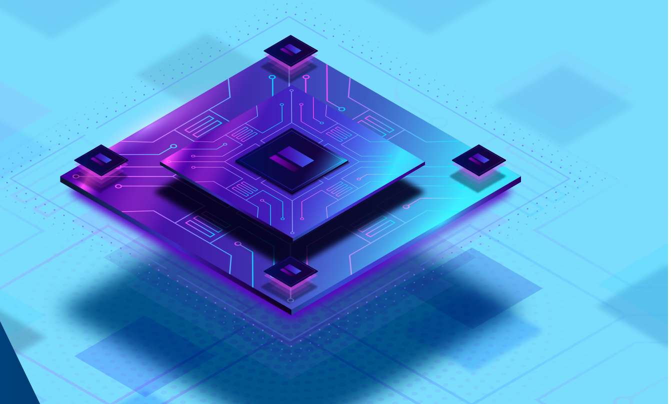
JioCloud

GPU Worker Node AMD



Overview

JioCloud NVIDIA GPU Compute Node offers a Kubernetes-native, high-performance K8s Worker Node solution built on NVIDIA H200 NVL 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—optimized 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. JioCloud makes GPU sharing simple and safe, with built-in support for Time Slicing and MIG (Multi-Instance GPU). Teams can run multiple workloads on the same hardware without compromising performance or security. With CUDA drivers pre-installed, deep observability built in, andnative i ntegration with Kubernetes, you can move from setup to production without delay.

Key Features

- **High-performance GPU worker nodes**
  Provision powerful NVIDIA H200 NVL GPU worker nodes directly into a Kubernetes cluster for maximised performance.
- **NVIDIA H200 NVL**
  Purpose-built for AI workloads with 141GB HBM3e memory and exceptional bandwidth for memory-intensive analytics and LLM training.
- **Configurable worker node sizes**
  Right-size your resources. Deploy GPU VMs with 1, 2, 4, or 8 GPUs to perfect lmatch job size and budget.
- **GPU concurrency with MIG and time slicing**
  Maximise hardware value by running multiple workloads securely and efficiently on shared NVIDIA GPUs.

- **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 workloads and namespaces.
- **Pre-installed CUDA drivers and plugins**
  Start training or inference immediately on a fully configured platform—no manual setup.

Benefits

- **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.

Technologies Supported

Specification	NVIDIA H200 NVL	AMD MI 300X
GPU Architecture	NVIDIA Hopper	AMD CDNA 3
Memory	141 GB HBM3e	128 GB HBM3
Memory Bandwidth	Up to 4.8 TB/s	Up to 3.2 TB/s
GPU Interconnect	NVLink, NVSwitch	Infinity Fabric, XGMI
Compute Performance	TFLOPS (FP16/FP8/INT8)	TFLOPS (FP64/FP32/FP16)
MIG (Multi Instance GPU)	Profile Based Single Strategy	NA
Power Efficiency	Optimized performance-per-watt	

Use Cases

- **eCommerce - real-time personalisation with MIG**
  An online retailer uses NVIDIA H200 NVL GPU VMs to train transformer-based recommendation models while simultaneously serving live traffic via MIG slices. This reduces iteration time and improves customer experience without underutilising GPUs.

- **FinTech - fraud detection with CUDA acceleration**
  A financial services company deploys NVIDIA GPU VMs to run real-time fraud detection models on transaction streams. Time-slicing enables multiple risk models to share GPUs efficiently while maintaining sub-millisecond response times.

- **Gaming - AI-driven content generation**
  A game studio uses NVIDIA H200 NVL VMs for AI-assisted asset generation and procedural content creation. MIG partitioning allows simultaneous training of different AI models for textures, animations, and level design.