RAFAY

Managed Kubernetes Clusters as a Service

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

Market Challenge, Target Audience & Solution

CSPs and Neoclouds face mounting pressure to monetize GPUs while delivering a Kubernetes experience that meets enterprise expectations. Running Kubernetes at scale is operationally complex: multi-version upgrades, patching, high availability, and security controls strain engineering teams.

Tenants, meanwhile, expect instant access, elastic scaling, and compliance-ready clusters without the burden of managing infrastructure.

Rafay Managed Kubernetes Clusters as a Service enables providers to expose fully governed, GPU-backed Kubernetes clusters as elastic SKUs. With self-service provisioning, zero-trust access, and automated lifecycle management, providers deliver a hyperscaler-grade experience while ensuring sovereignty, security, and monetization.

Target Audience

CSPs adding managed K8s services to expand portfolio margins.

Neoclouds offering sovereign K8s services for compliance-heavy 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 K8s 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	Hands-off provisioning, scaling, patching, and upgrades 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 via 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.
Load Balancer Integration	Native integration with 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 setup.

Business Outcomes

Outcome	Impact
Accelerated Kubernetes Monetization	Launch fully managed GPU K8s services in weeks, not years. Convert complex infrastructure into consumable SKUs with immediate tenant demand.
Higher Efficiency & Utilization	Multi-version support, 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 struggle to serve.