Architecting the Right Foundation for AI

Simplifying and Optimizing GPU Clouds with WEKA and Rafay. Delivering performance, efficiency, and simplicity in one cohesive platform

Challenges

- Power and space constraints limit density and increase costs.
- Complex multi-tenancy slows deployment and adds to overhead.
- Multi-tenant isolation and governance at exabyte scale

Solution

WEKA and Rafay transform GPU infrastructure into automated, self-service AI clouds with unified orchestration, multi-tenant isolation, and power-optimized efficiency.

Benefits

- Efficiency - power-optimized storage + infrastructure utilization
- Simplicity - automated deployment and Day-2 operations
- Consistency - unified experience across environments

Building GPU accelerated clouds is no longer just about stacking compute. To succeed, providers need to deliver performance, efficiency, and simplicity in one cohesive platform.

AI Clouds, GPU-as-a-Service (GPUaaS) providers, sovereign AI initiatives, and private AI clouds face mounting pressure to maximize infrastructure density while managing complex multi-tenant environments at exabyte scale. Infrastructure teams struggle with power and space constraints that limit density, complex orchestration across bare metal and Kubernetes, and the challenge of delivering secure multi-tenant governance. Traditional approaches force impossible trade-offs between performance, cost, and operational complexity.

This paper shows how NeuralMesh by WEKA and the Rafay Platform together deliver the foundation needed to build efficient, automated AI clouds.

Storage Built for AI

NeuralMesh is designed from the ground up for AI workloads, validated through the NVIDIA Cloud Partner (NCP) Reference Architecture, and trusted by leading AI Clouds and sovereign AI Clouds.

Key Capabilities of NeuralMesh

- **Multi-Tenancy Flexibility** – Offers both logical isolation and full resource isolation using composable clusters, ensuring secure and predictable performance for every tenant.
- **Cloud-Native Ready** – Supports both bare metal and Kubernetes deployments. Use the WEKA Operator for automated WEKA cluster deployment and lifecycle management, or the WEKA CSI plugin for Kubernetes storage integration.
- **Multi-Protocol Access** – Delivers high-speed data via POSIX bare metal / K8s clients (RDMA, RoCEv2, GPUDirect Storage, DPDK) while also supporting S3 and NFS for flexible workload compatibility.
- **Rapid Tenant Onboarding** – With the WEKA Operator, providers can spin up composable WEKA environments in minutes per tenant, accelerating time to value and simplifying operations at scale.

Orchestration at Scale

The Rafay Platform provides the orchestration and governance layer that makes GPUaaS and private AI clouds consumable at scale.

Key Capabilities of the Rafay Platform:

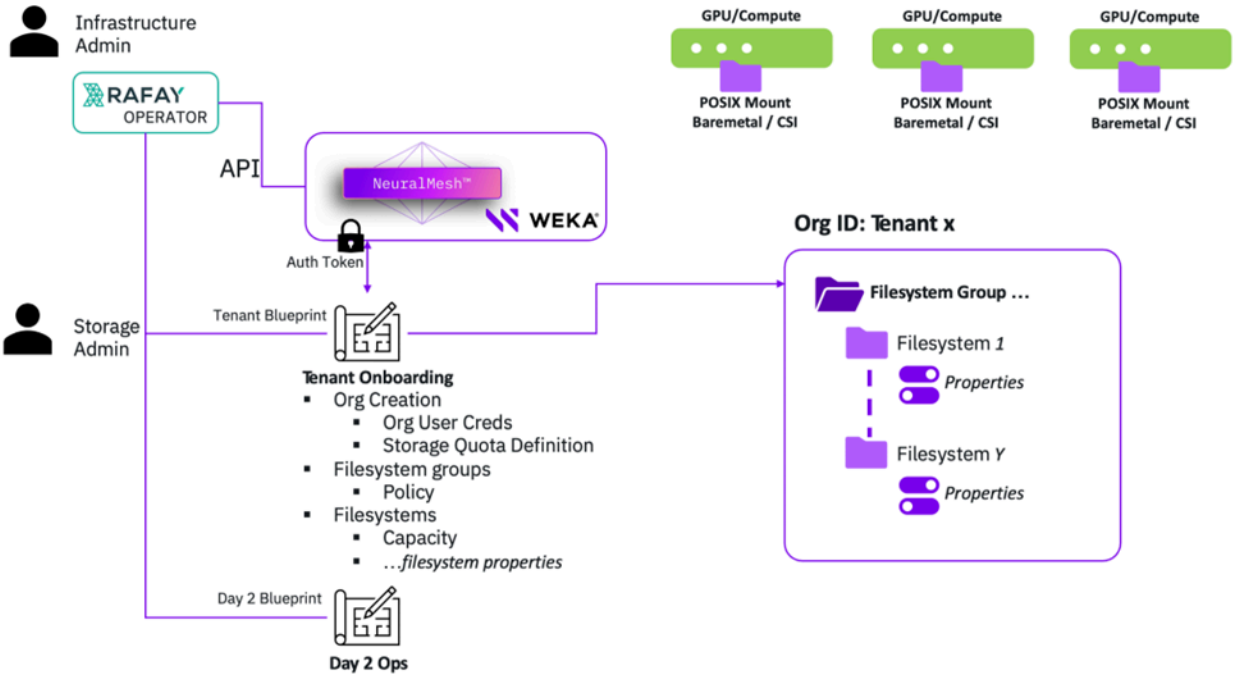
- **Self-Service Consumption** – Rafay overlays AI infrastructure with a self-service portal, API, and CLI, enabling users to provision GPU environments from standardized, pre-approved SKUs without manual tickets or ad-hoc workflows.
- **Declarative SKU-Based Provisioning and Quota Enforcement** – GPU infrastructure and AI environments are productized as version-controlled SKUs that define compute profiles, GPU slicing (including MIG), storage, policies, and lifecycle controls, with real-time quota enforcement.
- **Secure Multi-Tenancy and Governance by Design** – Native multi-tenancy with hierarchical RBAC, namespace isolation, policy enforcement, and centralized audit logging enables secure and compliant sharing of GPU infrastructure across teams and tenants.
- **Turnkey AI Integration** – One-click deployments of NVIDIA NIM, NeMo, Cloud Functions, and NVIDIA Run.ai allow providers to expand services rapidly.

With Rafay, operators gain predictability and consistency across environments, whether public AI Clouds, private AI deployments, or sovereign platforms.

WEKA and Rafay: Purpose-Built for AI Clouds

AI Clouds choose WEKA and Rafay when they need:

- **Efficiency** – Built on WEKA's power-optimized storage with Rafay's automated orchestration for maximum infrastructure utilization.
- **Simplicity** – Rafay abstracts away orchestration complexity, automating deployment and day 2 operations.
- **Consistency** – Tenants consume GPUs, storage, and AI services through a unified, governed experience on bare metal, in containers, or across hybrid environments.



Integration that Streamlines Operations

The Rafay Platform orchestrates NeuralMesh via scoped API tokens. Tenant onboarding creates a dedicated Organization per Tenant, applies quotas/policies, and provisions per-tenant filesystems; apps consume storage via CSI (Kubernetes) or direct POSIX mounts on bare metal. Day-2 changes are executed as Rafay blueprints/workflows.

What's Automated

- **Tenant boundary:** Create Organization per tenant; generate org user creds.
- **Governance:** Apply quotas, security policies, and filesystem groups per tenant.
- **Provisioning:** Create filesystems with requested capacity & properties; publish to clusters.
- **Day-2 ops:** Expand/shrink, snapshot, tiering, and QoS updates; surface health/events.

Typical Onboarding Workflow: When a new ML team onboards, Rafay automatically creates a dedicated Organization with scoped credentials, applies tenant-specific quotas and security policies, provisions filesystems with requested capacity, and publishes storage to the team's cluster—resulting in a ready-to-use environment in minutes.

Reference: WEKA Operator deployments https://docs.weka.io/kubernetes/weka-operator-deployments

Conclusion

By combining NeuralMesh by WEKA with the Rafay Platform, providers can deliver GPU-accelerated clouds that are:

- **Efficient** – optimized power efficiency for AI workloads,
- **Simple** – from deployment to day-2 operations, and
- **Consistent** – across tenants, workloads, and environments.

For AI Clouds, GPUaaS providers, sovereign AI platforms, and enterprises, this means faster time-to-market, lower operational costs, and sustainable differentiation in the rapidly evolving AI cloud economy.

To learn more about Rafay, visit: https://rafay.co/

To learn more about NeuralMesh by WEKA visit: https://www.weka.io/