RAFAY

Rafay-powered Virtual Machines as a Service (VMaaS)

Deliver GPU-accelerated virtual machines as secure, scalable, consumption-based services

Rafay enables organizations to deliver GPU- and CPU-based virtual machines as secure, scalable, consumption-based services.

The platform transforms existing compute capacity into fully managed, tenant-ready VM SKUs, with governance, visibility, and automated lifecycle control, built on mature, battle-tested KVM virtualization.

Why choose Rafay

Rafay transforms existing compute capacity into high-performance VM SKUs that tenants consume on demand, with faster service rollout, higher utilization, and greater operational efficiency.

- **Deliver near bare-metal performance.** NUMA topology-aware GPU passthrough on KVM provides full GPU performance with the flexibility of virtual machines, validated against the NVIDIA reference architecture.
- **Isolate tenants at the hardware level.** NVIDIA DPU-based hard tenancy separates VM and storage traffic, with SELinux-enforced security boundaries and per-tenant network and storage isolation.
- **Consume on demand, with metering built in.** Tenants self-provision GPU VMs through the portal, allocating one to eight GPUs per VM, with per-second metering and billing.

Designed For

Neoclouds: Offer GPU VMs as flexible, metered SKUs that tenants provision on demand, expanding the portfolio beyond bare metal rentals.

Sovereign AI Clouds: Deliver in-region GPU VM services with hardware-level tenant isolation and full data residency control.

Enterprises: Give internal teams instant, governed access to GPU-accelerated VMs for training, inference, and HPC, with near bare-metal performance.

*Aligned with the NVIDIA VM Performance Reference Architecture and NVIDIA Cloud Partner (NCP) software reference architecture

Key Capabilities

Capability	Description
Automated VM Provisioning & Smart Scheduling	Intelligent matchmaking places each VM on the right server by NUMA topology and GPU layout, then provisions and configures it through one workflow.
Network Automation & Tenant Isolation	Per-tenant VRF-based isolation, public and private IPAM, and automated firewall and NAT rules, programmed as tenants onboard. VRFs can be created dynamically through the Netris integration.
Hardware-Level Multi-Tenancy	NVIDIA DPU-based hard tenancy isolates tenant VM and storage traffic, with DPU acceleration for both. Confidential VMs add TEE-secured execution for sensitive workloads.
NVIDIA Reference Architecture Performance	NUMA topology-aware GPU passthrough delivers near bare-metal performance, validated on H100, B200, B300, A100, A10, and L40S.
Flexible GPU SKUs & Sharing	Define VM SKUs with 1, 2, 4, or 8 GPUs, including SXM GPUs and ARM guest VMs on GB200 and GB300, plus vGPU sharing for higher density.
GPU VM Clustering	Cluster GPU VMs across nodes over InfiniBand with GPUDirect RDMA for low-latency, multi-node distributed training and HPC.
Storage Integration & Lifecycle	Turnkey boot, block, and shared file storage across Ceph, DDN, Weka, Vast, and Dell PowerStore, with ephemeral disk support and VM migration for maintenance and lifecycle operations.
Self-Service with Metering	Tenants provision, access through an in-browser console, and manage VMs on demand through the portal, with per-second usage metering and billing built in.

Business Outcomes

Outcome	Impact
Faster Path to a Live Service	Stand up a governed, multi-tenant GPU VM service in weeks, not the months a hand-built platform requires.
Near Bare-Metal Performance	NUMA-aware passthrough delivers full GPU performance with the flexibility of VMs.
Higher Utilization & Accurate Chargeback	Flexible SKUs and vGPU sharing drive density across the fleet, with per-second metering and per-tenant billing for precise cost recovery.

RAFAY
Rafay Systems, Inc. | 530 Lakeside Dr. Ste 220 Sunnyvale, CA 94085 | 669.336.4800 | rafay.co