Privacy Preserving Solid LLMs

Solid Symposium 2025

Davi Ottenheimer
VP Trust and Digital Ethics

“It’s TEE Time”

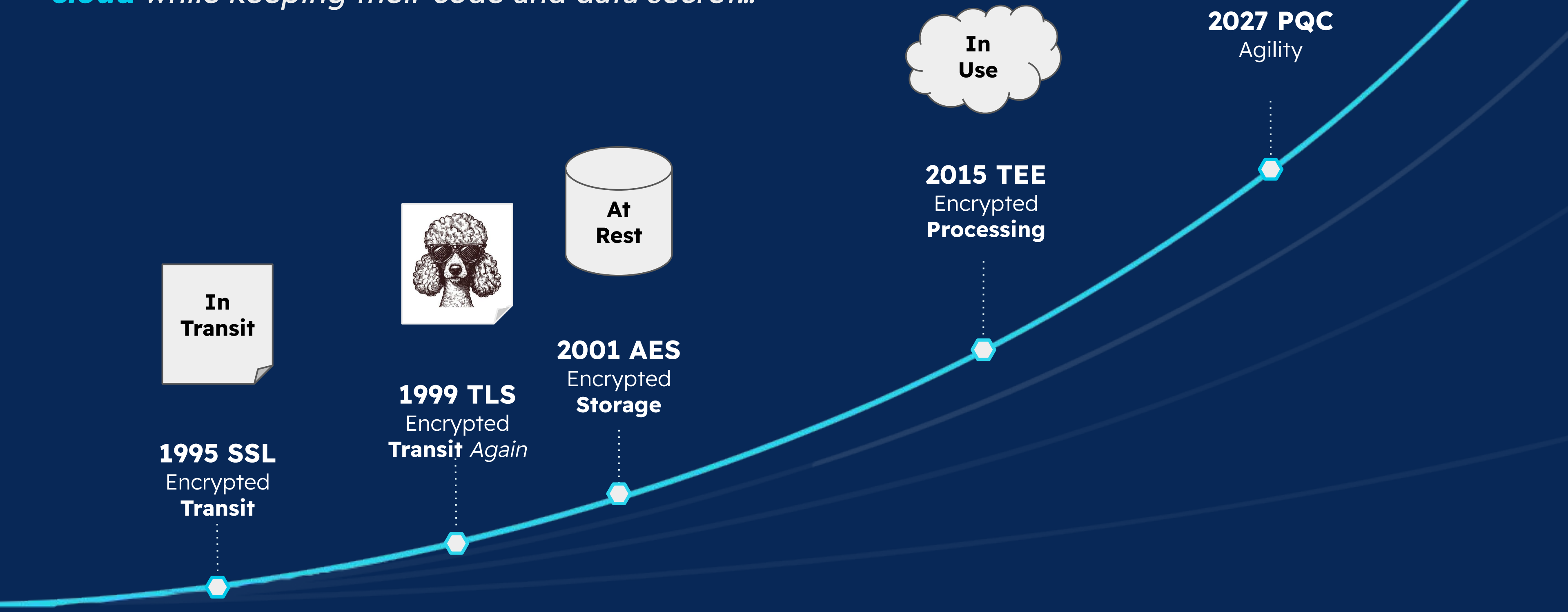
Kermit the Frog
Privacy Thought Leader

inrupt

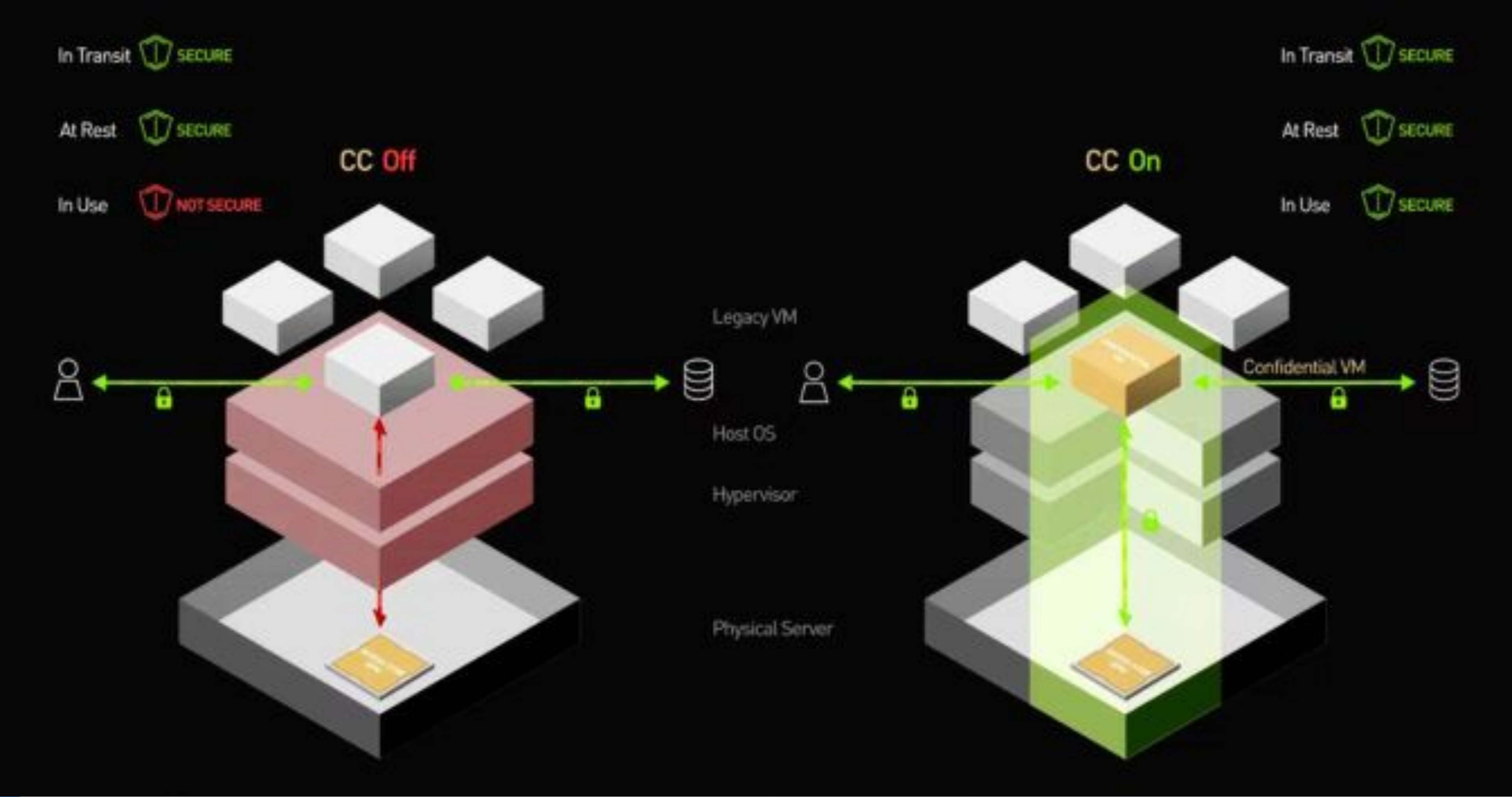
Verifiable Confidential Cloud Computing

https://ieeexplore.ieee.org/document/7163017

“…run distributed MapReduce ***computations in the cloud*** *while keeping their code and data secret…”*



inrupt



Source: NVidia

Terminology

Enclaves, Containers, Pods, Vaults, Buckets, Wallets, Boxes, etc

Personal Data Storage (Pods)

- Pods as W3C standards
- User-controlled with granular access permissions
- Data sovereignty emphasis (independent of ID)

Confidential Computing (CC) and Trusted Execution (TE) Environments

- Hardware-based abstraction and isolation for processing
- Runtime memory encryption protects “data in use”
- Uses "secure enclave" even in untrusted environments

Terminology

Semantics of Emerging Digital Identity

- Pods are the Personal Data Storage open standard
- Wallets are a storage ecosystem affiliated with a value system
- Vaults refer to storage services for sensitive and “holistic” personas

late 14c., *walet*, "**bag, knapsack, large purse**," especially one used by travelers, a word of uncertain origin... Germanic word in Anglo-French or Old French, from Proto-Germanic **wall-* "roll".

A "flat pocket-case for paper money" recorded 1834, **American English**.

Problem Statement

BIG TECH

- ***Forced Exposure:*** AI personalization demands handing over sensitive personal data
- ***False Consent:*** "Better AI = Less Privacy" assumption traps users in bad choice (false consent)
- ***Known Gaps:*** Vulnerabilities exist across the entire data lifecycle

BETTER TECH

- Verifiable security guarantees across all data states (rest, transit, and use)
- AI that respects data boundaries while maintaining high utility
- End-to-end attestation that proves the system works as promised

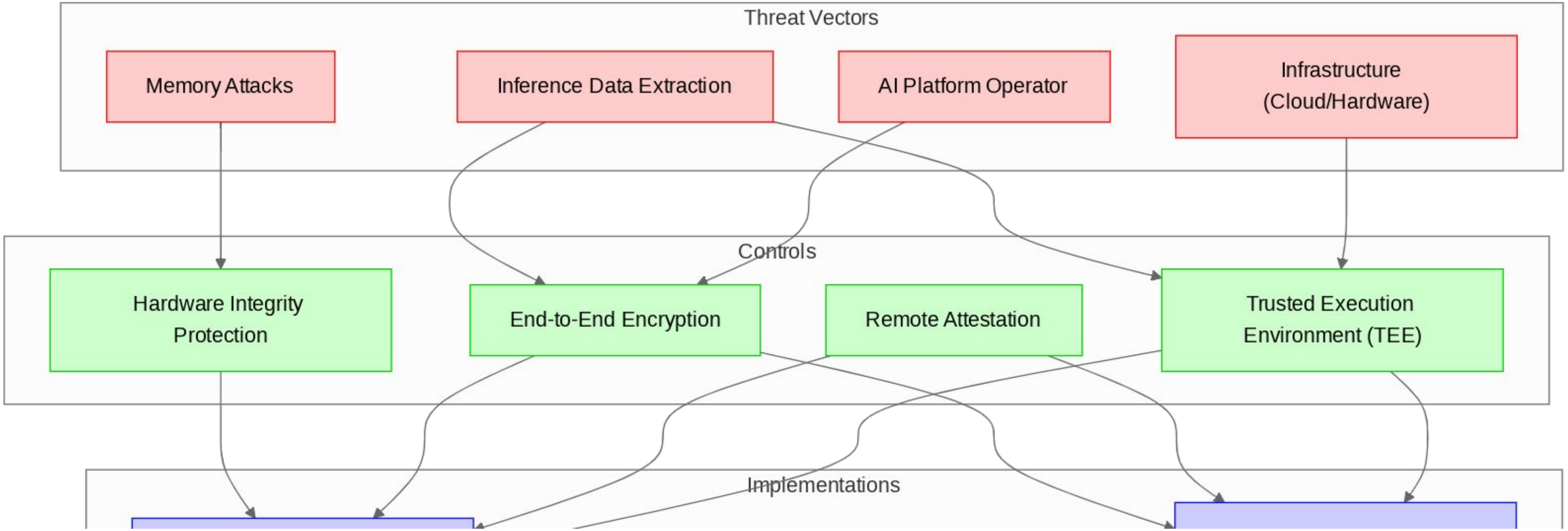
Security Analysis

Threat Model Coverage (Complexity Balance With Usability)

- Protection against infrastructure provider (cloud, hardware)
- Defense against processing data exposure (AI platform operator)
- Prevention of inference data extraction attacks
- Mitigation of memory attacks through hardware integrity protection

Security Analysis

Threat Model Coverage



Future Work

Known Limitations

- Implementation complexity in Kubernetes environments
- Limitations in Azure attestation (MAA) *requires* 3rd-party solution
- Independent key management should be separate from cloud option
- Optimization of deployment and performance across providers and hardware
- Memory constraints with single-GPU implementation (~70B parameters maximum)
- Performance tests: 15-20% for memory encryption operations (vendors claim less)
- Hardware compatibility (Azure AKS with AMD SEV-SNP and NVIDIA H100)

Future Work

Roadmap

- Multi-GPU support enabling larger models and faster inference
- Field-level encryption of routing metadata
- Chip Neutral (Intel)
    - Attestation protocol (DCAP vs. AMD SEV-SNP)
    - Memory allocation constraints (enclave size limitations)
    - Modified key management for Intel enclaves
- Cloud Neutral (Google)
    - Confidential GKE modifications
    - AMD SEV support in GCP N2D instances
    - TEE verification through GCP Attestation Service

Architecture

Data Wallet With Privacy-Preserving Processing

Architecture

Chain-of-Trust Considerations

1. Runtime memory protection utilizing AMD SEV-SNP capabilities, providing hardware-enforced isolation with integrity verification
2. Component integrity verification through measured boot process and cryptographic attestation chains with known limitations in verification scope
3. Layered security approach combining hardware-level memory encryption with application-level cryptographic protections

Architecture

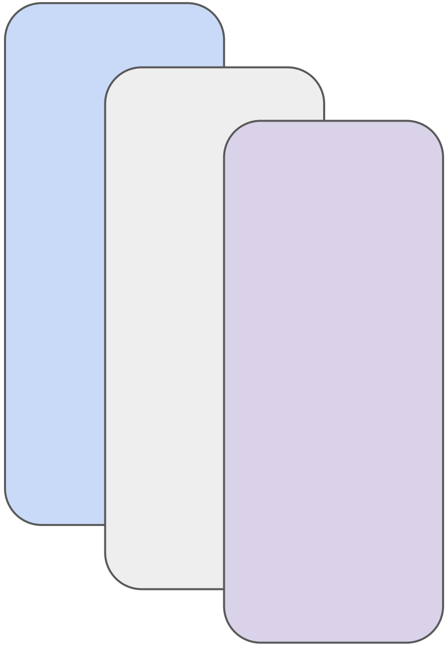
Data Flow Considerations

1. Remote attestation mechanism for hardware and software integrity verification, utilizing hardware-based cryptographic measurements
2. Secure key exchange protocol using hardware-root with isolation
3. Application-level encrypted communications using authenticated encryption schemes (e.g., AES-GCM)
4. Proposed secure memory boundaries for GPU compute operations with confidential computing extensions
5. Complete encrypted data path including response re-encryption within the protected execution environment

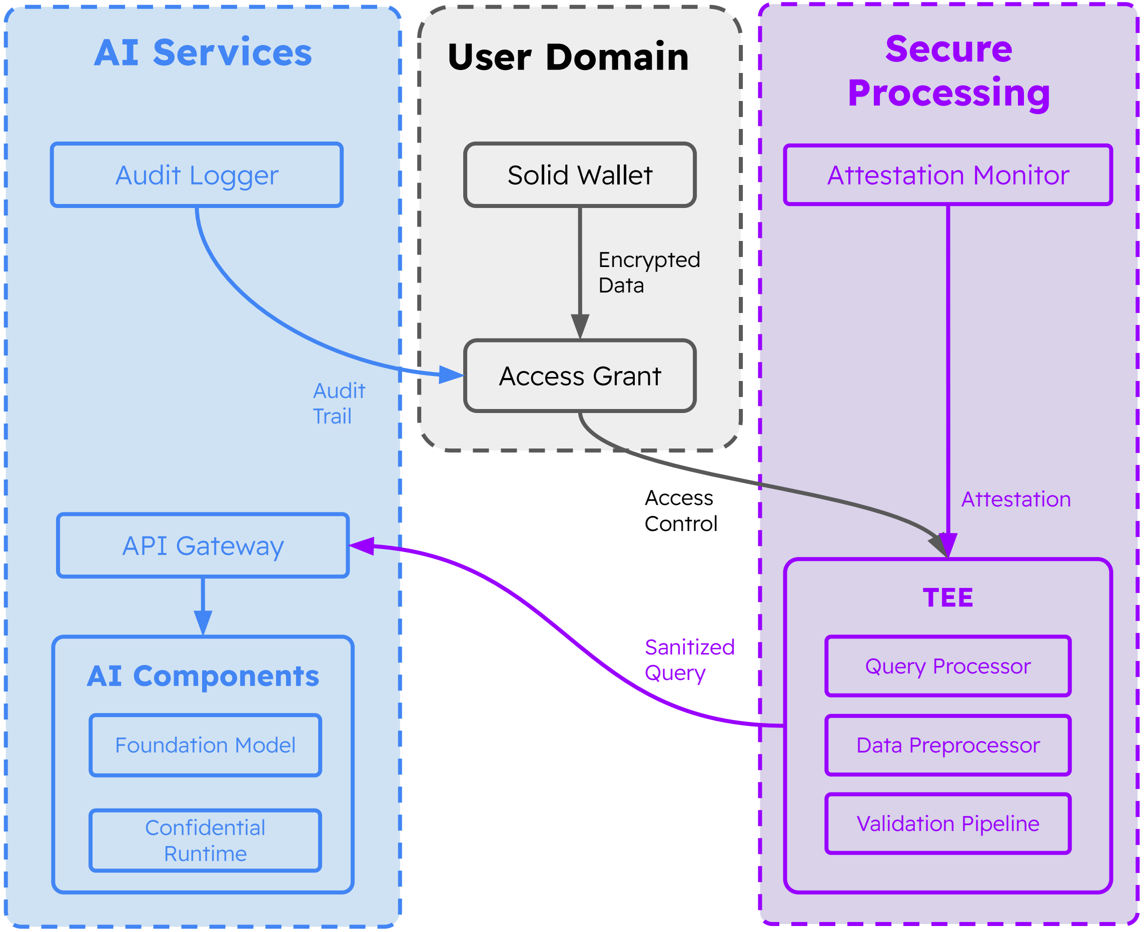
Architecture

Three Abstractions

1. **AI Service Layer:** Isolated inference service with verified attestation chains
2. **User Data Layer:** Solid for secure data storage using W3C protocol and granular access controls
3. **Processing Layer:** Confidential Computing Environment with hardware-backed TEE



Three Abstractions



inrupt

Implementation

Data Wallet With Privacy-Preserving Processing

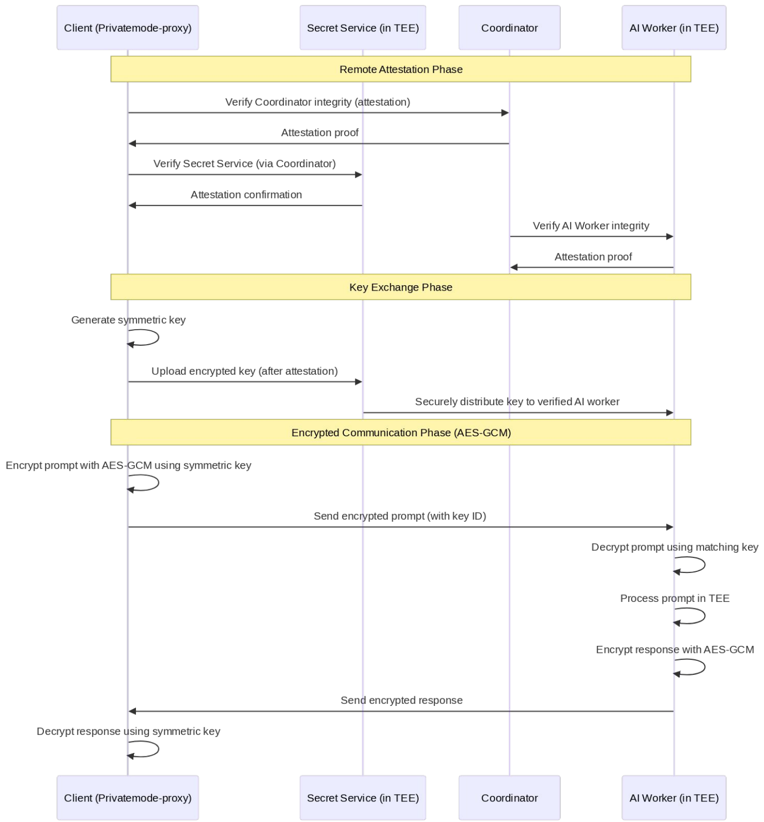
Implementation

TEE Details (Microsoft Azure March 2025)

- **AMD SEV-SNP** for memory encryption with integrity protection
- **NVIDIA H100** confidential computing capabilities for GPU workloads
- Container isolation with verified attestation chain
- DM-Verity for continuous disk integrity verification

AMD	intel	NVIDIA	intel
DCasv5 & ECasv5 AMD SNP CVMs Generally available DCasv6 & ECasv6 Gated preview	DCsv2 & DCsv3 Intel SGX VMs Generally available	NCCH100v5 VMs NVIDIA GPUs Generally available	DCesv5 & ECesv5 Intel TDX CVMs Public preview

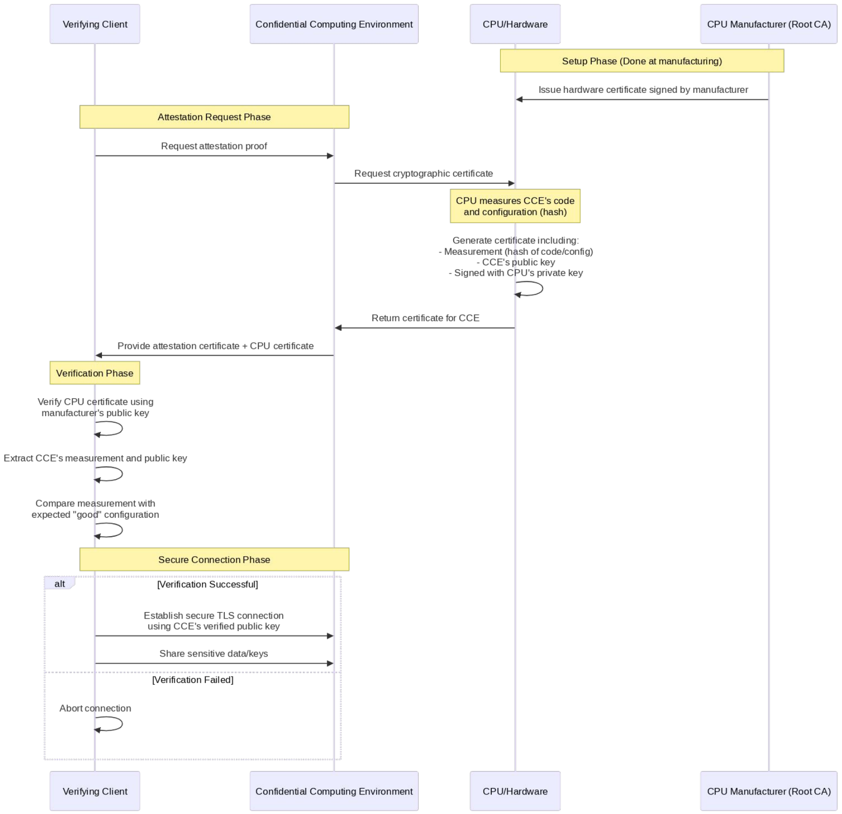
Implementation



Encryption

- AES-GCM authenticated encryption at application layer (API)
- AMD SEV-SNP memory encryption at hardware layer
- Key rotation protocols and separation of encryption domains

Implementation



Remote Attestation Protocol

- Hardware-based cryptographic measurements establishing root of trust
- Multi-stage verification chain: client → coordinator → workers → GPU
- Reproducible builds enabling transparent code verification
- Challenge-response protocol preventing replay attacks

inrupt

Implementation

TEE Service Component for AI

- Privatemode-proxy: Trust anchor for attestation, encryption, and authentication
- Contrast Coordinator: attestation service running in CCE on Azure AKS
- Secret Service: KMS with attestation-based authentication
- AI Workers: Confidential containers in Azure DC-series VMs with NVIDIA H100 integration
- Network: Azure Virtual Network with dedicated subnet
- Storage: Confidential PostgreSQL with hardware-level encryption

Deployment

Data Wallet With Privacy-Preserving Processing

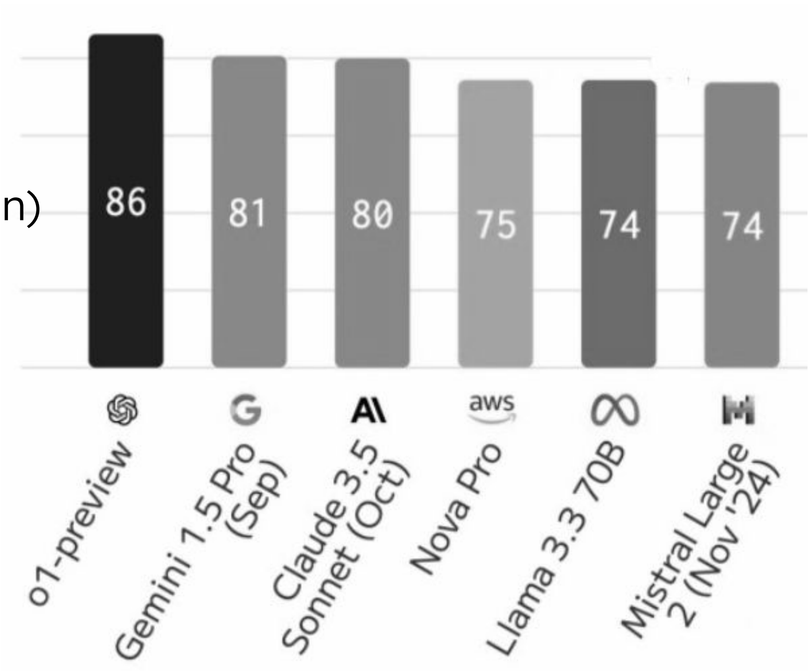
Deployment

LLM on a TEE as a Service

Immediate access via the Privatemode API

- 1 million total tokens (prompt + completion)
- 20,000 prompt tokens per minute
- 10,000 completion tokens per minute
- 20 requests per minute

Llama 3.3 70B within TEE
OpenAI-compatible API



Deployment Choice

A. INTEGRATED

- Containerized ESS microservice deployment in Azure AKS
- Contrast/Edgeless framework on Azure with DC-series VMs
- Privatemode-proxy as sidecar container integrated with ESS Pod
- Direct communication via OpenAI-compatible API
- Complete attestation chain using third-party verification instead of MAA

B. EXTERNAL TEE (HYBRID)

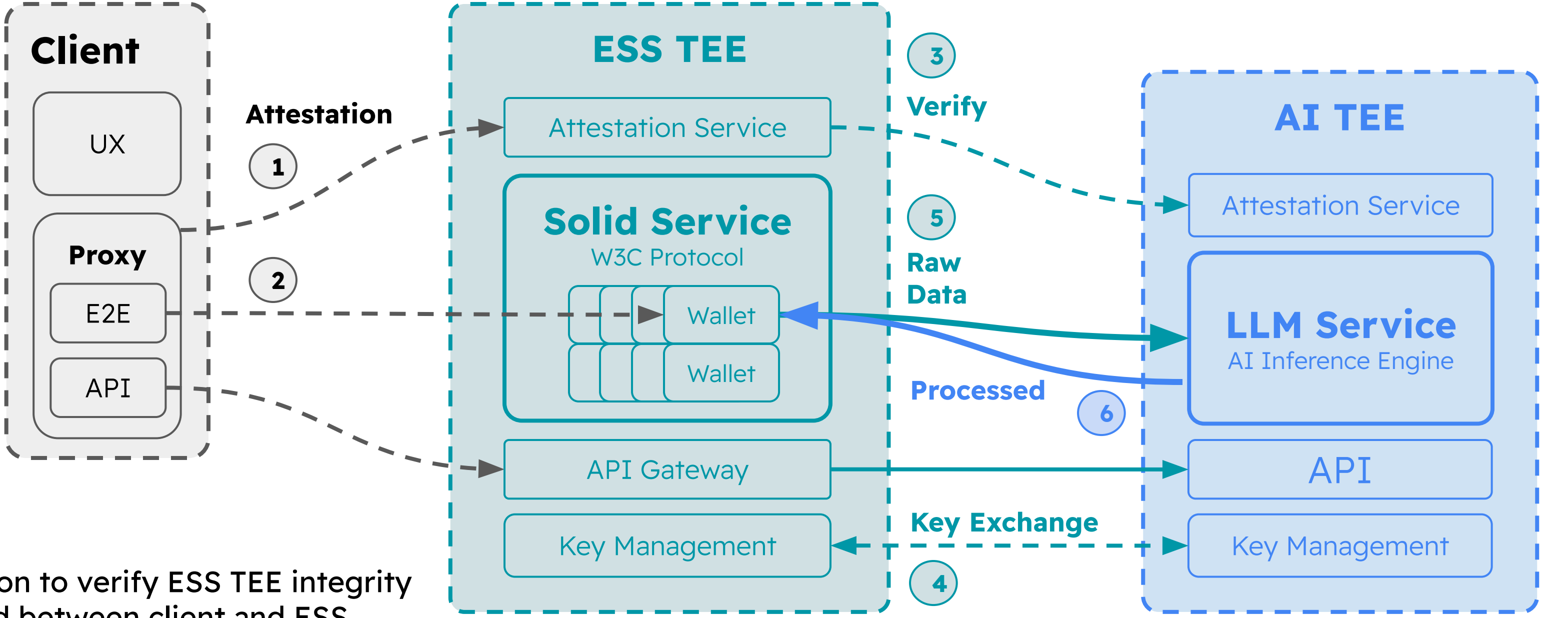
- Standard ESS deployment in Azure without TEE requirements
- Privatemode-proxy deployed via Helm chart in Azure Kubernetes
- Connection to external Privatemode service for confidential AI
- Data remains encrypted outside the verified TEE

Deployment

A. INTEGRATED – Deploy Steps

1. Kubernetes deployment of ESS container
2. Contrast configuration for Azure confidential containers
3. Sidecar privatemode-proxy integration
4. API endpoint configuration and authentication setup

Solid (ESS) and AI Sitting in a TEE



1. Client attestation to verify ESS TEE integrity
2. E2E established between client and ESS
3. ESS TEE verifies AI TEE through attestation chain
4. Secure key exchange between trusted environments
5. Raw data flows only between verified secure environments
6. Processed results returned through secured channel

Client Proxy

Solid clients access transparent TEE attestation

1. **Attestation Verification:** Step 1 when a client initiates attestation with ESS TEE. It sits between Solid apps and ESS, leveraging existing Solid tokens, to verify the integrity and authenticity of server-side components through remote attestation before data exchange is authorized.
2. **E2E (End-to-End Encryption):** Leveraging attestation, the proxy ensures encryption of outgoing data and decryption of incoming responses have knowledge of the ESS TEE.
3. **Authentication Management:** The proxy adds necessary authorization tokens to inference requests, establishing the client's identity and permissions within the secure environment.

Real-World Example

Data Wallet With Privacy-Preserving Processing

Example

Health Data Processing App

- Solid storage of health information
- API for secure AI inference
- Personal insights from LLM without data exposure
- Verifiable processing for technology-based (measurable) regulatory compliance

Implementation

- User-controllable processing boundaries (Solid)
- Privatemode-proxy for API access
- Data processing within verified TEE

Privacy Preserving Solid LLMs

Davi Ottenheimer
VP Trust and Digital Ethics