RAFAY

Rafay Managed Kubernetes as a Service (MKS)

Deliver fully managed, GPU-ready Kubernetes clusters as a governed, self-service offering

Rafay-powered Managed Kubernetes as a Service enables service providers and enterprises to deliver fully managed Kubernetes clusters optimized for modern workloads.

The platform delivers self-service, governed clusters that combine agility with control, built on Rafay MKS, an upstream Kubernetes distribution for bare metal and virtual machines.

Why choose Rafay

Rafay streamlines every stage of Kubernetes operations, removing the burden of manual cluster management while maintaining consistent performance and governance.

- **Operate the full cluster lifecycle without manual effort**. Run multiple Kubernetes versions with automated, non-disruptive upgrades and rollbacks across the fleet from one control plane.
- **Govern access and tenancy by default**. Enterprise-grade multi-tenancy, RBAC, and Zero Trust Kubectl provide secure access without VPNs or bastions, backed by audit logging and identity-based controls.
- **Run AI and ML workloads on a conformant foundation**. Rafay MKS holds CNCF Kubernetes AI Conformance for v1.35, with integrated GPU scheduling for training and inference.

Designed For

Neoclouds: Offer managed Kubernetes as a premium, self-service SKU tenants consume on demand, with fleet-wide standardization and lifecycle automation.

Sovereign AI Clouds: Deliver compliant, governed clusters with Zero Trust access and sovereign deployment for regulated workloads.

Enterprises: Give platform teams a managed Kubernetes experience across on-prem and hybrid infrastructure, with governance built in.

*Aligned with the NVIDIA Cloud Partner (NCP) software reference architecture.

Key Capabilities

Capability	Description
Automated Cluster Lifecycle	Provision clusters on demand and manage upgrades, scaling, and Day-2 operations fleet-wide, with multi-version support across the cluster fleet.
Highly Available Control Plane	Resilient, highly available cluster control plane for production workloads.
Network Automation & CNI	Flexible pod networking with Cilium, Calico, or Flannel, or bring your own CNI, plus native cloud load balancer and ingress integration.
Multi-Tenancy & Access Control	Per-tenant isolation, RBAC, and zero-trust kubectl access enforced across every tenant.
AI/ML Ready	Integrated GPU operator and GPU scheduling for training and inference workloads out of the box.
Storage Integration	Turnkey CSI integration with Rook/Ceph, DDN, Weka, and Vast Data for persistent AI and ML storage.
Integrated Observability & Metering	Cluster and GPU metrics for SLA enforcement, with usage metering and billing for chargeback.
Self-Service Portal	Tenants request and operate clusters on demand, once the platform, tenancy, and governance are in place. Runs across public, private, air-gapped, and sovereign environments.

Business Outcomes

Outcome	Impact
Faster Path to a Live Service	Deliver a governed managed Kubernetes service in weeks, not months.
Standardization at Lower Operational Cost	Lifecycle automation removes config churn across clusters, and one control plane operates the fleet with a small platform team.
Enterprise & Sovereign Adoption	Zero-trust access, RBAC, and sovereign deployment open regulated markets hyperscalers cannot serve.