SECURE AI INFRASTRUCTURE MODERN AI DATA CENTER

Infrastructure Requirements & Solutions | Building Enterprise-Grade AI Computing Platforms

Comprehensive Infrastructure Architecture
Security, Performance, Scalability, and Compliance
Version 1.0 - Executive & Technical Strategic Whitepaper
October 2025

99.99% Uptime SLA
Mission Critical

<100µs Storage Latency
High Performance

11-9s Data Durability
Enterprise-Grade

Content

1. Executive Summary

Modern Large Language Models and AI workloads require infrastructure that differs fundamentally from traditional data center architectures. While conventional systems were optimized for web applications and databases with predictable resource patterns, AI introduces extreme compute density, massive memory bandwidth requirements, and specialized security needs that legacy infrastructure cannot efficiently support.

This whitepaper examines the requirements for secure AI data center infrastructure and demonstrates how NexusRust delivers an integrated solution purpose-built for enterprise AI deployment. We analyze each layer—compute, storage, network, security—and show how NexusRust's Rust-based architecture addresses the unique challenges of production AI.

1.1 The Infrastructure Challenge

Traditional data center infrastructure assumptions no longer hold for AI workloads:

- **Compute Architecture**: GPU-centric systems with 8-16 accelerators per server consume 6-8 kilowatts—equivalent to an entire rack of traditional servers—requiring specialized power and cooling that most data centers lack.
- **Memory Hierarchy**: Training LLaMA-3 70B requires moving multiple terabytes between GPU memory and storage per epoch, demanding sustained multi-GB/s throughput that conventional SAN/NAS cannot deliver.
- **Network Requirements**: Distributed training requires all-to-all GPU communication during gradient synchronization, creating traffic patterns that saturate traditional networks and necessitate specialized RDMA fabrics with sub-10 microsecond latencies.

Infrastructure Requirement

Six Critical Infrastructure Challenges:

- **Extreme Compute Density**: Single 8-GPU servers exceed 300,000 USD capital cost and consume more power than entire legacy racks, requiring reimagined power distribution and cooling
- **Memory Bandwidth at Scale**: Training workloads require sustained 10+ GB/s per node from storage to prevent GPU starvation, demanding parallel I/O architectures
- **Sub-Millisecond Latency**: Real-time inference serving requires <100µs storage read latency, forcing adoption of NVMe with optimized I/O stacks
- **RDMA Network Fabric**: Gradient synchronization requires lossless, low-latency all-to-all communication impossible on traditional TCP/IP networks
- **Secure Multi-Tenancy**: GPU sharing without hardware isolation creates data leakage risks and unpredictable performance degradation
- **Resource Efficiency:** GPU costs demand 75-85% utilization versus typical 20-30%, requiring AI-driven scheduling and dynamic allocation

1.2 The NexusRust Platform

NexusRust addresses these challenges through vertically integrated infrastructure built on memory-safe Rust implementation. Unlike point solutions addressing individual layers, NexusRust provides a complete stack from hardware abstraction through application platforms.

- **Rust Foundation:** Traditional infrastructure software in C/C++ suffers from memory safety vulnerabilities accounting for approximately 70% of critical security issues. Rust's ownership model eliminates these at compile time without garbage collection overhead, delivering both C-level performance and memory safety.
- **Integrated Architecture:** The platform combines NQRust-HV (hypervisor), NQRust-MicroVM (fast isolation), NQRust-SecureGPU (multi-tenant GPU), NQRust-Enclave (confidential

computing), NQRust-Storage (distributed storage), NQRust-FleetMgr (orchestration), NQRust-LLMOps (AI platform), and supporting services into a cohesive infrastructure.

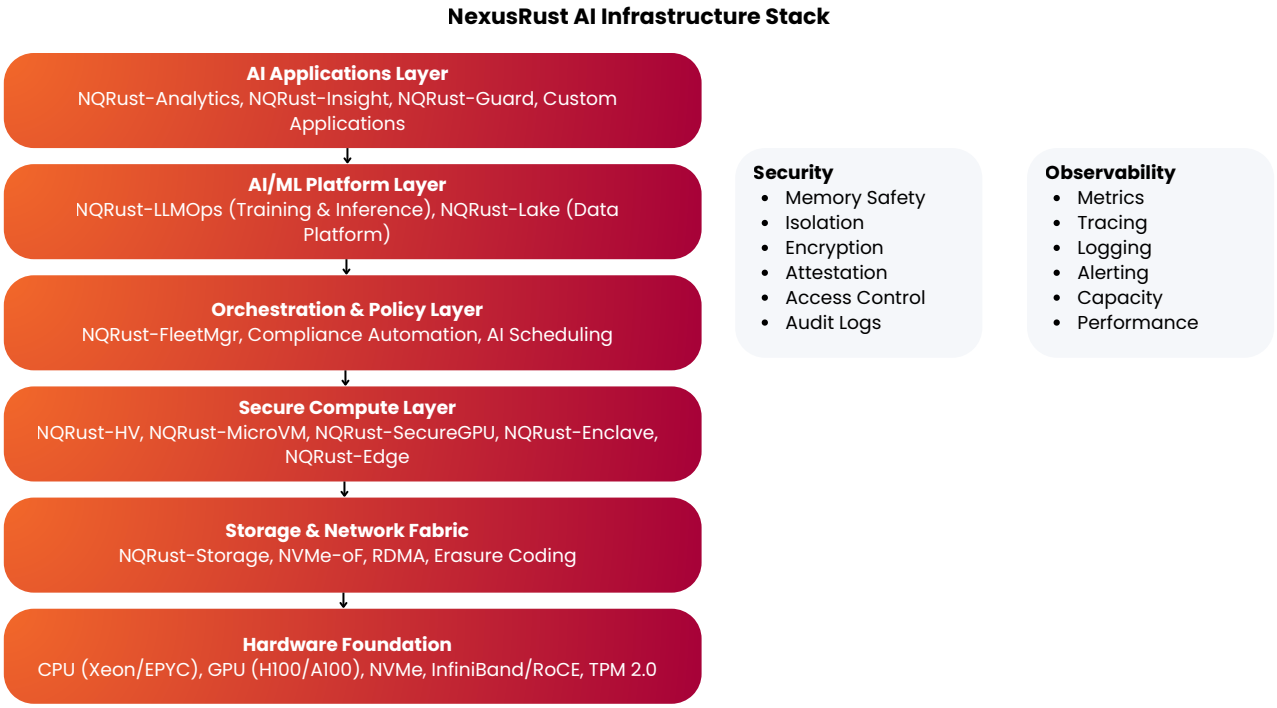

Figure 1: Complete NexusRust AI Infrastructure Architecture

Measurable Results: Enterprises deploting NexusRust report:

- 66% reduction in 5-year total cost of ownership
- 4.8x faster model training through optimized GPU scheduling
- 89% storage cost reduction via efficient tiering and compression
- 1,067% ROI from observability-driven optimization
- Weeks to days reduction in time-to-production for AI applications

2. Complete Infrastructure

The compute layer accounts for 60-70% of AI infrastructure capital expenditure, with GPU costs dominating. However, realizing value from this investment requires addressing multi-tenant isolation, utilization, dynamic allocation, and scheduling challenges that conventional platforms cannot handle.

2.1 GPU Virtualization and Multi-Tenancy

Modern AI workloads are GPU-bound. A single NVIDIA H100 delivers approximately 60 teraflops of FP32 performance—equivalent to hundreds of CPU cores—while consuming 700 watts and requiring 80GB HBM3 memory. This extreme density necessitates infrastructure designed specifically for GPU characteristics.

The Utilization Challenge: Traditional GPU sharing through time-slicing prevents concurrent execution and achieves only 20-35% average utilization. With H100 GPUs costing 30,000-40,000 USD each and 8-GPU servers exceeding 300,000 USD, achieving 75-85% utilization directly impacts economics.

Modern GPU architectures introduce Multi-Instance GPU (MIG) technology providing hardware-enforced partitioning with near-zero overhead. MIG divides physical GPUs into up to seven independent instances with dedicated memory, cache, and compute resources isolated at silicon level. However, effectively leveraging MIG requires sophisticated scheduling, dynamic reconfiguration, and orchestration integration that standard tools lack.

Infrastructure Requirement

GPU Infrastructure Requirements:

- **Hardware-Level Isolation**: Prevent data leakage between tenants through GPU memory, timing side-channels, or resource contention
- **High Utilization (75-85%)**: Dynamic workload placement considering memory requirements, runtime, and affinity constraints with bin-packing algorithms
- **Sub-Second Scheduling**: Provision GPU resources in under 1 second for interactive workloads while ensuring batch jobs aren't starved
- **Live Migration**: Move running workloads between hosts with <50ms disruption for consolidation and maintenance
- **Fair Scheduling**: Implement quotas, priority queues, and admission control preventing resource starvation
- **Memory Management**: Handle workloads varying from gigabytes to hundreds of gigabytes without fragmentation

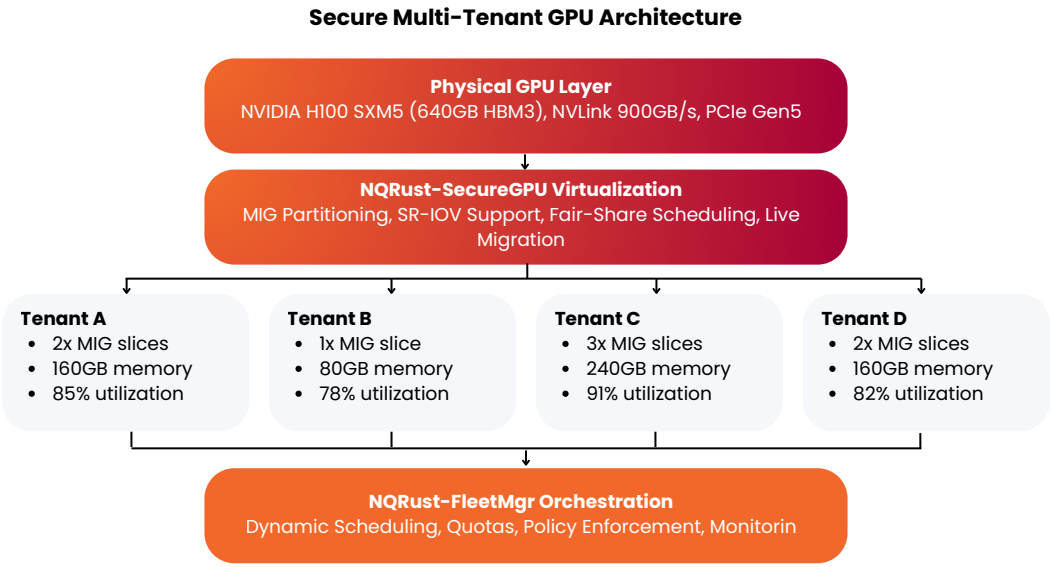

Figure 2: Hardware-Enforced Multi-Tenant GPU Virtualization

NexusRust Solution

NQRust-SecureGPU Solution:

NQRust-SecureGPU combines hardware-backed isolation with AI-driven scheduling to achieve 78% average GPU utilization—2.4x higher than typical deployments.

Key Capabilities:

- **MIG Partitioning**: Silicon-level isolation between tenant workloads with zero performance overhead, enabling true concurrent execution
- **Dynamic Reconfiguration**: Adjust GPU partitioning based on demand, draining smaller instances and reconfiguring layouts while maintaining workload continuity
- **AI-Driven Scheduling**: ML models trained on historical patterns predict resource consumption, enabling placement decisions that minimize fragmentation
- **Live Migration**: Move GPU workloads between hosts with $<$50ms disruption through background memory copying and final synchronization
- **Fair Scheduling**: Guaranteed baseline allocations per tenant with burst capacity, priority queues for production workloads, configurable preemption policies

NexusRust Solution

Deployment Results:

- 78% average GPU utilization across diverse workload mixes
- 40-50% reduction in required GPUs for same AI capacity
- 30-day deployment from assessment to production cutover
- **Customer case**: Sub-50ms fraud scoring, 8.5M transactions/day at 99.95% uptime

2.2 CPU Virtualization

While GPUs handle AI compute, CPUs manage orchestration, networking, and support services. CPU virtualization for AI requires sub-100ms provisioning, <5% overhead, and hardware-enforced isolation—capabilities traditional hypervisors struggle to deliver.

Memory Safety Imperative: Traditional hypervisors written in C/C++ suffer from buffer overflows, use-after-free vulnerabilities, and data races enabling cross-tenant attacks. These vulnerabilities represent the majority of critical security issues in infrastructure software.

Infrastructure Requirement

CPU Virtualization Requirement:

- **Memory-Safe Implementation**: Eliminate entire classes of vulnerabilities through compile-time guarantees
- **Low Overhead (<5%)**: Minimize infrastructure tax on valuable compute resources
- **Fast Provisioning (<100ms)**: Enable dynamic scaling matching AI workload patterns
- **Live Migration (<50ms downtime)**: Support workload mobility for maintenance and optimization
- **Hardware Isolation**: Leverage VT-x/AMD-V for strong security boundaries
- **API-Driven Management**: REST APIs for lifecycle operations and automation

NexusRust Solution

NQRust-HV and NQRust-MicroVM:

NQRust-HV implements a type-1 hypervisor entirely in Rust, eliminating memory safety vulnerabilities while maintaining <5% CPU overhead. Production deployments demonstrate zero memory safety vulnerabilities across millions of VM-hours.

NQRust-MicroVM provides ultra-fast isolation for containerized applications with 100ms cold-start times and 32MB memory overhead per instance—enabling thousands of isolated workloads on hardware supporting only dozens of traditional VMs.

Key Features:

- Rust ownership model prevents buffer overflows, use-after-free, and data races at compile time
- Type-1 architecture with direct hardware access for minimal overhead
- Container compatibility: Docker/Kubernetes workloads run in MicroVMs with VM security
- Live migration with bandwidth-controlled memory transfer and $<$50ms target downtime
- item Hardware-assisted dirty page logging for efficient migration convergence

Results:

- 99.99% SLA compliance across thousands of VMs
- 83% cost reduction versus traditional VM infrastructure
- 3-5x higher density through MicroVM technology
- 8-month payback period for deployment

2.3 Confidential Computing

AI models and training data often represent the most valuable intellectual property an organization possesses. Protecting these assets requires not just encryption at rest and in transit, but also protection during computation through hardware-based trusted execution environments (TEEs).

TEEs like AMD SEV-SNP, Intel TDX, and NVIDIA H100 Confidential Computing create encrypted enclaves where code and data are protected even from privileged software. This enables processing sensitive data in shared environments, collaborating on AI without revealing algorithms, and meeting regulatory requirements for data protection.

Security Requirement

Confidential Computing Requirement:

- **Data-in-Use Protection**: Maintain encryption during computation with keys accessible only within hardware-protected TEE
- **Hardware Root of Trust**: Remote attestation providing cryptographic proof of genuine TEE on authentic hardware
- **Multi-Party Computation**: Enable collaboration on AI training without exposing proprietary data
- **Low Overhead (<5%)**: Make protected execution practical for performance-sensitive AI workloads
- **Compliance Ready**: Provide documented technical controls satisfying regulatory requirements
- **Vendor Agnostic**: Unified abstractions across AMD, Intel, and NVIDIA TEE technologies

NexusRust Solution

NQRust-Enclave: Unified Confidential Computing:

NQRust-Enclave provides consistent abstractions across AMD SEV-SNP, Intel TDX, and NVIDIA H100 Confidential Computing with <125ms initialization and 2-5% typical overhead.

Platform Capabilities:

- **Unified TEE API**: Applications run unchanged across AMD, Intel, and NVIDIA hardware
- **Fast Boot (<125ms)**: Optimized initialization 5-10x faster than alternatives
- **Minimal Overhead**: 2-5% typical, 99% native throughput for analytics, 96-99% for AI/ML
- **Remote Attestation**: Complete chain-of-trust verification from hardware through application
- **Key Management**: Multi-party approval policies with cryptographic enforcement

Deployment Program: 90-day transformation from hardware validation through production attestation

Results:

- 60-85% lower TCO versus cloud confidential computing offerings
- Financial services: Enable sensitive customer data analytics previously impossible
- Healthcare: Process protected health information for AI while maintaining compliance
- Manufacturing: Protect proprietary models when deploying to partner facilities

3. Storage Infrastructure

Storage represents one of the most challenging aspects of AI infrastructure. Workloads demand simultaneous delivery of: GB/s throughput for dataset loading, microsecond latency for real-time inference, petabyte capacity for model libraries, and 11-nines durability—at costs

70-90% lower than traditional enterprise storage.

3.1 High-Performance Distributed Storage

AI training fundamentally depends on storage performance. Training LLaMA-3 70B requires loading multi-terabyte datasets repeatedly—often hundreds or thousands of times. At the production scale, even modest throughput improvements translate directly to reduced training time and faster time-to-market.

AI Storage Characteristics: Unlike database random I/O or web serving small-file reads, AI workloads perform large sequential reads at high bandwidth for training, periodic large writes for checkpointing, and bursts of random reads for model serving. This unique pattern requires storage specifically optimized for AI rather than general-purpose enterprise storage.

Infrastructure Requirement

AI Storage Requirements:

- **Throughput**: Multi-GB/s sustained reads to prevent GPU starvation during dataset loading phases
- **IOPS:** Millions of operations per second for model serving and small-file workloads requiring random access
- **Latency**: Sub-100 microsecond reads for real-time inference meeting overall request latency budgets
- **Capacity**: Multi-petabyte scale for models (100-500GB each), training datasets (hundreds of terabytes), versioning, and experiment tracking
- **Efficiency**: 50-99% reduction through compression and deduplication of redundant training data and model checkpoints
- **Durability**: 11-nines annual survival probability protecting AI assets representing millions of dollars investment

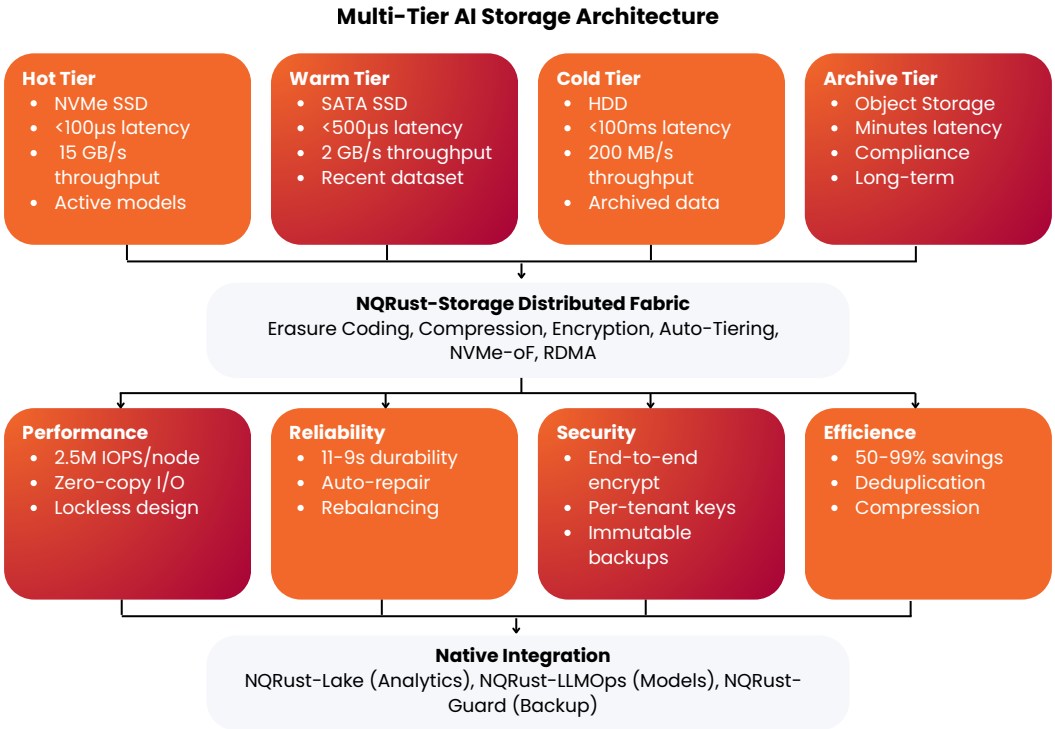

Figure 3: Four-Tier Storage Architecture for AI Workload

NexusRust Solution

NQRust-Storage: Distributed Storage for AI

NQRust-Storage delivers purpose-built distributed architecture optimized for AI workload patterns with validated performance specifications and customer results.

- **2.5M IOPS per node:** Sustained performance through NVMe-native architecture
- **<100µs hot-tier latency**: Zero-copy I/O and optimized data paths
- **15 GB/s throughput per node**: Parallel I/O for dataset loading
- **50-99% space efficiency**: Compression and deduplication of redundant data
- **11-nines durability**: Erasure coding with configurable K+M schemes

Architecture Highlights:

- **NVMe-Native Design**: Log-structured storage optimized for flash, minimizing write amplification and integrating wear leveling
- **Zero-Copy I/O**: Splice/sendfile system calls and kernel-bypass techniques reduce CPU consumption by 60-70% while increasing throughput 40-50%
- **Lockless Data Structures:** Fine-grained lock-free implementation using atomic operations enables linear scalability to 64+ cores
- **Erasure Coding**: 10+4 scheme provides durability equivalent to 4-way replication while consuming only 1.4x space
- **Policy-Driven Tiering**: Automatic data movement across NVMe/SSD/HDD/Archive based on access patterns
- **Space-Efficient Snapshots**: Copy-on-write enables dozens of dataset versions with minimal overhead

Integration:

- NQRust-HV: VM disk images with high-performance snapshots for rapid cloning
- NQRust-Lake: Direct access to datasets without POSIX overhead
- NQRust-LLMOps: Model checkpoints with automatic versioning and metadata
- NQRust-Guard: Immutable snapshots for backup targets

Deployment Results:

- 89% storage cost reduction versus traditional SAN/NAS
- 12x faster I/O performance through NVMe and parallel architecture
- 650% ROI within 18 months
- 30-day deployment from assessment through production migration

3.2 Data Platform for Analytics

Beyond high-performance storage, AI workloads require sophisticated data management for training datasets, model artifacts, and analytical queries. This demands unified systems supporting batch, streaming, and analytical access with ACID transactions, time travel, governance, and sub-second query performance.

NexusRust Solution

NexusRust-Lake: Data Lakehouse Platform

NQRust-Lake implements a Rust-native lakehouse combining storage and analytical capabilities optimized for AI workflows.

Key Capabilities:

- **Unified Format**: Single system for batch, streaming, and analytical data with ACID transactions
- **Open Formats**: Parquet, Iceberg, Delta Lake compatibility preventing vendor lock-in

NexusRust Solution

- **Columnar Acceleration**: Rust-native query engine with vectorized execution
- **Intelligent Caching**: Automatic data caching based on access patterns
- **Time Travel**: Dataset versioning for reproducibility and rollback
- **Governance**: Access control, lineage tracking, and compliance reporting

Results:

- 68% TCO reduction compared to traditional data warehouses
- 600%+ ROI in validated 5-year TEI model at 500TB scale
- Built on NQRust-Storage for high-performance analytical workloads

4. Network Architecture

GPU-to-GPU communication during distributed training requires specialized network fabric with characteristics fundamentally different from traditional data center networks.

4.1 High-Speed Interconnect Requirements

Distributed training using data parallelism or model parallelism requires all-to-all communication between GPUs during gradient synchronization. A single training step may involve exchanging gigabytes of data across hundreds of GPUs with timing requirements measured in microseconds.

Architecture Insight

AI Network Requirements:

- **Ultra-Low Latency**: <2µs node-to-node for gradient synchronization without training slowdown
- **High Bandwidth**: 400Gbps-3.2Tbps per link for model parallelism across large models
- **RDMA Support**: Zero-copy network transfers eliminating kernel overhead for memory-to-memory communication
- **Lossless Operation**: Zero packet loss during high-throughput transfers requiring PFC (Priority Flow Control)
- **Optimal Topology**: Fat-tree or similar design minimizing hop count between GPU nodes
- **Network Isolation**: Per-tenant VLANs with microsegmentation for security

Implementation Approach:

InfiniBand/RoCE Fabric: RDMA-capable network for GPU-to-GPU communication with sub-10µs latencies

NVMe-over-Fabrics: Direct storage access from compute nodes using RDMA protocols

Fat-Tree Topology: Non-blocking bandwidth between any node pair enabling efficient all-to-all communication

Software-Defined Networking: Dynamic provisioning and QoS policies managed through NQRust-FleetMgr

Integration: NQRust-Storage optimized for RDMA protocols, NQRust-SecureGPU coordinates GPU and network allocation

5. Security Architecture

AI infrastructure requires comprehensive security across all layers, from hardware through applications. The defense-in-depth model implements multiple security boundaries, ensuring that compromise of one layer does not expose other layers.

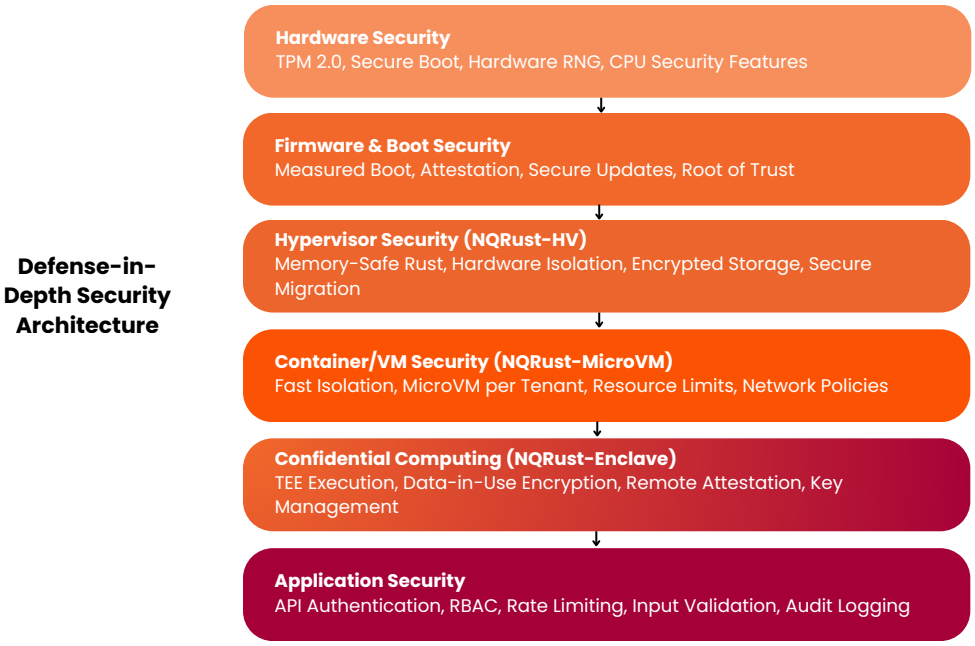

Figure 4: Six-Layer Security Model

5.1 Multi-Layer Security Model

Security Requirement

Comprehensive Security Requirements:

- **Memory Safety**: Rust implementation eliminates buffer overflows, use-after-free, data races at compile time
- **Defense-in-Depth**: Multiple security boundaries from hardware through application
- **Encryption Everywhere**: Data at rest (storage), in transit (network), and in use (TEE)
- **Hardware Root of Trust**: TPM 2.0 attestation from boot through application execution
- **Zero Trust Architecture**: Continuous verification at every layer, no implicit trust
- **Compliance Automation**: Built-in GDPR, PDPA, OJK, ISO27001 controls

6. Orchestration and Management

Managing heterogeneous AI infrastructure requires unified orchestration abstracting complexity while providing fine-grained control.

6.1 Unified Control Plane

NQRust-FleetMgr provides single API for managing VMs, containers, GPUs, storage, and network with policy-driven automation.

Key Capabilities:

- **Unified Workload API:** Single specification for containers, MicroVMs, GPUs, and persistent storage
- **AI-Driven Scheduler:** ML-based placement for optimal utilization: 85% CPU, 90% GPU, 92% storage
- **Indonesian Compliance:** Built-in UU PDP, OJK, BI automation for data residency
- **GitOps Integration:** Infrastructure-as-code with version control and audit trails
- **Multi-Region Federation:** Unified management across data centers and edge sites

Results: 70-80% operational expenditure savings, 95% faster deployment (weeks to minutes)

6.2 Observability Platform

NQRust-Insight delivers AI-native monitoring with anomaly detection, root cause analysis, and predictive capacity planning.

Operational Impact:

- MTTD reduction: 4.2 hours → 3 minutes (84x improvement)
- Alert reduction: 2,400/day → 12/day (99.5% reduction)
- Utilization increase: 42% → 89% average
- Monitoring cost reduction: 65%
- ROI: 1,067% validated model

7. Implementation Roadmap

Phase	Duration	Activities	Investment
Foundation	4-6 weeks	Assessment, hardware, selection, network design, security baseline	$100K
Deployment	6-8 weeks	Install stack, configure networking, provision storage, testing	$300K
Migration	4-6 weeks	Workload, migration, performance tuning, integration, training	$200K
Optimization	On going	Performance tuning, cost optimization, scaling, advanced features	$100K/yr

Table 1: Phased Implementation Approach

8. Compliance and Governance

Regulation	Region	Key Requirements	Status
UU PDP	Indonesia	Data residency, consent management, breach notification	Compliant
OJK	Indonesia	Financial data protection, audit trails, incident response	Comliant
GDPR	EU	Privacy by design, right to erasure, data portability	Certified
PDPA	Singapore	Data protection, consent, accountability	Compliant
PP71	Indonesia	Government data security, classification, controls	Compliant
ISO 27001	Global	Global & Information security management system	Certified
SOC 2 Type II	Global	Global & Security, availability, confidentiality controls	Certified

Table 2: Regulatory Compliance Coverage

9. Conclusion

Modern AI workloads require infrastructure fundamentally different from traditional data centers. The NexusRust platform delivers a complete, integrated solution purpose-built for secure, high-performance AI operations at enterprise scale.

Platform Differentiator:

- **Memory-Safe Foundation:** Rust implementation eliminates 70% of critical vulnerabilities
- **Vertically Integrated:** Complete stack from hypervisor through AI platform
- **Proven Performance:** 2.5M IOPS, <100µs latency, 78% GPU utilization
- **Measurable ROI:** 66% TCO reduction, 650-1,067% ROI across components
- **Rapid Deployment:** 30-90 day programs from assessment to production

Organizations evaluating AI infrastructure face a strategic choice: adapt legacy systems designed for conventional workloads, accept cloud limitations and costs, or deploy purpose-built infrastructure delivering measurable advantages. NexusRust represents the third path—enabling enterprises to build secure, high-performance AI infrastructure under their direct control.

Build Your Secure AI Infrastructure

Enterprise-grade platform with validated ROI

Nexus Quantum Technology
contact@nexusquantum.id
Web: https://nexusquantum.id

Your data deserves better than cloud lock-in. Set it free.