RAFAY

HPC Use Case

Rafay-powered SLURM as a Service (SLURMaaS)

Bring HPC job scheduling into the cloud era with GPU-accelerated, fully managed SLURM clusters.

With Rafay, customers can offer tenants secure, self-service access to on-demand SLURM environments, with governance, visibility, and lifecycle automation built in.

From Bottleneck to Business Value

SLURM (Simple Linux Utility for Resource Management) is the de facto standard for high-performance computing (HPC), but traditional deployments are siloed, static, and resource-intensive to manage. Operators struggle with:

- Complex cluster setup and manual lifecycle management.
- Underutilized GPU resources locked in dedicated HPC silos.
- Security and multi-tenancy gaps when scaling to multiple teams or tenants.

Rafay-powered SLURM as a Service bridges HPC and Kubernetes through Project Slinky, enabling providers to expose SLURM job scheduling as a fully managed, multi-tenant service. The result: HPC-grade scheduling with cloud-like agility, delivered securely across shared GPU clusters.

Designed For

Cloud & Service Providers: Expand into HPC and AI/ML markets by offering SLURM clusters as a service with sovereign deployment options.

Enterprises: Empower research and engineering teams with governed, GPU-backed SLURM environments that scale elastically.

Sovereign & Regional Clouds: Deliver SLURM-based services in-country, ensuring compliance, auditability, and secure multi-tenancy.

Key Capabilities

Capability	Description
Self-Service Access	Tenants launch SLURM clusters instantly via portal or API.
Kubernetes Integration	Project Slinky bridges SLURM with Kubernetes, enabling containerized HPC workloads.
GPU-Optimized Scheduling	Allocate GPUs dynamically across SLURM jobs with Kubernetes-backed orchestration.
Multi-Tenant Isolation	Secure separation of workloads across teams, projects, or customers.
Lifecycle Automation	Automated provisioning, scaling, patching, and teardown of SLURM environments.
Integrated Monitoring	Visibility into job performance, GPU utilization, and tenant usage.
Chargeback & Governance	Per-tenant usage tracking with built-in billing and policy enforcement.

Business Outcomes

Outcome	Impact
Accelerated Monetization of HPC Resources	Convert siloed HPC infrastructure into elastic, revenue-ready services.
Higher Utilization & Efficiency	Dynamically allocate GPUs across jobs, reducing idle capacity and maximizing ROI.
Enterprise & Sovereign Adoption	Deliver compliant, governed SLURM environments for research, engineering, and regulated industries.