JioCloud

NVIDIA GPU VM

Overview

JioCloud GPU Virtual Machine NVIDIA is built to supercharge your most demanding AI, deep learning, and HPC workloads at enterprise scale. Powered by NVIDIA H200 NVL GPUs, each VM supports up to 8 GPUs with 141 GB HBM3e memory and 4.8 TB/s NVLink bandwidth. Teams can train, fine-tune, or infer with speed, flexibility, and control. With MIG (Multi-Instance GPU), you can split GPUs for granular workloads. CUDA and AI frameworks are pre-installed, so setup is minimal. All usage is observable in real time, and billing remains transparent with both on-demand and reserved plans.

Key Features

- **NVIDIA H200 NVL GPUs**
  Built for AI with 141 GB HBM3e memory and high NVLink bandwidth.
- **Multi-GPU scaling**
  Configure 1 to 8 GPUs per VM based on your training or inference needs.
- **MIG support**
  Divide a GPU into virtual slices for smaller tasks and cost-efficient inference.
- **Ready-to-use environments**
  Includes prebuilt images with PyTorch, TensorFlow, and CUDA drivers.
- **Lifecycle control via API or portal**
  Start, stop, pause, delete, or snapshot VMs easily.
- **Real-time observability**
  Monitor GPU, memory, power usage, and VM health with built-in tools.
- **Flexible pricing plans**
  Choose on-demand or reserved options (1, 3, 6, or 12 months) to control costs.
- **Instant availability**
  Access GPUs without procurement delays or queue times.

JioCloud

Benefits

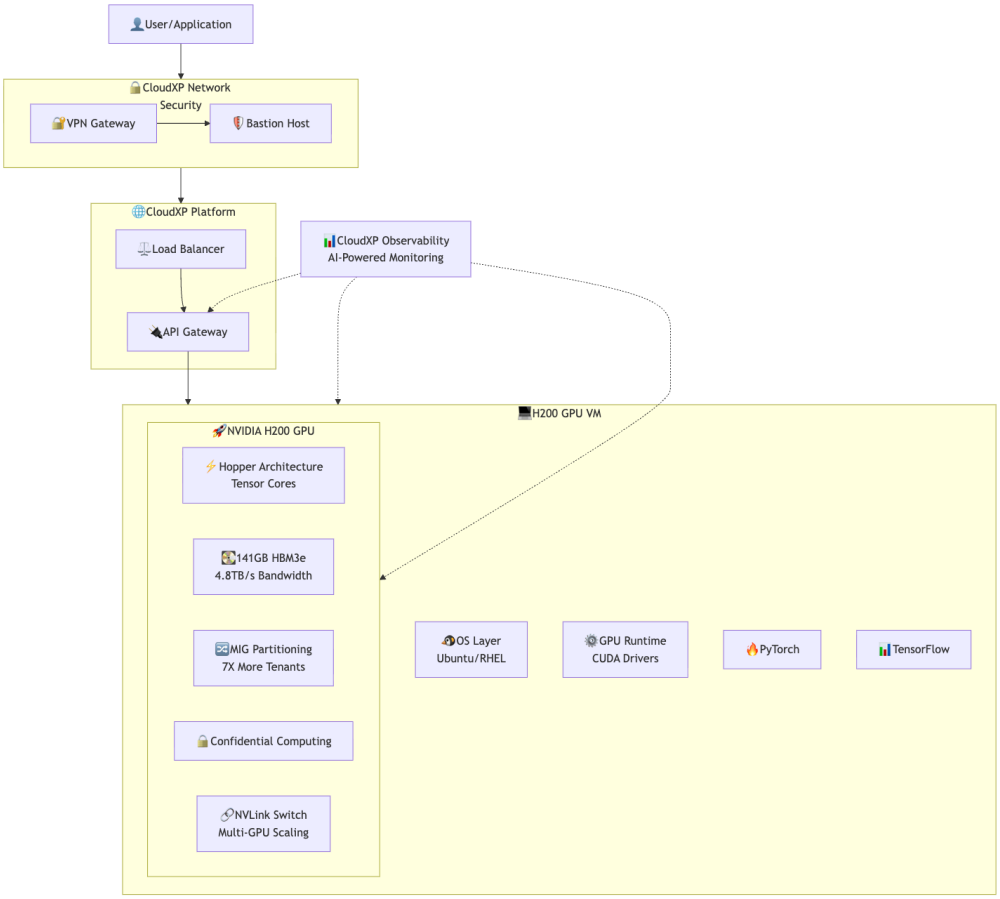
- Train and fine-tune large AI models faster.
- Process image, text, and tabular data more efficiently.
- Skip setup delays with pre-installed environments.
- Scale compute from small inference to full training runs.
- Control costs with flexible pricing and no lock-ins.
- Improve performance with real-time workload insights.

Technical Specifications

Specification	Details
GPU Architecture	NVIDIA Hopper
Memory	141 GB HBM3e
Memory Bandwidth	Up to 4.8 TB/s
GPU Interconnect	NVLink, NVSwitch
Compute Performance	TFLOPS (FP16/FP8/INT8) *
Supported OS	Ubuntu, RHEL
AI Frameworks	PyTorch, TensorFlow
Storage	100GB
Power Efficiency	Optimized performance-per-watt

JioCloud

Architecture Diagram



Use Cases

- **E-commerce recommendations**
  Train personalization engines quickly and scale inference during high-traffic periods.

- **Medical research and genomics**
  Speed up deep learning models for drug discovery and real-time clinical prediction.

- **Climate simulation and forecasting**
  Run high-volume climate models and reduce simulation times with multi-GPU setups.

- **Kubernetes log chatbots**
  Use AI-driven chat interfaces to analyze and troubleshoot K8s logs using plain English.