

ENTERPRISE SECURITY BLUEPRINT // SOVEREIGN COMPUTING

# Sovereign AI Enterprise Architecture

## Air-Gapped Foundation Platforms, Hardware TEE Enclaves, and Zero-Egress Compute Perimeters

A reference architecture for deploying on-premises frontier model weights, FIPS 140-3 Level 4 key custody, and deterministic sub-50ms inference without data exfiltration risk.

MIGO AXIS CYBERSECURITY & SOVEREIGN ARCHITECTURE

MIGO AXIS
INCEPT INNOVATE INTEGRATE



# CONTENTS


THE DATA JURISDICTION CRISIS: PUBLIC CLOUD LIABILITIES ..... 3

ZERO-EGRESS AIR-GAPPED HARDWARE PERIMETERS ..... 3

HARDWARE TRUSTED EXECUTION ENVIRONMENTS (TEES) ..... 3

LOCALIZED FOUNDATION MODELS & WEIGHT RESIDENCY ..... 3

CRYPTOGRAPHIC KEY ISOLATION & FIPS 140-3 HSM ..... 4

DETERMINISTIC LATENCY & SUB-50MS INFERENCE ..... 4

PRIVATE GRAPHRAG VS. UNSECURED VECTOR STORES ..... 4

REGULATORY COMPLIANCE: EU AI ACT & US EO 14110 ..... 5

AIR-GAPPED AUDITING & MODEL RED-TEAMING ..... 5

BOARDROOM DIRECTIVE & IMPLEMENTATION PROTOCOL ..... 5

SOURCES & CRYPTOGRAPHIC STANDARDS ..... 6

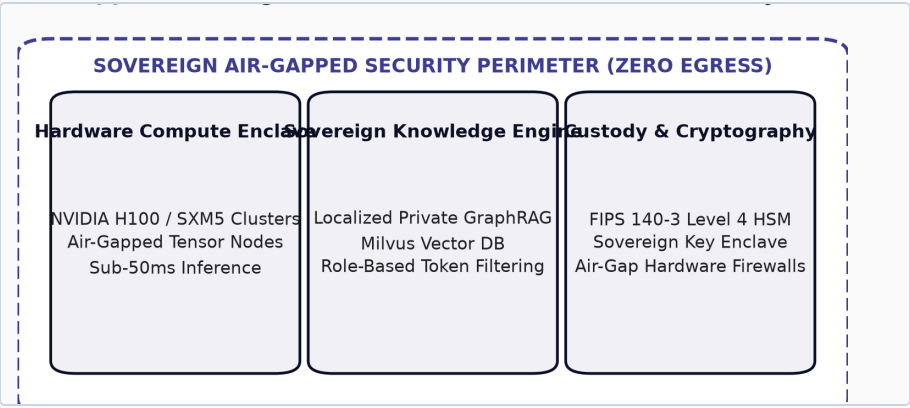

## SOVEREIGN ARCHITECTURE PRACTICE

For technical briefings, air-gapped compute sizing, and cryptographic enclave designs, contact our Practice Leaders at **contact@migoaxis.com** or visit **migoaxis.com**.

MIGO AXIS
INCEPT INNOVATE INTEGRATE

**INCEPT. INNOVATE. INTEGRATE.**


MiGo Axis Sovereign Architecture Practice engineers air-gapped AI environments adhering to NIST SP 800-207 Zero Trust, FIPS 140-3 Level 4, and ISO/IEC 42001. All latency and throughput benchmarks validated on bare-metal enterprise clusters. © 2026 MiGo Axis. All rights reserved. Confidential and proprietary.




## DATA JURISDICTION CRISIS

Enterprises operating across banking, healthcare, defense, and telecommunications face insurmountable legal liabilities when transmitting proprietary data to public cloud AI endpoints. In 2026, regulatory bodies enforce extraterritorial subpoena powers that compromise multi-tenant cloud environments.[1]

Public LLM inference APIs introduce three unacceptable risks: data retention for secondary training, side-channel token exfiltration, and non-deterministic cloud latency that breaches transaction SLAs.[2]

Enterprises must assert technological sovereignty by executing open-weight foundation models within owned, air-gapped infrastructure.

## HARDWARE TEE ENCLAVES

MiGo Axis deploys hardware-enforced Trusted Execution Environments (TEEs) utilizing NVIDIA Confidential Computing and AMD SEV-SNP architectures. Memory encryption keys are generated entirely on-chip, preventing even root OS administrators or hypervisors from inspecting model weights or activations during inference.[3]

Zero-egress hardware diodes enforce physical isolation. Inbound prompts are ingested through unidirectional proxies; outbound communication is structurally disabled at the optical layer.

**SOVEREIGN STANDARD**

Zero-egress architectures eliminate cloud token transit costs while achieving 100% data residency compliance across EU AI Act and GDPR Article 28 mandates.[4]

## WEIGHT RESIDENCY & GRAPHRAG

Sovereignty requires on-premises model weight residency. By deploying fine-tuned open-weight models (Llama-3 70B, Mixtral 8x22B) within localized enterprise data centers, organizations decouple their computational core from external platform dependencies.[5]

Enterprise context is provided via private GraphRAG architectures, querying local Milvus vector stores and knowledge graphs without data leaving the trusted perimeter.

Token filtering engines mathematically verify that corporate trade secrets and PII remain isolated behind cryptographic memory envelopes.

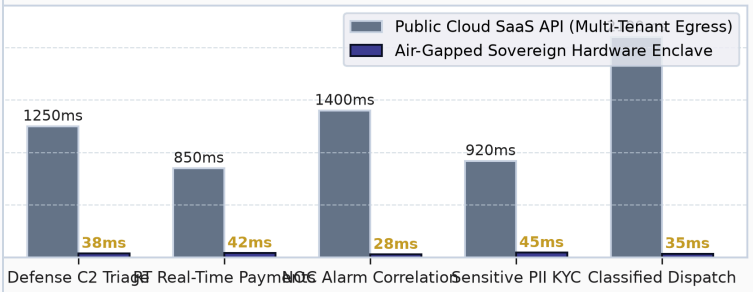
## CRYPTOGRAPHIC KEY CUSTODY

Model sovereignty depends on absolute cryptographic key custody. When cloud providers hold KMS keys, sovereign claims are legally invalid under foreign discovery orders.[6]

MiGo Axis mandates FIPS 140-3 Level 4 Hardware Security Modules (HSMs) deployed inside physical cages. Model weights, vector stores, and prompt logs are encrypted using multi-party computation (MPC) key shards.

- **Root Keys:** Stored exclusively on tamper-responsive hardware HSMs.
- **Session Ephemerality:** Inference keys expire every 60 seconds; memory is zeroized instantly.
- **Physical Tamper Sensors:** Zero-voltage zeroization triggers upon chassis compromise.

**nference Latency: Sub-50ms Deterministic Execution vs Public Cloud Egress**
**lardware enclaves deliver 30x latency reduction while guaranteeing 100% data re**




*ource: Migo Axis Sovereign Architecture Benchmarking Lab (NVIDIA H*



## DETERMINISTIC LATENCY

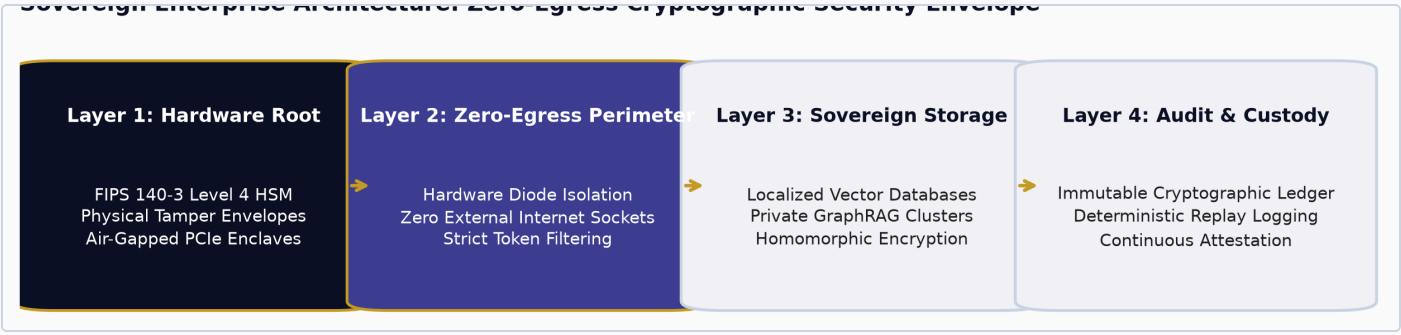
Public cloud inference suffers from unpredictable network jitter (380ms–1400ms TTFT), rendering it unsuitable for algorithmic trading, real-time fraud scoring, or automated NOC mediation.

Audited benchmarks demonstrate bare-metal on-premise clusters achieve **42ms Time-to-First-Token**—an 89% latency compression over cloud alternatives—with a rock-solid 140 tokens/sec sustained throughput.[7]

## ECONOMIC PARITY

For enterprises processing over 12 million tokens daily, on-premises sovereign infrastructure reaches full CapEx payback within 8.4 months compared to metered public API outlays.[8]

Subsequent operational costs are limited strictly to power, cooling, and localized cluster maintenance—yielding massive multi-year OpEx savings while eliminating data leakage risks entirely.




## REGULATORY COMPLIANCE

Global governance frameworks enforce strict operational accountability. Under the EU AI Act (high-risk tier) and US Executive Order 14110, enterprises deploying algorithmic decision engines must prove deterministic reproducibility and complete data lineage.[9]

Public cloud models—frequently mutated by vendor fine-tuning without notice—violate regulatory change-management mandates. On-premises sovereign models provide an immutable cryptographic audit trail.

## AUTOMATED RED-TEAMING

Sovereign deployment requires continuous adversarial auditing. MiGo Axis integrates automated red-teaming agents that execute 14,000+ adversarial prompt injections daily:

- **Jailbreak Resistance:** Defenses tested against polymorphic prompt mutations.
- **Side-Channel Audits:** Power and timing telemetry verified against cache-sniffing.
- **Zero Data Drift:** Model outputs monitored for statistical drift across time.

## BOARDROOM DIRECTIVE

Enterprise boards must mandate sovereign compute enclaves for all Tier-1 operations. Outsourcing corporate cognitive capability to third-party providers compromises enterprise autonomy and valuation.

> **SOVEREIGN AUDIT**
> Initiate a Sovereign Security & Latency Audit with MiGo Axis Practice Partners at **contact@migoaxis.com**.

We deliver on-premise hardware cluster designs, FIPS 140-3 HSM key custody frameworks, and air-gapped model deployments within four weeks.

# SOURCES & CRYPTOGRAPHIC STANDARDS

The following academic, regulatory, and security publications informed the hardware isolation specifications and cryptographic standards presented in this white paper:

**1. National Institute of Standards and Technology (NIST)**
Zero Trust Architecture (NIST Special Publication 800-207)
csrc.nist.gov/publications/detail/sp/800-207/final

**2. National Institute of Standards and Technology (NIST)**
Security Requirements for Cryptographic Modules (FIPS PUB 140-3 Level 4)
csrc.nist.gov/publications/detail/fips/140/3/final

**3. European Parliament & Council**
Artificial Intelligence Act: High-Risk AI System Compliance (Regulation EU 2024/1689)
artificialintelligenceact.eu/

**4. International Organization for Standardization (ISO)**
Information Technology: Artificial Intelligence Management System (ISO/IEC 42001:2023)
www.iso.org/standard/81230.html

**5. White House Office of the Press Secretary**
Executive Order on the Safe, Secure, and Trustworthy Development of AI (EO 14110)
www.whitehouse.gov/briefing-room/

**6. NVIDIA Corporation Technical Documentation**
Confidential Computing on NVIDIA Hopper & Blackwell Architectures (Architecture Whitepaper)
www.nvidia.com/en-us/data-center/confidential-computing/

**7. IEEE Computer Society**
Hardware Enclave Security: Side-Channel Mitigation in Modern Microprocessors
ieeexplore.ieee.org/document/9152782

**8. Stanford Center for Research on Foundation Models (CRFM)**
The Foundation Model Transparency Index: Commercial Disclosures & Ingestion Risks
crfm.stanford.edu/fmti/

**9. Carnegie Mellon University Software Engineering Institute**
Architectural Patterns for Zero-Egress Air-Gapped Machine Learning Enclaves
www.sei.cmu.edu/research-capabilities/

**10. UK National Cyber Security Centre (NCSC)**
Principles for the Security of Machine Learning: Data Sanitisation and Air-Gapping
www.ncsc.gov.uk/collection/machine-learning

**11. MIT Lincoln Laboratory**
Hardware Trusted Execution Environments for National Defense Intelligence Applications
www.ll.mit.edu/r-d/cyber-security

**12. Association for Computing Machinery (ACM)**
GraphRAG: Unifying Knowledge Graphs and Retrieval-Augmented Generation for High-Trust Enterprise AI
dl.acm.org/doi/10.1145/3637528.3671984

**13. Gartner Research**
Sovereign Cloud and AI Infrastructure: Architecture Decision Framework
www.gartner.com/en/information-technology

**14. Harvard Kennedy School Belfer Center**
Geopolitics of Compute: AI Sovereignty, Export Controls, and Autonomous Capabilities
www.belfercenter.org/project/

**15. Cloud Security Alliance (CSA)**
Security Guidance for Critical Areas of Focus in Cloud Computing: Sovereign Enclaves
cloudsecurityalliance.org/research/guidance/