

NVIDIA GPU VM

Accelerate LLMs and AI Workloads —
with NVIDIA H200 NVL GPU VMs

JioCloud GPU Virtual Machine NVIDIA gives you instant access to fully managed GPUs - with MIG flexibility, observability, and zero setup delays.

The Challenge

- Training large models like LLMs takes weeks, delaying product timelines.
- Image, video, and medical models need high compute that's often unavailable.
- Processing large datasets on CPUs is slow and resource intensive.
- Physical GPU setups add complexity in procurement, power, and maintenance.
- Limited cloud GPU availability causes provisioning delays.
- Teams spend time configuring CUDA, frameworks, and dependencies manually.
- Scaling infrastructure mid-project is often rigid and difficult.
- Poor GPU-level visibility makes it hard to optimize workloads.



The JioCloud Solution

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.

What You Gain

- 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.



Use Cases in Action

eCommerce 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.

Who It's For

AI/ML engineers and data scientists
MLOps, DevOps, and data platform teams
HPC and simulation researchers
CIOs and infrastructure leaders
Academic and research institutions

Why JioCloud

Fully managed GPU VMs: Focus on your models, not your infrastructure. We handle the complexities of provisioning, monitoring, and lifecycle management.

Lowest-cost H200 NVL access

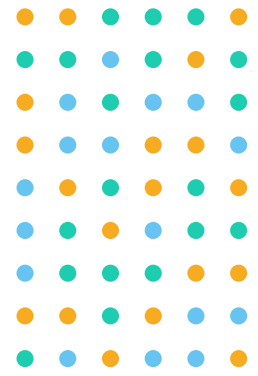
Achieve more for less. We offer the most competitive pricing for NVIDIA H200 NVL power, with both on-demand and reserved instance plans to optimise your costs.

Guaranteed GPU availability

Eliminate procurement delays. Scale your GPU resources instantly to meet project demands, guaranteed.

Enterprise-grade infrastructure

Build on a secure, high-performance cloud platform with deep integration across the JioCloud stack for a seamless workflow.



Build, Train, and Scale with JioCloud GPU Virtual Machine NVIDIA

Contact us at jpl.cloudsales@ril.com or visit [\(website\)](#) to get started.

