JioCloud

Bare Metal NVIDIA

Overview

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, with NVLink 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-GPU training without communication bottlenecks.
- **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

Benefits

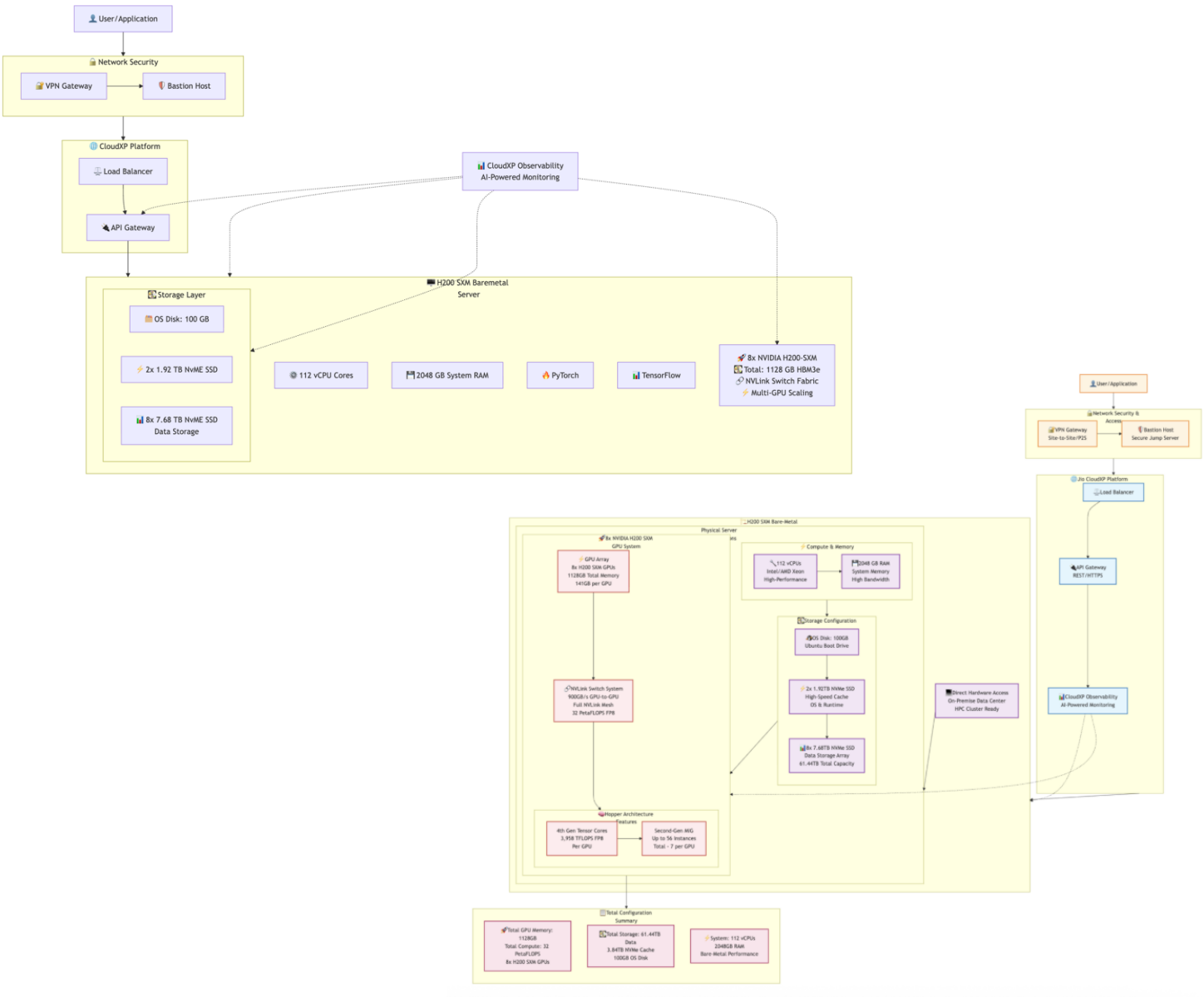
- **Peak AI performance**
  With 8x H200 SXM GPUs and full NVLink.
- **Consistent compute**
  With no virtualisation, no sharing, and zero throttling.
- **Mission-critical readiness**
  With physical isolation and compliance alignment.
- **Faster time-to-training**
  With preconfigured, ready-to-use 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.

Technologies Supported

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 (FP64/FP32/FP16)
Supported OS	Ubuntu
Storage	100 GB
Power Efficiency	High performance-per-watt
Deployment	BareMetal Server

JioCloud

Architecture Diagram



Use Cases

- **E-commerce - vision-language models for virtual try-on**
  Train and deploy AI-poweredtry-on and fit recommendation models with full memory and bandwidth access for fast, photorealistic renderingload.
- **Healthcare - genomics and medical imaging**
  Run sensitive workloads like 3D image segmentation and genome sequencing on isolated, compliant infrastructure with consistent throughput and fullobservability.
- **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.