Rafay-powered Managed Kubernetes as a Service (K8SaaS)

GPU-accelerated, fully managed Kubernetes clusters optimized for AI and modern workloads.

Provide tenants with self-service access to Kubernetes clusters backed by GPUs, complete with governance, monitoring, and lifecycle automation while operators retain control and monetization levers.

From Bottleneck to Business Value

Enterprises and service providers want to monetize GPU investments while meeting the high bar enterprises set for Kubernetes. Yet running Kubernetes at scale is complex: multi-version upgrades, patching, availability, and security controls stretch engineering teams to their limits.

Tenants, meanwhile, expect cloud-like agility:

- Instant access to GPU-backed clusters.
- Elastic scaling without downtime.
- Compliance-ready environments without infrastructure burden.

Rafay Managed Kubernetes Clusters as a Service makes this possible. Providers can expose GPU-accelerated clusters as elastic SKUs with built-in governance, self-service provisioning, zero-trust access, and full lifecycle automation. The result: hyperscaler-grade Kubernetes experiences delivered locally, securely, and on your terms.

Designed For

Cloud & Service Providers: Add managed Kubernetes to your portfolio to expand margins and attract enterprise tenants.

Enterprises: Empower teams to build, deploy, and scale AI workloads on governed, GPU-backed Kubernetes without operational bottlenecks.

Sovereign & Regional Clouds: Deliver in-country Kubernetes clusters with zero-trust access, RBAC enforcement, and compliance-ready controls for regulated industries.

Key Capabilities

Capability	Description
End User Self-Service	Tenants launch and manage GPU-backed K8s clusters instantly via portal or API.
Multi-Version Support	Run multiple Kubernetes versions with automated upgrades and rollbacks.
High Availability Control Plane	Multi-master setup with automated failover and distributed etcd for reliability (99.9% SLA)
Automated Cluster Lifecycle	Provisioning, scaling, patching, and upgrades fully automated with zero downtime.
CNI Plugin Support	Flexible networking with Cilium, Calico, or Flannel for pod networking and policies.
RBAC & Security Policies	Enforce enterprise-grade security with RBAC, pod security standards, and network policy enforcement.
Rafay Zero Trust Kubectl	Secure developer access with centralized authorization, audit logging, and identity-based controls — no VPN or bastion required.
Load Balancer Integration	Native support for cloud load balancers and ingress controllers.
Latest GPUs	Support for NVIDIA, AMD, and Intel GPUs including H100, A100, and more.
AI/ML Ready	Pre-installed GPU operator for Kubernetes accelerates AI/ML workload setup.
Integrated Metrics	Real-time monitoring of cluster health, GPU utilization, and tenant workloads.
Metering	Low-granularity usage metering for accurate chargeback and billing.
Storage Integration	Turnkey integration with CSI providers like Rook/Ceph, DDN, Weka, and Vast Data.

Business Outcomes

Outcome	Impact
Accelerated Kubernetes Monetization	Launch fully managed GPU Kubernetes services in weeks, not years. Convert operational complexity into consumable SKUs with immediate tenant demand.
Higher Efficiency & Utilization	Multi-version support, lifecycle automation, and elastic scaling reduce operator overhead while maximizing GPU ROI.
Enterprise & Sovereign Adoption	Deliver secure, compliant clusters with zero-trust access, RBAC enforcement, and sovereign deployment options. Expand into regulated industries hyperscalers cannot serve.

© 2025 Rafay Systems, Inc. All rights reserved. All other trademarks and service marks are property of their respective owners.