Bare Metal NVIDIA

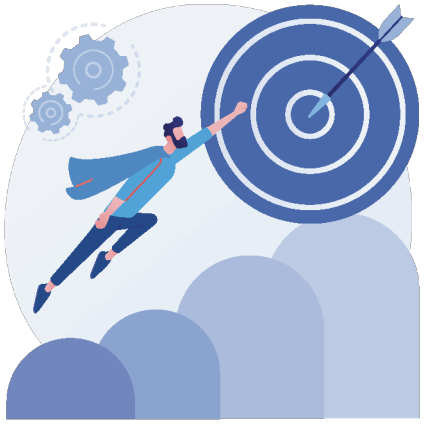
Run AI at Full Power — with 8x H200 SXM Bare-Metal GPU Servers

JioCloud

Get peak performance, zero overhead, and full control — for training, simulation, and mission-critical workloads.

The Challenge

- Conventional CPU infrastructure lacks the memory and parallelism needed for LLMs, simulations, and high-performance AI.
- Virtualised GPU environments add latency, throttle performance, and restrict access to critical GPU features.
- Shared infrastructure leads to quota delays, fragmented memory, and unpredictable runtime behaviour.
- Sensitive workloads in regulated sectors require dedicated, physically isolated hardware to meet compliance and security needs.



The JioCloud Solution

JioCloud Bare-Metal GPU Servers are purpose-built for the most demanding AI and HPC use cases — with full access to 8x NVIDIA H200 SXM GPUs per node, 141 GB HBM3e per GPU, and 4.8 TB/s NVLink bandwidth.
There is no hypervisor, no container runtime interference — just direct-to-metal performance. Ideal for training LLMs, running simulations, or deploying inference at scale, with full telemetry and lifecycle control via API or dashboard. Whether you are running 24x7 AI pipelines or short-burst experiments, JioCloud ensures your workload gets the power, speed, and consistency it demands — every time.

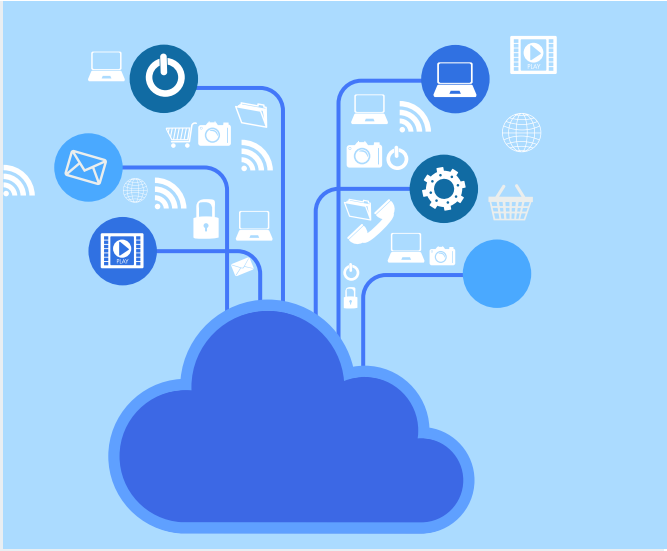
Key Features

- **8x NVIDIA H200 SXM GPUs per server**
  Hopper architecture, 141 GB HBM3e per GPU, withNVLink interconnect for massive parallelism.
- **No virtualization, no overhead**
  Run on bare metal for consistent, unthrottled performance across long-running jobs.
- **Physically isolated infrastructure**
  No shared tenancy. Fully dedicated servers ideal for compliance and high-sensitivity AI.
- **Direct NVLink bandwidth (4.8 TB/s)**
  Enable fast tensor parallelism and multi-GPUtraining without communication delays.
- **MIG and time-slicing support**
  Facilitate concurrent inferencing or multi-tenant GPU usage on demand.
- **Full telemetry and observability**
  Monitor GPU utilisation, power draw, memory bandwidth, thermals, and performance metrics in real time.
- **Preinstalled CUDA stack**
  Train or deploy on day one — no setup delay.

JioCloud

What You Gain

- **Peak AI performance**
  With 8x H200 SXM GPUs and full NVLink.
- **Consistent compute**
  With no virtualisation, no sharing, and zero throttling.
- **Mission-critical readiness**
  With physical isolation and compliance alignment.
- **Faster time-to-training**
  With preconfigured, ready-to-use GPU servers.
- **Advanced debugging and RCA**
  Via full-stack GPU and workload telemetry.
- **No quotas, no wait time**
  Your full-GPU servers are always ready.



Use Cases in Action

eCommerce - vision-language models for virtual try-on

Train and deploy AI-powered try-on and fit recommendation models with full memory and bandwidth access for fast, photorealistic rendering.

Healthcare - genomics and medical imaging

Run sensitive workloads like 3D image segmentation and genome sequencing on isolated, compliant infrastructure with consistent throughput and full observability.

CloudOps - real-time cluster monitoring with LLMs

Deploy LLMs for autonomous Kubernetes monitoring and real-time root cause analysis without inference jitter or infrastructure noise.

Who It's For

- AI/ML Engineering Teams
- Data Scientists and LLM Modelers
- HPC and Research Labs
- CloudOps and Platform Teams
- Regulated Workload Owners (Healthcare, BFSI, Government)

Why JioCloud

- **Purpose-built for bare-metal H200 SXM** - No retrofits or shared infra — fully optimised for full-GPU workloads on 8x H200 SXM with NVLink.
- **Cost advantage without compromise** - Get bare-metal performance at lower total cost than public cloud equivalents.
- **Billing aligned to real usage** - Choose from on-demand, reserved, or capped billing — with full transparency and namespace-level control.
- **Developer-friendly operations** - Configure GPU quotas, MIG, and scheduling with Kubernetes-native tools — no node-level tuning required.
- **End-to-end observability** - Monitor GPU and workload metrics across power, thermals, and utilisation — to tune performance and resolve issues faster.

JioCloud

Get Bare-Metal Speed, Security, and Scale - All in One Stack

Reach us at jpl.cloudsales@ril.com or explore deployment options at (website) to get started.