SOLUTION BRIEF

GPU COMPUTE

CoreWeave

NVIDIA HGX B300 on CoreWeave Cloud

Unlock performance at scale for agentic AI.

Connect with CoreWeave

Explore all of the benefits the Essential Cloud for AI can bring to your most critical workloads.

Let's talk

Photo courtesy of Switch © 2026. All rights reserved.

Proven performance. Operational confidence.

NVIDIA HGX B300 represents a major architectural leap: doubling interconnect speed, expanding GPU memory, and redefining what's possible for agentic AI. CoreWeave's AI-native cloud is purpose-built to unlock this performance from day one, so you can run your production-scale frontier workloads with confidence.

NVIDIA HGX B300 key use cases

- AI reasoning and agentic systems
- Large-scale model training
- Complex, high-throughput inference

CoreWeave Cloud, your force multiplier

— **Partnership.** Blackwell-ready architecture co-designed and optimized with NVIDIA and direct-to-expert support trusted by AI leaders

— **Pace.** Reduced iteration cycles, faster issue resolution, and shorter path to production

— **Performance.** Memory and interconnect designed for scale and higher uptime driven by liquid cooling

Purpose-built for agentic AI

3.42x
higher token generation on Kimi K2.5

2.61x
higher token generation on DeepSeek-R1

1.95x
faster measured collective bandwidth on DeepSeek-R1

Built to run agentic AI at scale

NVIDIA HGX B300 represents a substantial step-up from previous generations within the same HGX form factor. Expanded memory capacity, inference throughput, and interconnect speed mean you can deploy, operate, and scale agentic AI with confidence.

> ...As we move toward NVIDIA HGX B300, that proven operating model and continuity in execution give us confidence to focus on building more capable AI code generation systems, rather than worrying about infrastructure risk.
>
> **Aman Sanger**
> *Co-founder, Cursor*

Specification	NVIDIA HGX B300	NVIDIA HGX H200	Impact
GPU memory	270 GB HBM3e	141 GB HBM3e	1.9x memory per GPU. Allows training of ~130B+ parameter models on a single node without model parallelism.
GPU power draw	1,100W	700W	Increased power budget enables higher performance and fuels larger memory subsystem.
Interconnect	NVIDIA Quantum-X800 InfiniBand	NVIDIA Quantum-2 InfiniBand	Improves node-to-node bandwidth and eliminates tail latency, reducing communication overhead for large, multi-node training runs.
Target workload	Frontier model training and inference, agentic systems and reasoning, video generation.	Gen AI and HPC workloads.	Services elite workloads that are constrained by memory, compute, or interconnect speed.

Purpose-built for partnership

CoreWeave delivers AI-native infrastructure co-designed and optimized with NVIDIA to support agentic AI at scale, backed by direct-to-expert support and trusted by leading AI pioneers.

Accelerated breakthroughs

NVIDIA HGX B300 on CoreWeave Cloud lets you accelerate from evaluation to production by giving you early, production-grade signals and the operational visibility to act on them. With Coreweave ARENA, W&B Weave, and Kubernetes-native orchestration via CoreWeave Kubernetes Service and SUNK, you can reduce iteration cycles, surface issues faster, and move from run to results with predictable speed at scale.

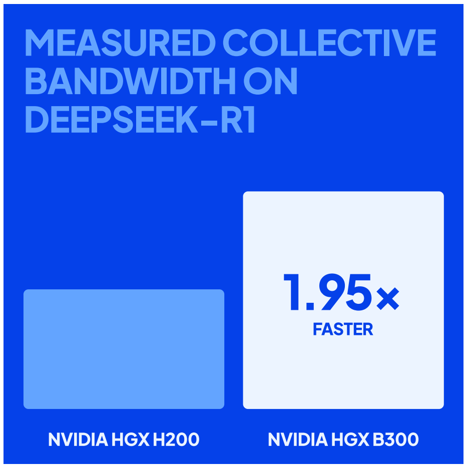
Engineered for performance and resilience

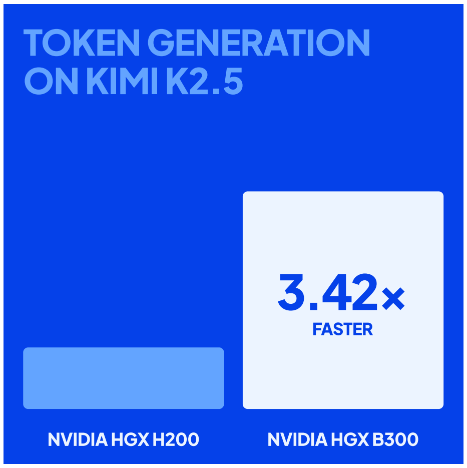
CoreWeave delivers sustained performance by pairing the expanded accelerator memory of NVIDIA HGX B300's with our non-blocking NVIDIA Quantum-X800 InfiniBand networking. This guarantees full bidirectional bandwidth with no oversubscribed links, meaning multi-node training runs experience zero bottlenecks in GPU-to-GPU communication. The result: 2x the node-to-node bandwidth across multi-thousand-GPU clusters.

Every NVIDIA HGX B300 cluster CoreWeave operates is liquid-cooled to prevent damage and thermal throttling, which causes GPUs to slow down. This enables larger systems to run efficiently and scale predictably across nodes. Large-scale systems that run cool operate more efficiently, so you can scale performance predictably and control your costs.

NVIDIA HGX B300 performance on CoreWeave

We benchmarked NVIDIA HGX B300 inference performance against NVIDIA HGX H200 and NVIDIA HGX B200 instances, all running on CoreWeave Cloud. We tested performance using DeepSeek-R1, one of the largest open-source Mixture of Experts models on the market, and Kimi K2.5, an open-source native multimodal agentic model.





DeepSeek-R1

We used native NVFP4 quantization on NVIDIA HGX B300 compared to FP8 block scaling on NVIDIA HGX H200, representing a "best-vs-best" configuration for each architecture.

We also measured distributed communication performance, testing NVIDIA HGX B300 instances (with NVIDIA Quantum-X800 XDR InfiniBand) against NVIDIA HGX H200 instances (with NVIDIA Quantum-2 NDR InfiniBand) to see how they handled heavy collective operations, such as NCCL all-to-all.

NVIDIA HGX B300 Results*

- **2.61x faster token generation**, for more reasoning steps, tool calls, and agent workflows at speed
- **4.93x faster end-to-end request latency**, so reasoning steps complete faster and agents can iterate more quickly
- **1.95x measured collective bandwidth**, for more performant agentic workloads

Kimi K2.5

We drove the system at maximum load while increasing the number of concurrent users to observe how performance changed as demand grew. The tests measured total token throughput and latency metrics, including P99 time to first token and time per output token, to understand how quickly responses started and how fast tokens were generated in the slowest cases users might see.

We recorded these metrics across multiple concurrency levels to evaluate peak performance, scaling under load, and the response speed consistency during inference.

NVIDIA HGX B300 Results

- **3.42x faster token generation,** for more reasoning steps, tool calls, and agent workflows at speed (vs. NVIDIA HGX H200)
- **1.41x improvement in concurrency** for less waiting under heavy load (vs. NVIDIA HGX B200)

Benchmarks

4.93x
faster end-to-end request latency

Source: Using DeepSeek-R1, vs. NVIDIA HGX H200

1.41x
higher concurrency

Source: Using Kimi K2.5, vs. NVIDIA HGX B200

**vs. NVIDIA HGX H200*

Scale agentic AI with CoreWeave Cloud

CoreWeave unlocks NVIDIA HGX B300 performance at scale for agentic AI. CoreWeave Cloud is the Essential Cloud for AI, purpose-built to power your most complex AI workloads. Accelerate every stage of development. Connect with us to reserve NVIDIA HGX B300 today.

Boost efficiency

96% Goodput

High effective compute driven by healthy fleets and fast recovery

Faster response speed

20% Higher MFU

Compared to public foundation model training benchmarks

Scaling under load

10× More stable

3.66 mean time to failure vs. an industry baseline of 0.33 days

Faster speed to market

10× Faster spin-up

Faster inference spin-up times so you can bring your solutions to market first

Engineered for mastery

Photo courtesy of Switch © 2026. All rights reserved.

Engineered for mastery

Explore all of the benefits CoreWeave's purpose-built AI cloud can bring to your most critical workloads.

Contact us

NVIDIA HGX B300 on CoreWeave

A deeper dive into NVIDIA HGX B300 on CoreWeave Cloud.

Read more →

What changes for agentic AI?

Watch our product briefing webinar about NVIDIA HGX B300.

Register →

CoreWeave Mission Control™

Reliability, transparency and insight for large-scale AI workloads.

Explore more →