Rafay-powered Bare Metal GPUs as a Service (BMaaS)

Turn GPU infrastructure into differentiated, high-value services with cloud-like agility and full control.

Dedicated, bare metal GPU servers with zero virtualization overhead deliver maximum performance for AI training, inference, and high-performance computing workloads, all while giving organizations complete control.

From Bottleneck to Business Value

Organizations everywhere are making significant investments in GPU infrastructure, yet too often these assets sit underutilized or trapped in multi-week (re)allocation procedures, or rigid rental models. The result: bad user experience, chronic underutilization, and shrinking margins.

What users want is clear:

- Instant, self-service provisioning with cloud-like agility.
- Transparent consumption models that align cost with value.
- Deployment options that respect sovereignty, compliance, and security requirements.

Rafay's **Bare Metal GPUs as a Service** offering makes this possible. With Rafay, organizations can transform raw GPU racks into differentiated, revenue-generating services. Elastic, dedicated servers come with governance, visibility, and billing built in, delivering hyperscaler-grade experiences, locally and securely.

Designed For

Cloud & Service Providers: Modernize infrastructure portfolios with GPU-as-a-Service offerings that go beyond commodity rentals.

Enterprises: Empower internal teams to accelerate AI, data science, and HPC projects with instant, governed access to dedicated GPUs.

Sovereign & Regional Clouds: Build in-region GPU services that meet data residency, compliance, and national AI requirements while retaining full control.

Key Capabilities

Capability	Description
Self-Service Portal	Tenants provision bare metal GPU servers instantly, without manual intervention.
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/NVSwitch for training and HPC.
Pre-Configured Images	CUDA, cuDNN, and ML frameworks ready out-of-the-box for immediate use.
NVMe Storage	High-performance SSDs for I/O-intensive workloads.
Real-Time Monitoring	Visibility into usage, GPU health, and power consumption for SLA enforcement.
Precise Usage Metering	low-granularity metering for accurate chargeback and revenue maximization.
Public Connectivity	High-bandwidth public IPs for external access.
Zero Overhead	Direct hardware access with no hypervisor tax.
High-Speed Interconnects	InfiniBand & Ethernet with RDMA for distributed training and HPC.

Business Outcomes

Outcome	Impact
Accelerated Monetization of GPUs	Reallocate expensive to the next customer in minutes, not weeks. Elastic provisioning and low-granularity 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 and defensible differentiation.
Enterprise & Sovereign Adoption	Meet the needs of regulated industries and sovereign AI projects with in-region, air-gapped deployments — competing in markets where hyperscalers cannot.