NVIDIA GB200 NVL72
on CoreWeave

Ready to harness the full potential of next-generation AI? CoreWeave Cloud is purpose-built for AI workloads to maximize performance, with incredible resilience and reliability.

CoreWeave is the first cloud provider to offer general availability of NVIDIA GB200 NVL72–based instances

- Available as bare metal instances through CoreWeave Kubernetes Services (CKS)
- Advanced scheduling mechanisms to help customers provision all 72 GPUs in a single rack
- Support for Slurm plugin in SUNK for intelligent job scheduling
- Paired with CoreWeave Tensorizer to enable faster check-pointing for training runs

Purpose-built AI Infrastructure for high performance

Advanced failure recovery

- Comprehensive observability tools for real-time performance insights at rack level
- Proactive node health checking effectively manages failures

Efficient resource usage

- Topology-aware scheduling with Slurm on Kubernetes (SUNK)
- Liquid cooling enables denser configurations and reduces operational costs

Superior performance & scale

- Advanced scaling to 100k+ GPU clusters with NVIDIA Quantum-2 InfiniBand interconnect for maximum throughput
- 13.5TB memory per rack, delivering ultra-fast inter-GPU communication
- Integrated with CoreWeave AI Object Storage, for performance levels up to 2 GB/s per GPU

Innovate faster with a reliable and resilient AI-first platform

- Extensive automated cluster validations for cluster readiness on Day One
- Early access to cutting-edge compute
- Comprehensive observability tools for real-time performance insights
- 24/7 MLOps and engineering support to alleviate the burden of troubleshooting and infrastructure management

Learn more

Transform your AI workloads with purpose-built infrastructure today by reaching out to us here.

coreweave.com