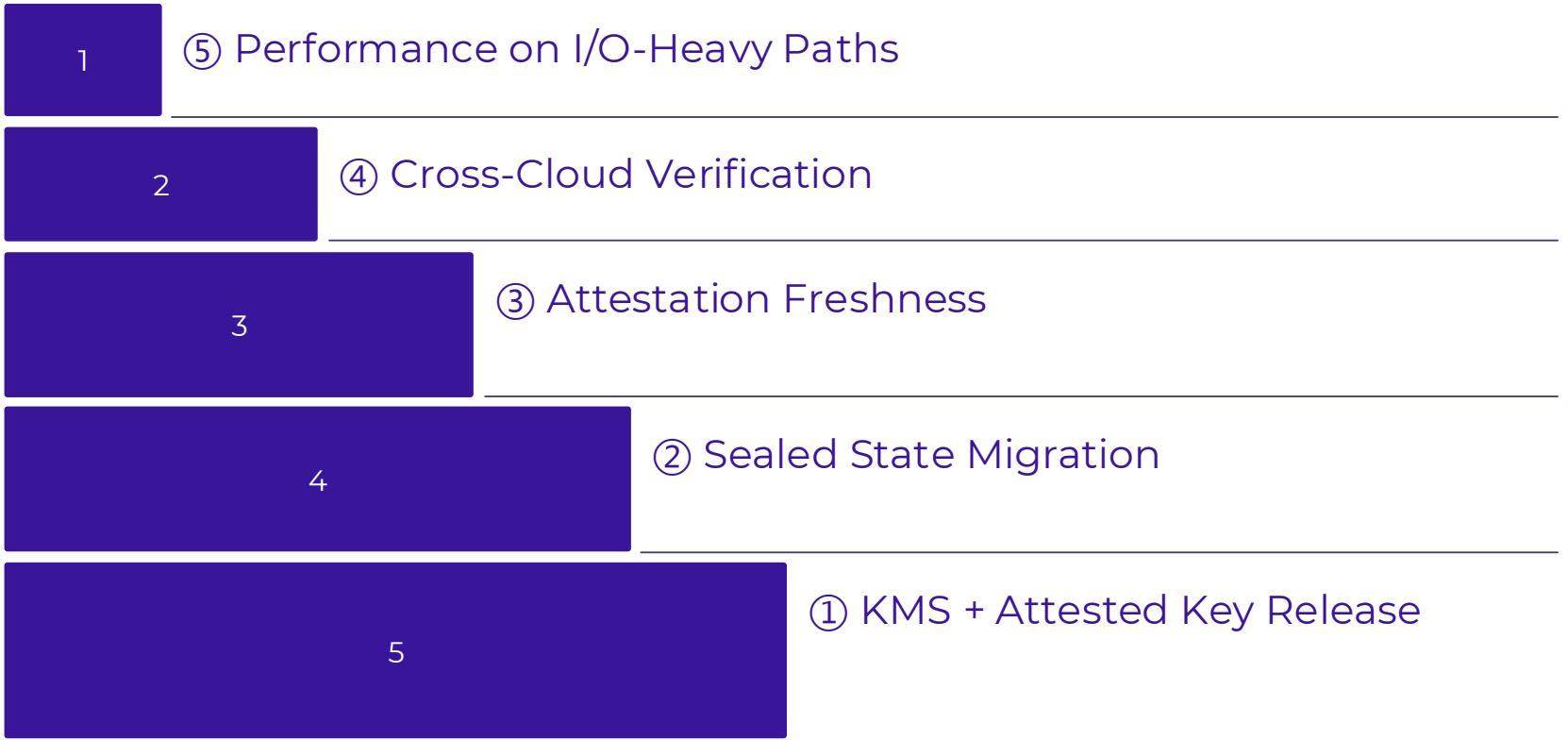
Confidential containers for AKS, GKE, and bare metal.

Phala dstack

Open-source confidential container framework..

Standards layer: Confidential Computing Consortium (CoCo, Veraison, Gramine)

The Hard Parts, Ranked by How Often They Bite You



"Isn't this just security theater?"

Partly. The CPU vendor is still in your TCB. If you assume a nation-state with supply chain capability, TEEs don't protect you. What you're getting is a **smaller, better-defined.**

"What about side channels?"

Real and ongoing. SGX has been broken many times. SEV-SNP and TDX too. The honest standard isn't "no side channels" - it's **public record of disclosures plus a clear patch process**.

🎛️ "Why not FHE or MPC instead?"

FHE: 4–5 orders of magnitude slower than plaintext. Not viable for LLMs in 2026.

MPC: 2–4 orders of magnitude. Specific primitives only.

TEEs: near-native performance with cryptographic confidentiality.

🎯 "Prompt injection inside the TEE?"

Not solved cryptographically. The TEE protects bytes, not meaning.

🪛 "What when a CVE drops?"

TCB Recovery: re-attest under new firmware, re-release keys. Build for this from day one.

Thank You

NUTANIX