RAFAY

Rafay-powered Virtual Machines as a Service (VMaaS)

Deliver GPU-accelerated VMs and storage as secure, scalable, consumption-based services.

Provide tenants with instant access to compute capacity for AI, ML, and HPC workloads, complete with governance, visibility, and monetization controls.

From Bottleneck to Business Value

Many organizations start their GPU cloud journey with virtual machines. But basic VM rentals quickly become a commodity: margins erode, provisioning is slow, and customer expectations go unmet.

Today's tenants demand more:

- Fast, self-service provisioning with cloud-like agility.
- Flexible GPU configurations and usage-based pricing.
- Strong network isolation, security, and compliance guarantees.

Rafay Virtual Machines as a Service (VMaaS) transforms raw GPU/CPU infrastructure into fully managed, tenant-ready VM SKUs. Organizations can deliver elastic, multi-GPU VMs with integrated monitoring, billing, security, and lifecycle automation, competing with hyperscalers while capturing sovereign AI and compliance-driven opportunities.

Designed For

Cloud & Service Providers: Expand beyond bare metal with scalable VM-based services that improve utilization and margins.

Enterprises: Empower teams to host, test, and prototype AI/ML projects on GPU-backed VMs without operational bottlenecks.

Sovereign & Regional Clouds: Deliver sovereign-ready VM services in-country with VPC isolation, data residency, and compliance built-in.

Key Capabilities

Capability	Description
End User Self-Service	Intuitive tenant portal to configure, launch, and use GPU-backed VMs instantly.
Latest GPU Models	Support for NVIDIA, AMD, and Intel GPUs, including H100, A100, and more.
Scaling	Allocate 1–8 GPUs per VM for training, inference, or HPC workloads.
AI/ML Ready OS	Pre-installed GPU/network drivers and ML images accelerate onboarding.
Storage	Turnkey integration with Ceph, DDN, Weka, Vast, Dell PowerStore, HPE Alltera, and others.
Integrated Metrics	Real-time monitoring of VM performance and GPU usage.
Metering	Low-granularity usage accounting for accurate billing and chargeback.
VPCs	Launch VMs inside isolated Virtual Private Clouds with unique CIDR blocks.
Public IPs	Assign direct public IPs without bastion or NAT.
High Performance	NVIDIA Reference Architecture compliance ensures optimal GPU passthrough.
vGPU Support	Share GPUs across multiple VMs with NVIDIA vGPU.
SXM GPU Support	Expose NVIDIA SXM-based GPUs (H100 SXM, A100 SXM) for maximum performance.
Routing	Intelligent traffic routing and load balancing across workloads and regions.
Firewall	Granular firewall rules and network protection for secure deployments.
VM Lifecycle Management	Automated provisioning-to-decommissioning workflows simplify operations.

Business Outcomes

Outcome	Impact
Accelerated Service Monetization	Convert idle infrastructure into cloud-grade VM services in weeks. Elastic provisioning and usage-based billing drive faster ROI on GPU/CPU investments.
Higher Utilization & Differentiation	Offer premium VM SKUs with scaling, GPU sharing, SXM support, and advancednetworking. Move beyond commodity rentals into differentiated, high-margin services.
Enterprise & Sovereign Expansion	Deliver sovereign-ready VM services with VPC isolation, firewall policies, and lifecycle automation. Capture regulated industries and workloads hyperscalers cannot serve.