RAFAY

Rafay-powered Bare Metal as a Service (BMaaS)

Deliver dedicated GPU servers with full host-level access, isolated and ready to use

Rafay gives customers direct access to the host operating system on dedicated GPU servers. Teams running large distributed training jobs, or those with highly opinionated software stacks, need control of the host that a virtualized environment cannot offer.

The platform ensures each server is delivered with secure access, networked with the other servers the customer is renting, and the right storage mounted, running with zero virtualization overhead for maximum performance on AI training, inference, and HPC.

Why choose Rafay

Rafay automates the full path from inventory to a dedicated, accessible GPU server, and operates the networking, isolation, storage, and access that a bare metal service requires.

- **Automated provisioning and bring-up.** Rafay matches the request against available inventory, boots and discovers the server, installs the operating system, and configures drivers and services.
- **Networking and tenant isolation, configured automatically.** Rafay configures north-south connectivity through the Tenant Access Network, so servers reach each other, the tenant's VPC, and the internet, and connects multiple rented servers east-west over InfiniBand for GPU-to-GPU communication. Separately, it isolates each tenant with EVPN on the Ethernet fabric and Pkey on InfiniBand via NVIDIA UFM, so one tenant's traffic never reaches another's.
- **Storage, access, and firewall control.** Rafay provisions per-tenant storage and mounts it to the server, enables SSH access via password or key, assigns a public IP from the tenant pool, and automates security groups and NAT rules to open ports for authorized internet-based users.

Designed For

Neoclouds: Offer dedicated bare metal GPU servers as premium SKUs, moving beyond commodity rentals with performance SLAs, metering, and billing.

Sovereign AI Clouds: Deliver in-region, air-gapped, compliant bare metal GPU services for regulated industries and national AI projects.

Enterprises: Give AI and HPC teams direct, host-level access to dedicated GPU servers for large training jobs and opinionated software stacks.

.

Key Capabilities

Capability	Description
Automated Bare Metal Bring-Up	Match servers from inventory, reserve IPs, boot and discover, install OS, and configure drivers and services through one workflow.
Network Automation & Tenant Isolation	Per-tenant VRFs, north-south tenant access, and optional firewall rules (security groups), programmed automatically as tenants onboard.
High-Speed Interconnects	InfiniBand and Ethernet with RDMA for low-latency distributed training and HPC.
Storage Integration	Attach high-performance network storage alongside direct-attached local NVMe for I/O-intensive workloads.
Hardware Inventory Management	Centralized inventory and data-center-aware placement across multiple locations.
Latest GPU Hardware	Expose NVIDIA H100, A100, L40S and more as premium bare metal SKUs.
Multi-GPU Configurations	Scale up to 8 GPUs per server with NVLink and NVSwitch for training and HPC.
Pre-Configured Images	CUDA, cuDNN, and ML frameworks ready out of the box.

Business Outcomes

Outcome	Impact
Faster Path to a Live Service	Stand up a governed, multi-tenant bare metal service in weeks, not the months a hand-built platform requires.
Accelerated GPU Monetization	Reallocate expensive GPUs to the next tenant in minutes, not weeks. Elastic provisioning and per-second billing maximize utilization and reduce depreciation risk.
Higher Margins & Differentiation	Move beyond commodity rentals with premium SKUs. Performance SLAs, analytics, and integrated billing support stronger price points.