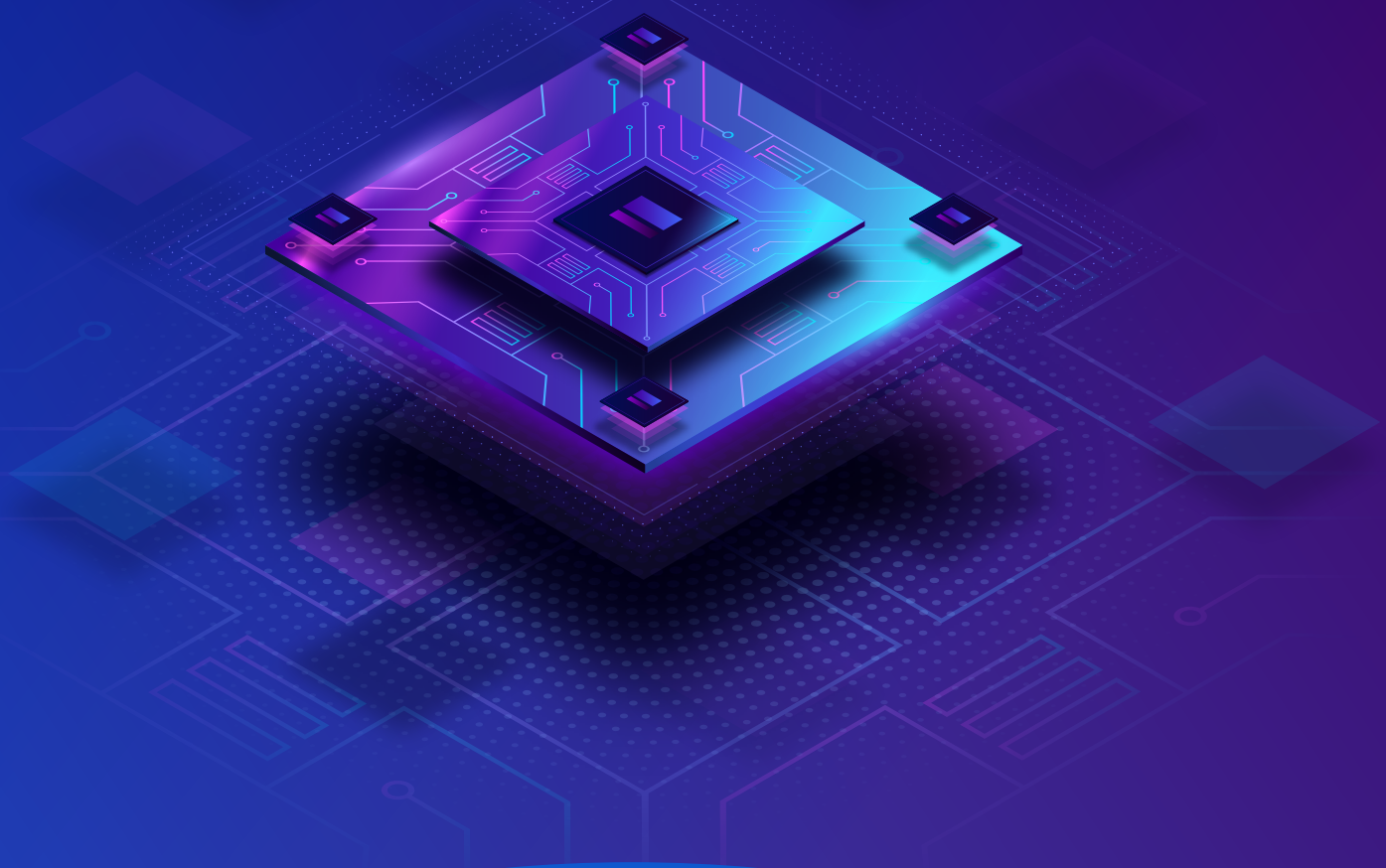
JioCloud



GPU Worker Node AMD

From Training to Deployment—
Run AI Workloads Faster on AMD GPUs

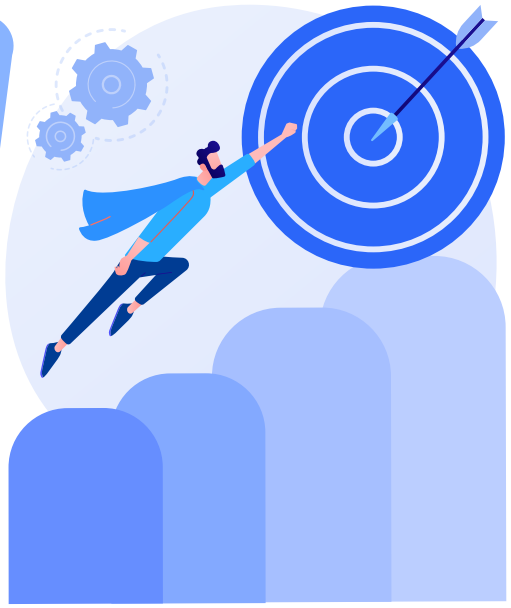
JioCloud

Power your AI journey with high-performance, flexible, and scalable GPU compute — built for training speed and deployment agility.

The Challenge

AI and simulation teams need high-performance computing, but traditional solutions stifle innovation with significant bottlenecks:

- CPU nodes are too slow for intensive workloads like LLMs, image processing, or simulations.
- GPU VMs are inflexible, introduce latency, and face quota or provisioning delays.
- Resources go underutilised due to fixed configurations, cold starts, and poor visibility.
- Manual setup of drivers, runtimes, and observability tools adds delays and overhead.



The JioCloud Solution

JioCloud AMD GPU Compute Node offers a Kubernetes-native, high-performance **K8S worker Node** solution built on AMD MI300X GPUs. It eliminates the delays, inefficiencies, and scaling issues of traditional CPU nodes and standard GPU-based virtual machines. You get direct access to powerful GPU VMs inside your Kubernetes cluster — optimised for performance, free from cold starts, or quota restrictions.

Select from flexible configurations - 1, 2, 4, or 8 GPUs per worker node — depending on your workload size and urgency. From large-scale training to real-time inference and simulation, the setup scales to match your team's needs.

Teams can run multiple jobs on the same hardware without compromising performance or security. With ROCm drivers pre-installed, deep observability built in, and native integration with Kubernetes, you can move from setup to production without delay.

JioCloud

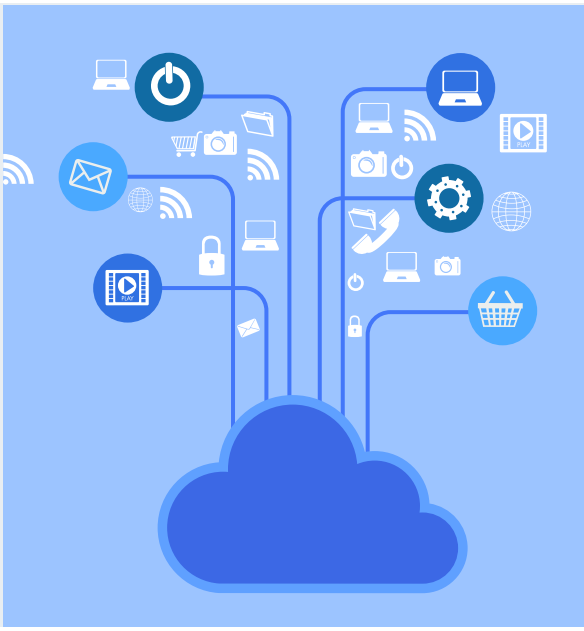
Key Features

- **High-performance GPU VMs**
  Provision powerful AMD MI300X GPU virtual machines directly into a Kubernetes cluster for maximised performance.
- **AMD MI300X**
  Industry-leading 192GB HBM3 memory with exceptional bandwidth, optimised for memory-intensive analytics and large-scale AI workloads.
- **Configurable VM sizes**
  Right-size your resources. Deploy GPU VMs with 1, 2, 4, or 8 GPUs to perfectly match job size and budget.
- **Autoscaling with kubernetes**
  Dynamically scale GPU resources up/down with industry-standard tools like Cluster Autoscaler or Karpenter.
- **Integrated observability**
  Monitor memory, usage, and power across jobs workloads and namespaces.
- **Pre-installed ROCm drivers and plugins**
  Start training or inference immediately on a fully configured platform — no manual setup.

Run AI, simulation, and data workloads at scale on Kubernetes — with high-performance VMs, zero cold starts, and full GPU control.

What You Gain

- **Accelerate training and inference**
  Deliver LLMs, generative AI, or real-time inference workloads faster.
- **Maximise GPU utilisation**
  Use MIG or time-slicing to run multiple jobs on the same GPU.
- **Lower cost per job**
  Choose the right-sized worker node, avoid overprovisioning, and scale only when needed.
- **Full GPU observability**
  Track usage metrics at the job, pod, or namespace level.
- **Secure multi-tenant sharing**
  Isolate workloads by team using namespace limits and GPU slices.
- **Run anywhere, scale freely**
  Avoid vendor lock-in and grow across your preferred Kubernetes environments.



JioCloud

Use Cases in Action

Healthcare - medical imaging with AMD MI300X

A research hospital uses AMD MI300X VMs to run 3D CNNs on MRI and CT scans. High-performance virtualisation speeds up diagnosis and genomics research while ensuring compliance through multi-tenant isolation and audit trails.

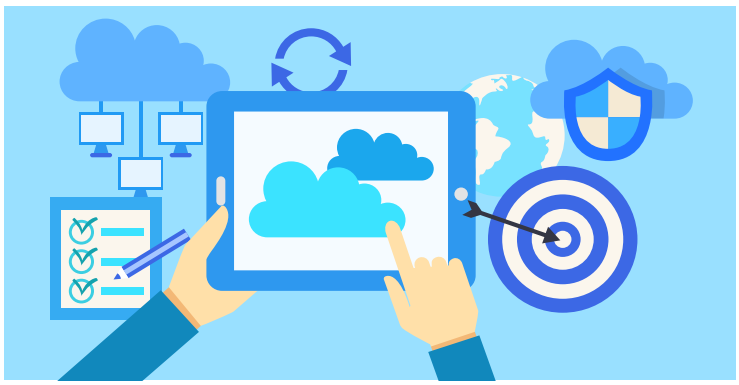
Scientific research - climate modeling with ROCm

A climate research institute leverages AMD MI300X's massive memory capacity for large-scale atmospheric simulations. The 192GB memory enables processing of detailed climate models that would exceed traditional GPU memory limits.

Financial analytics - risk modeling with memory-intensive workloads

A quantitative trading firm uses AMD GPU VMs for Monte Carlo simulations and portfolio optimisation. The superior memory bandwidth handles complex mathematical models across multiple market scenarios simultaneously.

Who It's For



- AI/ML Engineers and Data Scientists
- MLOps, DevOps, and Platform Teams
- Simulation and HPC Researchers
- Academic Institutions and Research Labs
- CIOs, Infra Leads, and GPU Platform Owners

Why JioCloud

- **Lower infra cost by design -** JioCloud owns the full stack — delivering GPU performance at better economics than public cloud VMs.
- **Full-stack control and faster provisioning -** Get tightly integrated orchestration, monitoring, and GPU scheduling from day one.
- **Guaranteed availability at scale -** Bypass public cloud quotas and delays with dedicated capacity and SLAs.
- **Cloud-native, kubernetes-first -** Centralised management, governance, and scaling via our self-serve cloud platform.
- **Secure, auditable GPU usage -** Configure quotas per namespace, monitor MIG assignments, and ensure workload separation with built-in controls.

JioCloud

Deploy High-Performance AMD GPU VMs—Without the Overhead

Talk to us at jpl.cloudsales@ril.com or visit (website) to provision your AMD GPU Compute Nodes with JioCloud today.