RAFAY

Rafay-powered SLURM as a Service

Provision, schedule, and govern SLURM clusters as an on-demand service

Rafay enables organizations to deliver fully managed, multi-tenant SLURM environments for high-performance computing workloads.

The platform delivers SLURM as a cloud-like, on-demand service through automated, BCM-based cluster bring-up, with secure per-tenant separation and governance built in.

Why choose Rafay

Rafay delivers the complete SLURM service: automated BCM-based cluster bring-up, multi-tenant isolation, and self-service access for research and engineering teams.

- **Convert siloed infrastructure into elastic, revenue-ready services**. Operators move from static, single-team SLURM deployments to multi-tenant HPC services consumed on demand.
- **Support both traditional HPC and modern AI/ML workflows on shared infrastructure**. A single offering serves research, engineering, and AI teams without separate environments.
- **Deliver compliant, governed SLURM for research and regulated industries**. Secure per-tenant separation and governance meet the requirements of regulated and sovereign workloads.

Designed For

Neoclouds: Convert siloed HPC infrastructure into elastic, revenue-ready SLURM services consumed on demand.

Sovereign AI Clouds: Deliver compliant, governed SLURM in-region for research and regulated industries.

Enterprises: Give research and engineering teams self-service SLURM access without managing the cluster lifecycle.

Key Capabilities

Capability	Description
Multi-Tenancy	Supports multiple tenants or teams with isolation and operational control.
Heterogeneous GPU Support	Supports mixed GPU types like H100 and L40 for flexible AI/HPC workloads.
Simplified Deployment	Automates BCM-based SLURM bring-up and reduces operational complexity.
Ready-to-Use Access	Provides users with login nodes for easy job submission.
Flexible Topologies	Supports dedicated or colocated SLURM components based on scale, cost, and isolation needs.

Business Outcomes

Outcome	Impact
Expand Into HPC at Higher Utilization	Offer elastic HPC and research services on existing GPU infrastructure.
Lower Operational Overhead	Automated BCM-based bring-up removes manual HPC cluster management.
Revenue-Ready HPC	Per-tenant metering and billing turn HPC scheduling into a priced service.