Operationalizing AI Fabrics with Aviz ONES, NVIDIA Spectrum-X Ethernet, and Rafay

Executive Summary

As AI becomes central to digital transformation, enterprises are investing heavily in GPU infrastructure. The opportunity is to operationalize them as a self-service, multi-tenant, production-ready platform.

Together, **Aviz ONES, NVIDIA Spectrum-X Ethernet, and the Rafay Platform** deliver that capability. This unified solution turns GPU fabrics into governed, cloud-native environments that can be consumed instantly by developers and data scientists, bridging the gap between high-performance hardware and enterprise-grade service delivery.

From fabric intelligence to cloud-native orchestration and self-service consumption, this collaboration empowers organizations to transform static infrastructure into a launchpad for AI innovation, achieving higher utilization, faster time-to-value, and greater agility.

The New AI Operations Model

Modern AI workloads require low-latency East-West (E-W) and North-South (N-S) networks, deterministic performance, and seamless visibility across every layer from network fabrics to Kubernetes clusters.

While **NVIDIA Spectrum-X Ethernet** provides the high-performance, lossless foundation, **Aviz ONES and Rafay** bring the operational intelligence that makes these environments run as a unified AI fabric:

NVIDIA Spectrum-X Ethernet Fabrics:
Lossless Ethernet architecture optimized for AI workloads, delivering congestion-free throughput and low latency with NVIDIA SuperNICs.

Aviz ONES:
Provides GPU-aware fabric orchestration, tenant segmentation, and lifecycle automation across Spectrum-X fabrics.

Rafay Platform:
Automates inventory allocation and lifecycle operations for compute and AI applications, and multi-tenant policy enforcement for end-users.

Together, these layers make GPU infrastructure instantly consumable, combining predictable performance with enterprise-grade governance and a self-service developer experience.

Integrated Operations: From Disparate Components to a Service Fabric

Enterprises no longer need to choose between control and speed. By combining Aviz, NVIDIA, and **Rafay** capabilities, platform teams can deliver AI services with both.

Unified Control:

Aviz ONES designs, deploys, and monitors GPU-aware fabrics with intent-driven automation.

Orchestrated Compute:

Rafay turns those networked GPU resources into secure, governed compute (e.g. Bare Metal as a Service, Kubernetes as a Service, VMs as a Service).

Cloud-Native Consumption:

Developers and data scientists gain immediate access to GPU compute, notebooks, and inference APIs through intuitive self-service catalogs.

The outcome: an operational model where fabric, orchestration, and consumption work in concert, enabling enterprises to **scale AI confidently and profitably.**

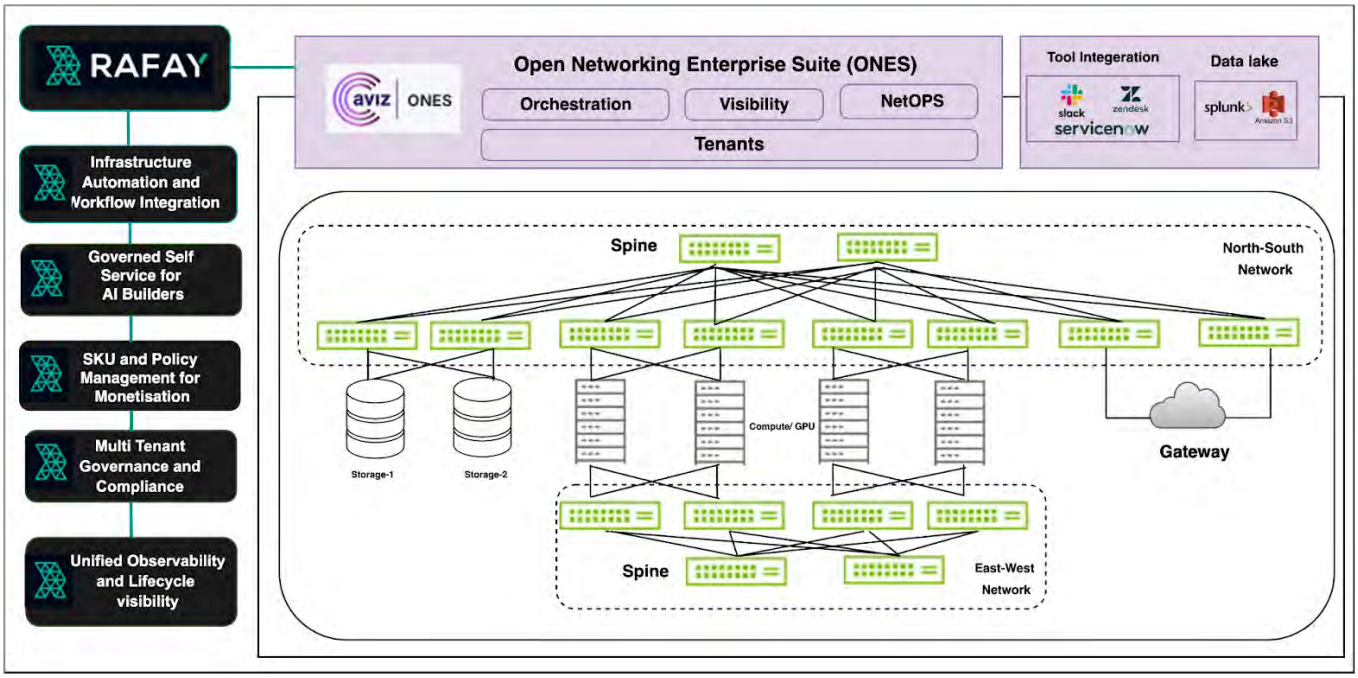

Figure 1: Unified Fabric-to-Cloud Architecture: Integrating Aviz ONES, NVIDIA Spectrum-X, and Rafay for Operational AI at Scale

Aviz ONES: GPU-Aware Fabric Orchestration

Aviz ONES delivers end-to-end automation for **NVIDIA Spectrum-X Ethernet** fabrics, ensuring predictable, lossless performance and simplified operations.

Key Capabilities

- **Intent-Driven Design:** Automates design and validation using NVIDIA AIR and ONES templates.
- **GPU-Aware Networking:** Allocates GPUs to tenants and maps the dedicated NVIDIA SuperNICs for deterministic communication.
- **Tenant Segmentation:** Logical and physical isolation through VXLAN/EVPN overlays per tenant or workload.
- **Lifecycle Automation:** Day-0 to Day-2 fabric setup, scaling, and upgrades with integrated telemetry.
- **Ecosystem Integration:** Native support for Splunk, ELK, ServiceNow, and Slack for visibility and incident response.

By automating the fabric layer, **Aviz ONES** establishes the performance and reliability baseline required for enterprise-grade AI operations

Rafay Platform: Infrastructure Orchestration and Self-Service for AI

Once the network fabric is optimized, the next step is enabling **self-service access** to that power.
The **Rafay Platform** provides the **infrastructure orchestration and workflow automation layer** that transforms GPU and CPU resources into a secure, multi-tenant, self-service Platform-as-a-Service (PaaS).

Core Capabilities

Lifecycle Orchestration:

Simplifies deployment, scaling, and upgrades of GPU attached compute and API applications across private, hybrid, and sovereign environments.

GPU-Aware Allocation:

Dynamically binds GPUs ensuring deterministic performance and maximum utilization.

Secure Multi-Tenancy:

Implements enterprise-grade RBAC, policy controls, and audit trails for tenant isolation.

Self-Service Workflows:

Gives developers and data scientists instant access to GPU attached compute, AI workbenches, and inference endpoints through a governed catalog.

Policy & Cost Governance:

Provides visibility into usage and exposes metrics for billing/chargeback

Ecosystem Integration:

Works natively with NVIDIA NIM, Run: AI, Kubeflow, Ray, and Jupyter for end-to-end AI workflows.

With **Rafay**, organizations elevate infrastructure into a launchpad for innovation, turning GPUs from a fixed cost into a shared, monetizable service that accelerates every stage of AI delivery.

NVIDIA Spectrum-X Ethernet and SuperNICs: Performance by Design

NVIDIA Spectrum-X Ethernet provides the deterministic, lossless foundation for large-scale AI training and inference.

Optimized for AI:

RoCEv2 transport ensures congestion-free throughput across thousands of GPUs.

Scalable Architecture:

Linear expansion from rack-level to AI factory scale.

SuperNIC Acceleration:

Delivers adaptive routing and in-band telemetry for consistent performance.

When paired with **Aviz ONES and Rafay, Spectrum-X** forms the physical and logical spine of an AI fabric that is both predictable and programmable.

End-to-End Workflow

- **Tenant Onboarding: Rafay** provisions new tenants and requests GPU and network resources.
- **GPU Allocation: Aviz ONES** maps GPUs to SuperNICs for low-latency connectivity.
- **Network Segmentation: ONES** establishes VXLAN/EVPN overlays per tenant.
- **Cluster Deployment: Rafay** deploys Kubernetes clusters and binds GPU resources automatically.
- **Operational Visibility:** Telemetry flows from **ONES** to enterprise monitoring systems while **Rafay** governs workload health, scaling, and chargeback.

The result is a fully automated Day-0 to Day-2 workflow that eliminates friction and accelerates AI service delivery.

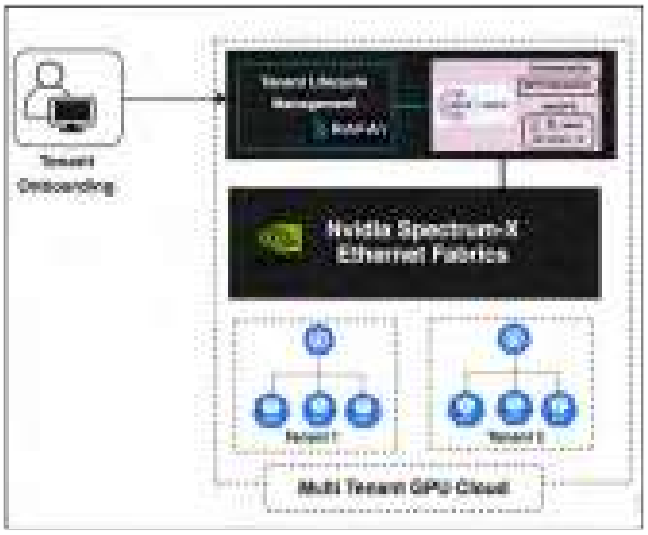

Figure 2: Rafay manages tenant orchestration and governance, Aviz ONES allocates GPU and network resources across NVIDIA Spectrum-X fabrics, enabling a secure, multi-tenant GPU cloud.

Joint Value and Key Benefits

The combined **Aviz ONES, NVIDIA Spectrum-X Ethernet, and Rafay** Platform stack delivers a complete spectrum of operational capabilities from fabric intelligence and lifecycle automation to self-service consumption and governance at scale.

Together, they transform static GPU infrastructure into a governed, cloud-native platform that is predictable, monetizable, and ready for production from day one.

Capability	Delivered By	Value to Customers
GPU-Aware Orchestration	**Aviz ONES + Rafay Platform**	Predictable GPU performance from fabric to pod with automated binding and lifecycle control.
Secure Multi-Tenancy	**Aviz ONES + Rafay Platform**	Isolation and governance across network and Kubernetes layers for sovereign and regulated deployments.
Inventory Management	**Rafay Platform**	Maintain real-time visibility into available compute, GPU pools, and service capacity for efficient resource allocation and chargeback.
Policy & Governance Framework	**Rafay Platform**	Enforce enterprise-grade RBAC, quota, and compliance policies across tenants and environments with full auditability
SKU Management	**Rafay Platform**	Define, package, and present GPU/CPU and AI service offerings as standardized SKUs to simplify consumption and monetization.
Automated Lifecycle	**Aviz ONES + Rafay Platform**	Unified Day-0/1/2 operations and scaling across fabric and workloads.
Self-Service GPU Consumption	**Rafay Platform**	Instant access to GPU clusters and AI workbenches via governed one-click catalogs.
Integrated Observability	**Aviz ONES**	End-to-end telemetry and alerting through Splunk, ELK, and Slack.
Scalable Architecture	**Joint Solution**	Modular scale from rack-level deployments to full AI factories.

Use Cases

Use Case	Business Challenge	Resulting Value to Customers
AI/ML Training & Inference	AI model training and inference pipelines often face unpredictable performance, fragmented GPU access, and slow provisioning cycles.	Deterministic, high-throughput GPU clusters across NVIDIA Spectrum-X Ethernet fabrics, orchestrated by Aviz ONES and Rafay for governed scheduling and lifecycle automation.
Enterprise AI Factories	Scaling multi-tenant GPU environments introduces complexity in isolation, governance, and cost attribution across business units.	Sovereign-ready, multi-tenant AI infrastructure with fabric-level segmentation (Aviz ONES) and governance **(Rafay)**. Enables secure, compliant, and streamlined AI operations.
GPU Platform-as-a-Service (PaaS)	GPU infrastructure often remains underutilized and difficult to monetize without self-service access or usage-based billing.	Governed self-service GPU catalog and billing controls delivered by Rafay atop **Aviz and NVIDIA Spectrum-X.** Converts GPU infrastructure into a scalable, revenue-ready service platform.
Cloud-Native AI Clouds	Hybrid AI workloads across on-prem and cloud environments are complex to orchestrate, monitor, and scale.	Unified, cloud-native control plane with dynamic GPU allocation and policy-driven MLOps automation powered by **Rafay and Aviz**. Delivers cloud-native agility, cost efficiency, and consistent governance across environments.

Conclusion: From Infrastructure to Innovation

The integration of **Aviz ONES, NVIDIA Spectrum-X Ethernet,** and the **Rafay Platform** establishes a clear path to **operational AI excellence.**

- **Aviz ONES** delivers fabric-level automation and tenant segmentation.
- **NVIDIA Spectrum-X Ethernet** guarantees lossless, deterministic network performance.
- **Rafay Platform** transforms those capabilities into a self-service, governed, multi-tenant GPU cloud experience.

Together, they provide a fabric-to-cloud continuum that allows enterprises to deploy, scale, and monetize AI workloads faster, more securely, and with higher utilization.