COREDGE

Dflare AI

Enterprise GPU Infrastructure Platform

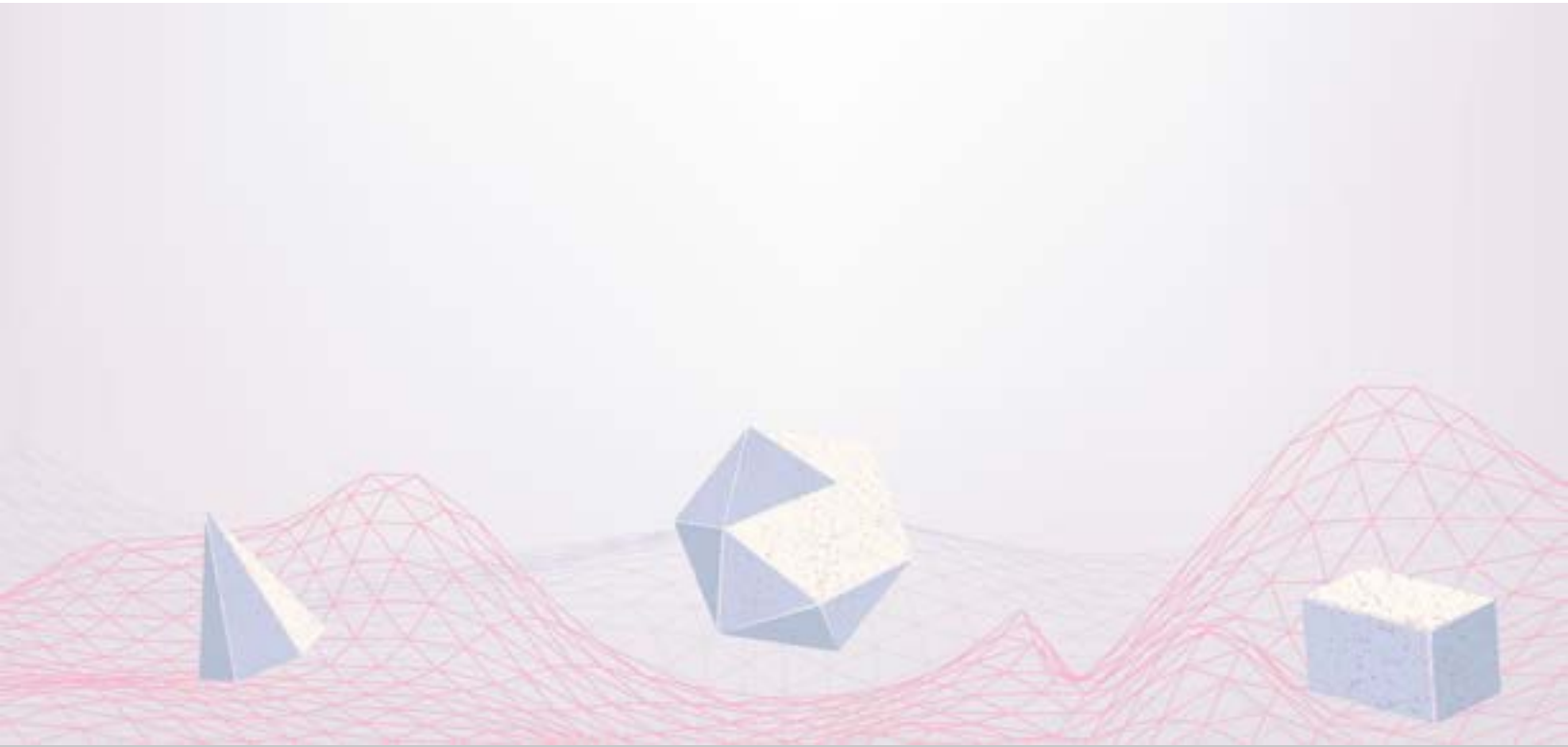
Whitepaper

Transforming bare metal GPU servers into production-ready, multi-tenant compute environments

Delivering the raw power of dedicated hardware with the operational simplicity of a managed cloud service.

Version 1.0

Contents



Executive Summary

The rapid advancement of artificial intelligence has fundamentally changed the requirements for compute infrastructure. Modern workloads—particularly large-scale model training, fine-tuning, and real-time inference—demand high-density GPU clusters, ultra-low-latency interconnects, and sustained high-throughput data access.

Traditional cloud architectures, designed for general-purpose compute, are increasingly misaligned with these requirements. At the same time, building and operating dedicated GPU infrastructure introduces significant complexity across hardware provisioning, cluster orchestration, networking, storage, and security.

The platform provides unified support for Kubernetes and Slurm workloads, high-performance storage integrated over InfiniBand fabric, hardware-enforced tenant isolation, and comprehensive lifecycle automation—all accessible through a single control plane.

KEY INSIGHT

Dflare AI addresses this gap by delivering a fully managed, enterprise-grade GPU infrastructure platform that transforms bare metal GPU servers into production-ready, multi-tenant compute environments.

Introduction

Artificial intelligence is transitioning from experimental workloads to mission-critical infrastructure. Organizations across industries are deploying AI systems that require:

Large-scale distributed training environments

Reliable and low-latency inference platforms

Secure, multi-tenant resource sharing

Transparent usage tracking and cost control

These requirements place significant pressure on infrastructure systems, exposing limitations in both public cloud offerings and traditional on-premises deployments. This paper outlines the challenges of modern AI infrastructure and presents Dflare AI as a purpose-built solution.

The AI Infrastructure Challenge

Hardware Complexity

Modern GPU servers are highly specialized systems. Each node may include multiple high-performance GPUs, high-bandwidth interconnects, and complex NUMA configurations. Achieving optimal performance requires precise tuning at both the hardware and operating system levels.

Fragmented Compute Paradigms

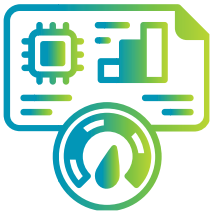
AI workloads typically span two operational models: Kubernetes for containerized pipelines, inference, and MLOps workflows, and Slurm (HPC) for batch scheduling and large-scale training jobs. Most platforms support one paradigm effectively, but not both in a unified environment.

Multi-Tenancy and Isolation

Sharing GPU infrastructure across teams or customers introduces challenges in ensuring strict isolation across network communication, storage access, compute resources, and identity management. Software-only isolation is insufficient in high-performance environments.

Data Throughput Requirements

AI workloads are highly data-intensive. Training processes require sustained access to large datasets, often at throughput levels that exceed traditional storage architectures.



Operational Overhead

Managing GPU infrastructure involves provisioning hardware, maintaining clusters, monitoring performance, enforcing compliance, and tracking usage for billing. These requirements create a significant operational burden for most organizations.

Dflare AI Platform Overview

Dflare AI is a GPU-native infrastructure platform designed to address the full lifecycle of AI compute operations. It enables organizations to:

- Provision bare metal GPU resources on demand
- Deploy Kubernetes clusters powered by CKP (Coredge Kubernetes Platform) and Slurm clusters through a unified interface
- Access high-performance storage integrated with compute fabric
- Enforce multi-tenant isolation across all infrastructure layers
- Monitor system performance and resource utilization
- Track and report usage with granular billing metrics.
- Run ML workloads through an integrated ML Platform with GPU notebooks, distributed training, LLM inference, fine-tuning, and experiment tracking

The platform abstracts infrastructure complexity while preserving direct access to hardware performance.

EIGHT OPERATIONAL PILLARS

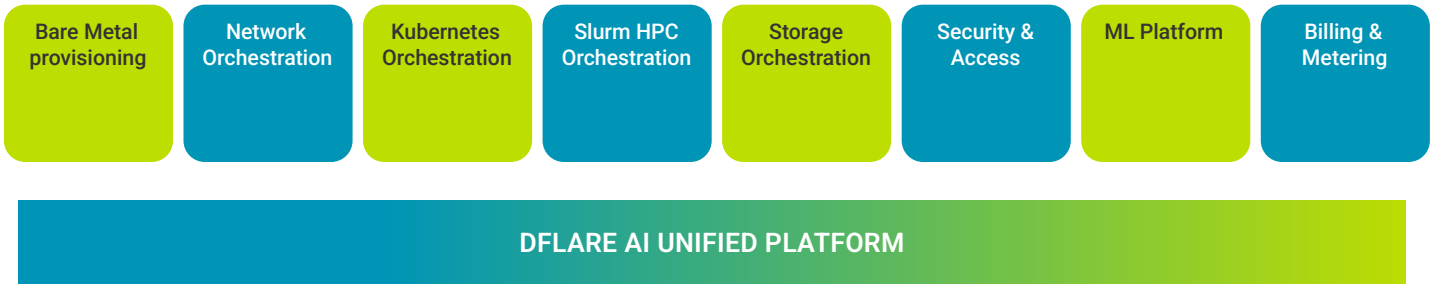


Exhibit 1: Dflare AI automates eight critical operational pillars.

Architecture Overview

Dflare AI is built on a modular, microservices-based architecture with clearly defined layers spanning control plane, compute, networking, storage, identity, and observability.

PLATFORM ARCHITECTURE

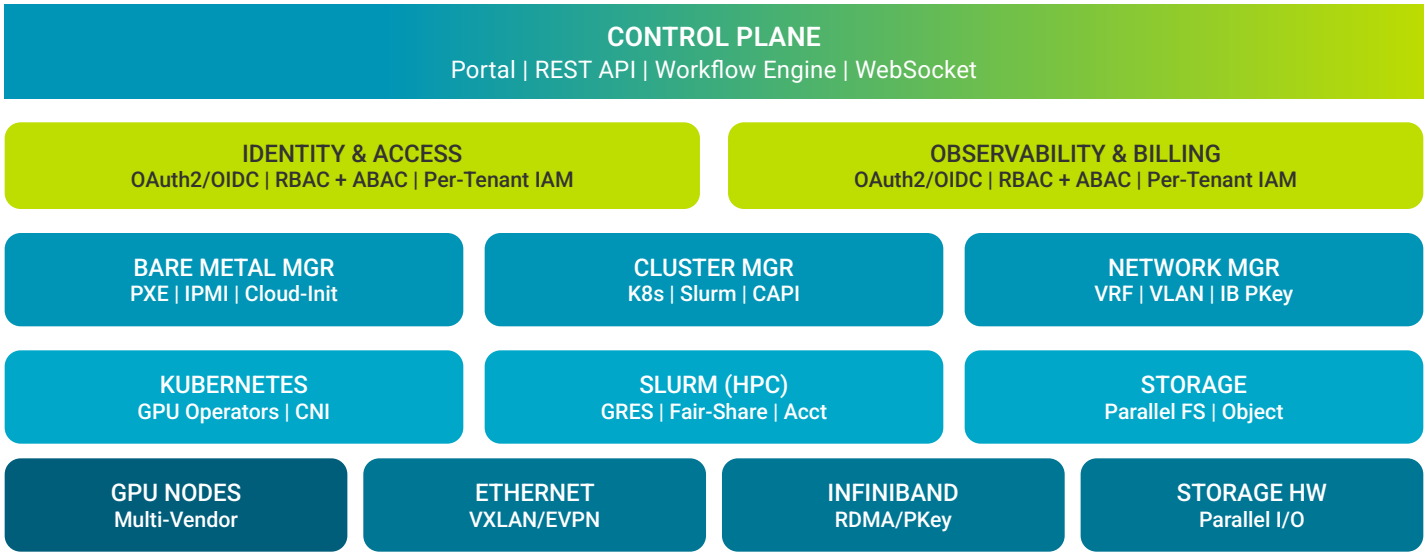


Exhibit 2: Platform architecture — from control plane to infrastructure.

Control Plane

Provides orchestration capabilities through a web-based portal, REST APIs, and workflow automation services.

Compute Layer

Manages bare metal GPU nodes, Kubernetes cluster orchestration, and Slurm-based HPC scheduling on shared physical infrastructure.

Network Fabric

Dual-fabric architecture: Ethernet frontend for control/management, InfiniBand backend for GPU-to-GPU and GPU-to-storage communication. Isolation via VRF, VLAN, and IB partition keys.

Storage Layer

High-performance parallel filesystem over InfiniBand for training workloads, and object/file storage for platform services. Dual-layer isolation.

Identity and Access

Enterprise IAM integration with OAuth2/OIDC, RBAC + ABAC, and per-tenant isolation of identity domains.

Observability and Metering

Real-time monitoring with GPU telemetry, node metrics, and cluster state. Granular usage aggregation for billing per tenant, project, and user.

Network Architecture

Dflare AI is built on a modular, microservices-based architecture with clearly defined layers spanning control plane, compute, networking, storage, identity, and observability.

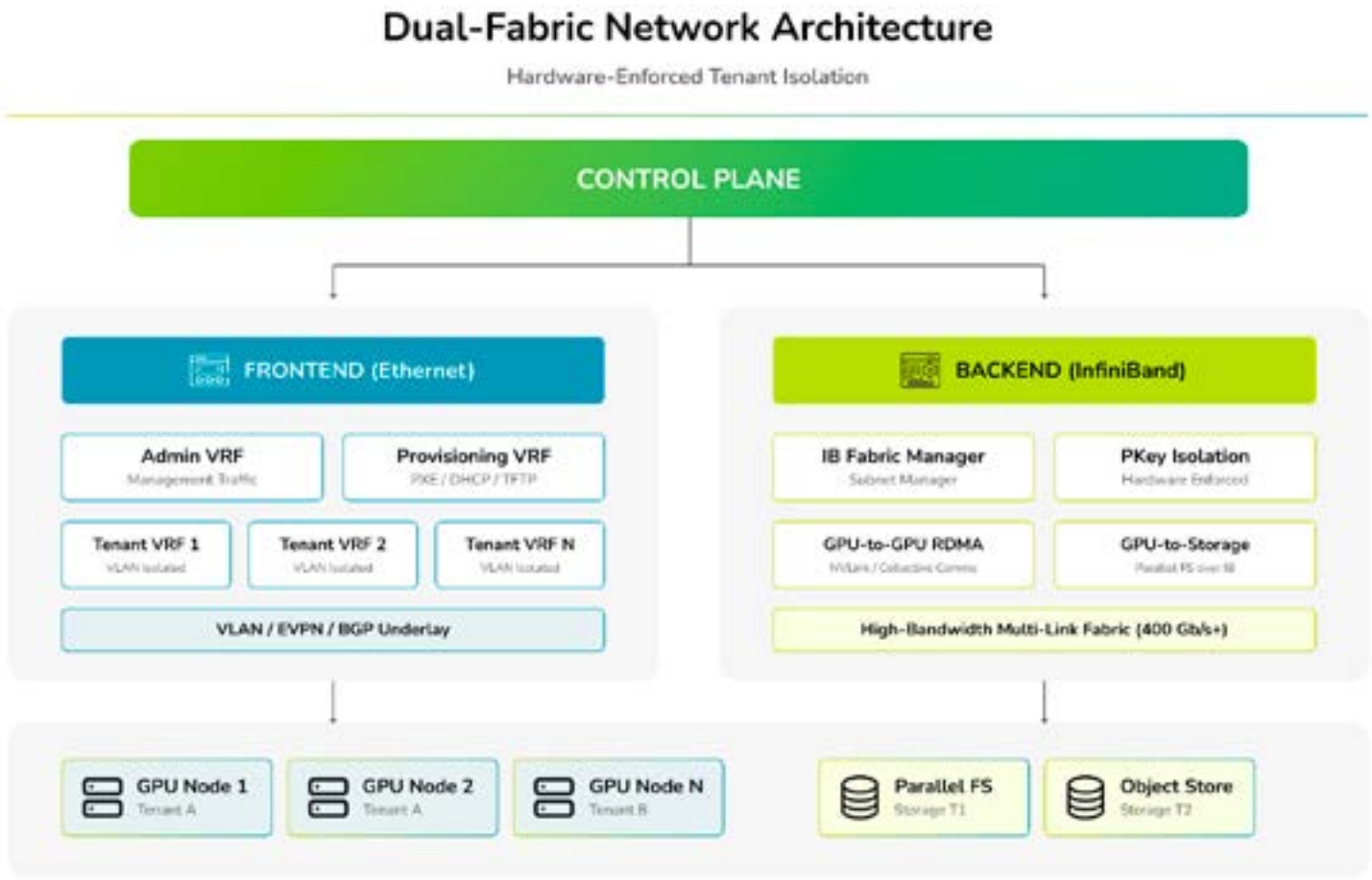


Exhibit 3: Dual-fabric network architecture with hardware-enforced tenant isolation.

Core Capabilities

Unified Kubernetes and Slurm Orchestration

Dflare AI supports both containerized and HPC workloads on the same infrastructure, enabling organizations to run diverse AI workloads without duplicating environments, standardize operations across teams, and optimize resource utilization.

Bare Metal Performance

The platform provides direct access to GPU hardware without virtualization overhead. System-level optimizations are applied automatically, ensuring consistent and predictable performance. Standard GPU slicing via NVIDIA MIG (Multi-Instance GPU) profiles enables partitioning supported GPUs into isolated instances for efficient resource utilization.

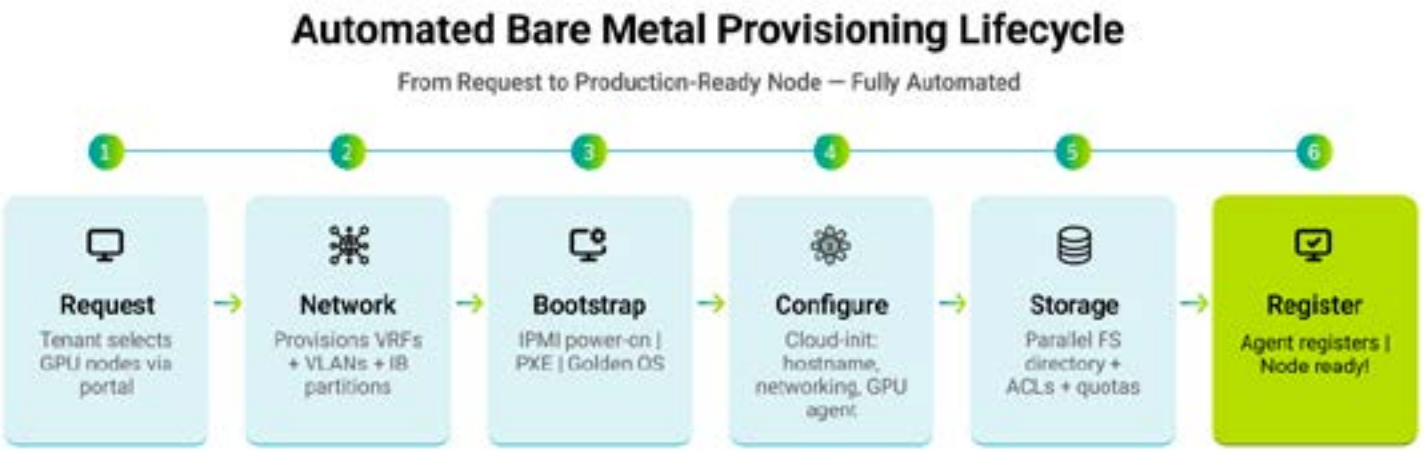


Exhibit 4: Automated bare metal provisioning lifecycle.

Multi-Tenant Isolation

Isolation is enforced across multiple layers: network segmentation using VRF and VLAN, InfiniBand partitioning for GPU communication, storage-level access controls, and compute-level resource isolation. This multi-layer approach ensures strong separation between tenants.

High-Performance Storage Integration

Storage is tightly integrated with the compute fabric, enabling high-throughput data access, parallel I/O operations, and efficient handling of large datasets.

Automated Lifecycle Management

The platform automates infrastructure provisioning, cluster deployment, scaling and upgrades, and resource deallocation — reducing operational overhead and accelerating time to production.

Monitoring and Billing

Comprehensive visibility into system performance, real-time resource usage metrics, and detailed billing and reporting capabilities. Every GPU-hour, CPU-hour, and storage byte is tracked.

ML Platform

Integrated machine learning environment providing GPU notebooks, distributed training, LLM inference with OpenAI-compatible APIs, model fine-tuning, experiment tracking with MLflow, and dataset management — enabling the complete ML lifecycle within workspace isolation.

Tenant Isolation Model

Dflare AI is built on a modular, microservices-based architecture with clearly defined layers spanning control plane, compute, networking, storage, identity, and observability.

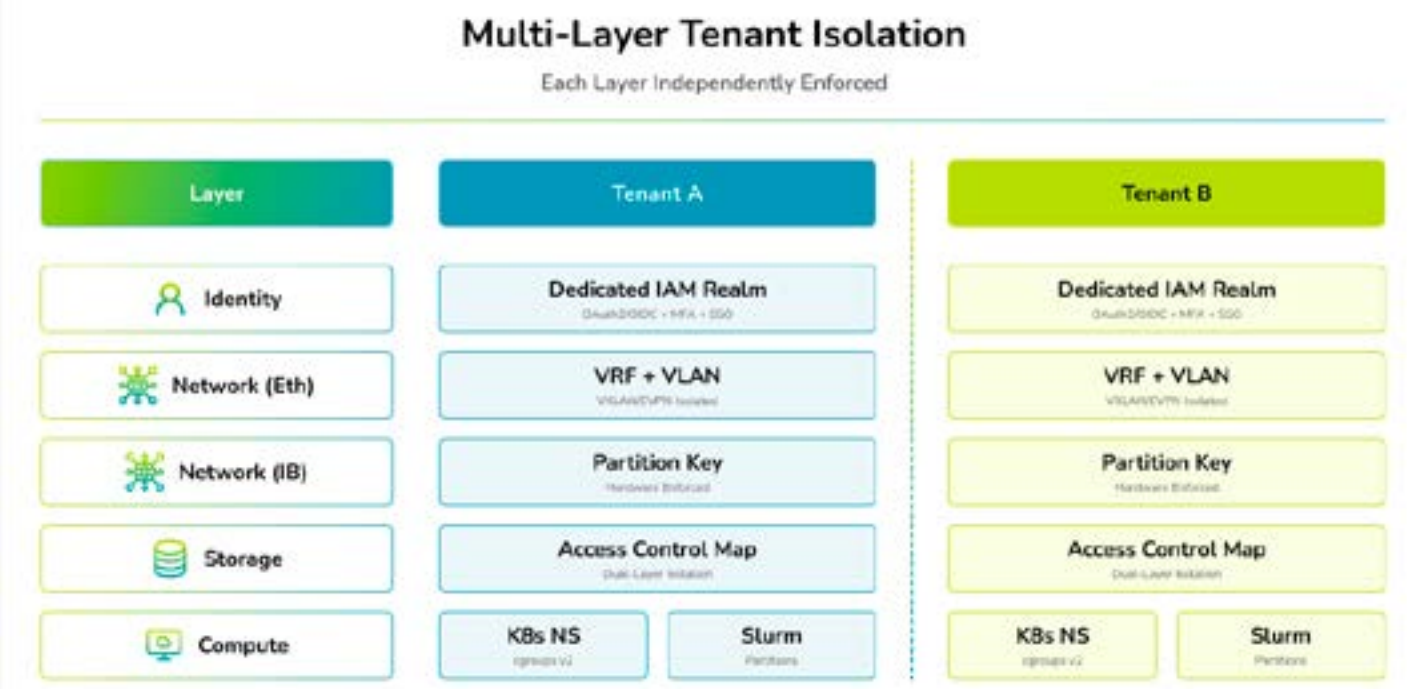


Exhibit 5: Multi-layer tenant isolation — each layer independently enforced.

Security and Compliance

Dflare AI implements a defense-in-depth security model based on zero-trust principles. No action is trusted by default; every request is authenticated and authorized at multiple levels.

- Authentication via OAuth2/OpenID Connect with short-lived JWT tokens and MFA support
- Authorization using RBAC and ABAC with per-tenant IAM realms
- Authorization using RBAC and ABAC with per-tenant IAM realms
- Encrypted communication using TLS 1.2+ and mTLS between internal services
- Immutable audit logging with correlation IDs for full traceability
- Hardware-enforced network isolation via VRF, VLAN, and InfiniBand partition keys

COMPLIANCE ALIGNMENT

Standard	Control Area	Implementation
NIST 800-53 Rev 5	Access Control	RBAC via IAM, SSO, scoped tokens, tenant realm isolation
NIST 800-53 Rev 5	Audit	Immutable logs, session tracking, telemetry pipeline
ISO/IEC 27001	Access / Crypto	IAM with RBAC, TLS/mTLS, PKI, certificate lifecycle
HIPAA	Access / Audit	IAM, ACLs, immutable logs, session playback

DEFENSE-IN-DEPTH SECURITY MODEL

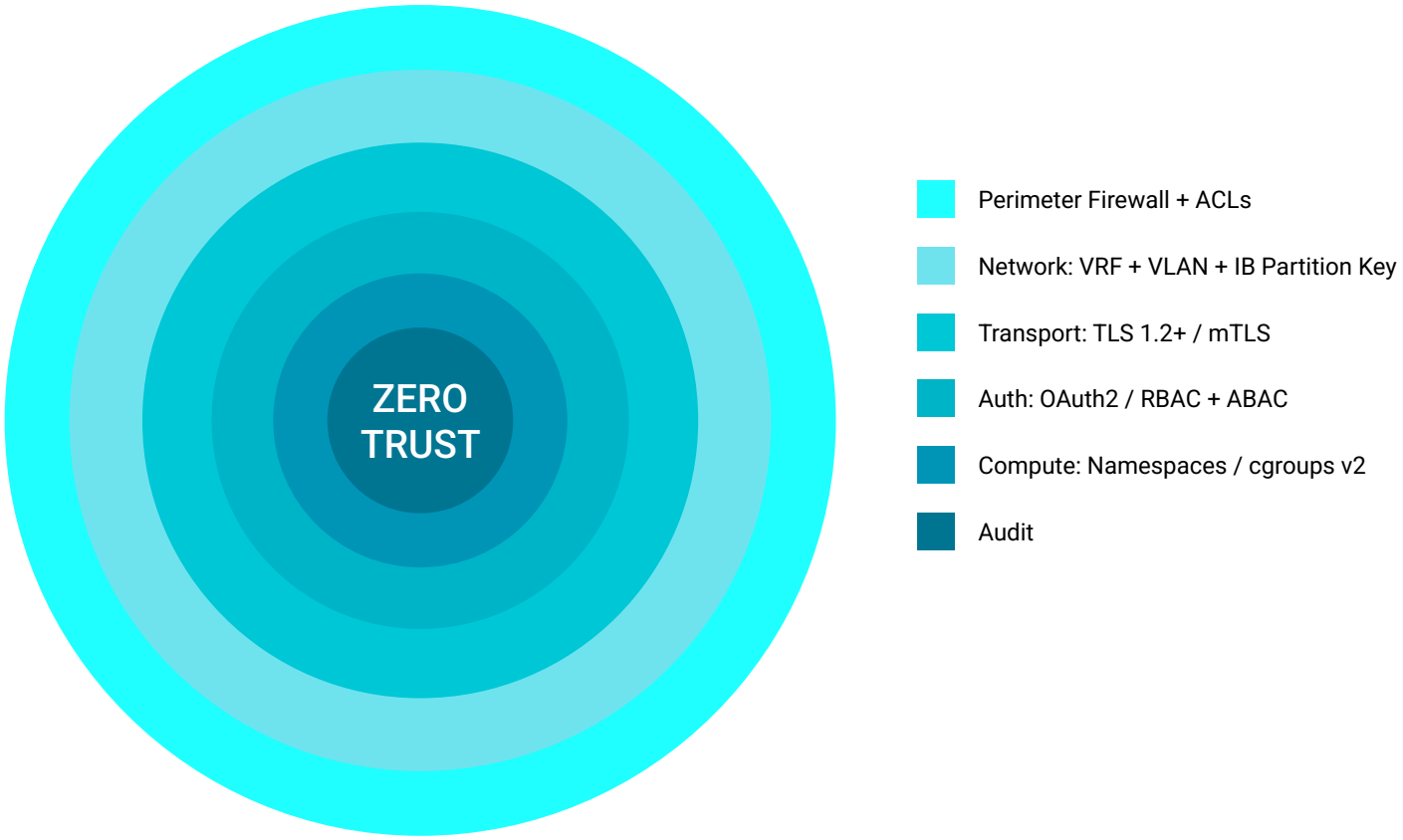


Exhibit 6: Concentric defense-in-depth security model.

Use Cases

1

Large-Scale Model Training

Distributed training across multi-node GPU clusters with high-speed interconnects and RDMA collective operations. Parallel filesystem delivers the storage throughput to keep GPUs fed.

2

Model Fine-Tuning and RLHF

Secure, isolated environments for adapting models to proprietary datasets with dedicated GPU allocation, isolated storage, and job-level accounting.

3

AI Inference at Scale

Scalable deployment of inference workloads with GPU resource requests, auto-scaling, load balancing, and performance monitoring through integrated dashboards.

4

High-Performance Computing

Scientific simulation, computational fluid dynamics, molecular dynamics, and climate modeling on Slurm with GPU-aware batch scheduling.

5

Enterprise MLOps

End-to-end pipelines for model development, deployment, and monitoring with Kubernetes-native workflows and persistent storage for datasets and model artifacts.

Target Customers

Dflare AI is designed for organizations requiring high-performance, secure, and scalable GPU infrastructure:

Segment	Use Case	Key Platform Benefit
Cloud Service Providers	GPU compute for end customers	Multi-tenant isolation, billing, scale
AI Research Institutions	Model training and experimentation	Slurm + K8s, fair-share scheduling
Enterprise AI Teams	Production ML and inference	K8s-native, monitoring, RBAC
Sovereign AI Programs	National AI infrastructure	Air-gapped, compliance, data sovereignty
Regulated Industries	Healthcare, finance	HIPAA/NIST alignment, audit trails

Key Differentiators

Dflare AI is designed for organizations requiring high-performance, secure, and scalable GPU infrastructure:

Unified Kubernetes + HPC.

Run both containerized and Slurm workloads on the same bare metal. infrastructure with unified networking, storage, security, and billing.

True Bare Metal Performance.

Direct GPU access without virtualization overhead. Hardware-level BIOS and OS tuning pre-applied via golden images for consistent, predictable performance.

Hardware-Enforced Multi-Tenant Isolation

Isolation at InfiniBand switch hardware (partition key), filesystem (access control map), and network fabric (VRF/VXLAN). Double isolation on storage ensures defense-in-depth.

Compute and Storage Over High-Speed Fabric.

Parallel filesystem over InfiniBand delivers storage throughput that matches GPU compute appetite.

Full Lifecycle Automation.

From bare metal power-on to production cluster—fully automated. No SSH, no manual configuration, no ticket-based provisioning.

Vendor-Agnostic GPU Support.

Supports NVIDIA, AMD, and Intel accelerators—no vendor lock-in.

KEY INSIGHT

Dflare AI uniquely combines unified K8s and HPC orchestration, bare metal performance, hardware-enforced isolation, and full lifecycle automation in a single platform—a combination not available from any single public cloud provider.

Conclusion

The increasing scale and complexity of AI workloads require a new approach to infrastructure—one that combines performance, flexibility, and operational efficiency.

Dflare AI provides a unified platform that simplifies the deployment and management of GPU infrastructure while maintaining the performance characteristics required for advanced AI workloads.

By integrating compute, storage, networking, and orchestration into a single system, Dflare AI enables organizations to focus on building and deploying AI solutions rather than managing infrastructure.

Get in touch with us

https://coredge.io info@coredge.io