

Back to Research

NVIDIA B200 vs B300: Why use one instead of the other?

March 10, 2026

In 2026 the AI compute market is no longer a simple generational upgrade cycle. With NVIDIA shipping two distinct Blackwell GPUs into overlapping demand, operators face a new kind of decision: not just which generation to deploy, but which specialization to optimize for. The B200 and B300 (Blackwell Ultra) share the same transistor count, the same manufacturing process, and much of the same silicon. But they are not the same chip. Their differences reflect a deliberate architectural bet by NVIDIA that training and inference workloads are diverging fast enough to justify distinct hardware.

This article compares the B200 and B300 across pricing, architecture, and workload fit, using data from Ornn's GPU rental benchmark and published

performance results, so operators and infrastructure buyers can make informed fleet composition decisions.

GPU Rental Costs: On-Demand Price Landscape

For the purposes of this analysis, we use pricing derived from Ornn's industry-leading GPU rental benchmark, which aggregates observed on-demand pricing across a broad set of cloud providers, GPU rental platforms, and private infrastructure operators.

Based on Ornn's benchmark data, current indicative on-demand rental prices are approximately:

- B200: $4.50 per hour (neocloud range) and $7.00 (hyperscaler)
- B300: $5.00 to $7.25 per hour (neocloud range)

These figures reflect prevailing market-level pricing for single-GPU, on-demand access across cloud and neocloud providers in 2026. Actual realized prices vary by provider, region, contract structure, and workload characteristics.

At the system purchase level, OEM pricing for a B200 SXM module sits at approximately $50,000 per GPU, while B300 modules are reportedly commanding around $53,000 per chip. A full GB200 NVL72 rack (72 GPUs plus 36 Grace CPUs) costs approximately $3.0 to $3.9 million, while a GB300 NVL72 rack runs an estimated $3.7 to $5.0+ million

Nvidia B200

The B200 is the original Blackwell data center GPU. It entered volume production in late 2025 and is now widely deployed across both hyperscaler and neocloud environments.

Key specifications include:

- 9 to 10 PFLOPS dense FP4 tensor performance
- 4.5 to 5 PFLOPS FP8 / FP6
- 2.25 to 2.5 PFLOPS FP16 / BF16
- 37 to 40 TFLOPS FP64
- 180 to 192 GB HBM3e (8-Hi stacks, 8,192-bit bus)
- 7.7 to 8 TB/s memory bandwidth

The B200 delivers strong performance across both training and inference. On inference workloads, a single DGX B200 has demonstrated over 1,000 tokens per second per user on Llama 4 Maverick (400B parameters) and roughly 72,000 tokens per second total throughput per server. On DeepSeek-R1 (671B), it achieves 250+ tokens per second per user. These figures represent approximately 15x the inference performance of DGX H100.

NVIDIA's continuous software optimization has been equally significant. B200 inference throughput improved roughly 5x since launch through TensorRT-LLM and Dynamo improvements alone, pushing effective

cost as low as $0.02 per million tokens on representative workloads.

Nvidia B300 (Blackwell Ultra)

The B300 is the Blackwell Ultra refresh, announced at GTC 2025 and shipping into production deployments in the second half of 2025. It makes targeted architectural trade-offs that reshape its workload profile relative to B200.

Key specifications include:

- 14 to 15 PFLOPS dense FP4 tensor performance (+50 to 55% vs B200)
- 4.5 to 5 PFLOPS FP8 / FP6 (same as B200)
- 2.25 to 2.5 PFLOPS FP16 / BF16 (same as B200)
- 1.2 to 1.3 TFLOPS FP64 (97% reduction vs B200)
- 270 to 288 GB HBM3e (12-Hi stacks, same 8,192-bit bus)
- 7.7 to 8 TB/s memory bandwidth (same as B200)

The B300's FP4 uplift comes from three sources: all 160 SMs enabled versus 148 harvested on B200, higher clocks under the increased power envelope, and tensor core optimizations for NVIDIA's proprietary NVFP4 format.

The most telling design choice is the 97% reduction in FP64 throughput. NVIDIA repurposed die area from double-precision units to boost FP4 compute and attention-layer throughput. INT8 performance also dropped roughly 93%. This is a statement about where NVIDIA sees AI workloads going: FP4

and FP8 will dominate, and double-precision is increasingly irrelevant for AI training and inference.

The 50% increase in HBM capacity uses taller 12-Hi memory stacks in the same physical bus configuration. At the rack level, a GB300 NVL72 delivers approximately 1.1 exaflops of FP4 and roughly 20 TB of GPU memory, compared to 0.72 exaflops and 13.5 TB for the GB200 NVL72.

Training Large Models

The B300 delivers a clear training advantage over the B200 due to higher memory capacity, faster networking, and support for new low-precision training modes. Its 50% larger HBM pool (2.3 TB in an 8-GPU system vs. 1.44 TB on B200) allows large models such as 70B-parameter architectures to fit more comfortably in memory, reducing the need for aggressive memory optimization and enabling larger batch sizes. ConnectX-8 also doubles inter-node bandwidth to 1.6 Tbps, lowering gradient synchronization overhead during distributed training. Combined with NVFP4 training support, these improvements produced roughly a 12–13% training speedup over B200 in MLPerf Training v5.1 benchmarks.

Inference and Throughput-Intensive Tasks

For inference workloads, the advantage largely disappears. Token generation in LLMs is primarily memory-bandwidth bound, and both B200 and B300 deliver roughly the same ~8 TB/s bandwidth. As a result, per-token throughput is similar across most serving scenarios. B200's 192 GB per GPU already supports models up to ~70B parameters in FP16 or ~130B in FP8 without sharding, covering the majority of production deployments. Because B200 GPUs are typically 15-40% cheaper per GPU-hour, they generally offer the better cost-performance tradeoff for inference unless model size exceeds the B200 memory envelope.

Cost-Performance Considerations in Practice

Comparing raw rental rates alone does not capture the full picture of cost effectiveness. When normalized for throughput, the relative economics shift depending on workload type.

For training workloads, B300's 12.6% speed advantage comes at a 50 to 100% premium in hourly rental cost. On a pure cost-per-training-job basis, B300 is less cost-efficient than B200 unless the workload specifically requires the extra memory capacity or benefits from FP4 precision. The B300 premium is justified when memory constraints force the alternative of more complex parallelism strategies on B200, or when the workload can exploit NVFP4 training recipes that are not available at B200's lower FP4 throughput.

For inference workloads, the calculation is more straightforward.Unless the model exceeds B200's memory capacity, the lower-cost chip wins on unit economics.

Effective cost comparisons should always account for throughput per GPU, memory and bandwidth bottlenecks, job wall-clock time, parallelism overhead, and aggregate resource utilization. The cheapest GPU per hour is not always the cheapest GPU per job.

Hyperscaler Procurement and the Emerging Fleet Split

The B200 versus B300 purchasing patterns across hyperscalers illuminate how the market is stratifying.

SemiAnalysis reported in December 2025 that all major hyperscalers had decided to move forward with GB300 for new orders, as initial GB200 production delays pushed procurement timelines into windows where the superior chip became available. By Q4 of NVIDIA's fiscal year 2026, GB300 accounted for roughly two-thirds of all Blackwell revenue.

But the underlying strategies differ substantially by buyer.

Microsoft moved fastest on GB300. Azure unveiled the world's first GB300

NVL72 supercomputing cluster in October 2025, a 64-rack deployment containing over 4,600 Blackwell Ultra GPUs purpose-built for OpenAI's workloads. Microsoft is scaling to hundreds of thousands of Blackwell Ultra GPUs globally.

Google takes a fundamentally different approach. Google's internal AI workloads run predominantly on custom TPUs, with TPU v7 Ironwood operating in pods of 9,216 chips. Google was the first cloud provider to offer both B200 and GB200 NVL72 instances to GCP customers, but these serve external demand for CUDA compatibility rather than Google's own fleet. Google's GPU procurement is primarily about making its cloud competitive for customers who need NVIDIA hardware, not about powering its own models.

Meta hedged with a dual-vendor strategy. In a single week in February 2026, Meta announced a multiyear deal with NVIDIA for millions of Blackwell and Rubin GPUs and separately committed $60 to $100 billion to AMD for custom MI450 GPUs across 6 gigawatts of capacity. The dual-vendor approach gives Meta leverage on NVIDIA pricing while ensuring supply diversity.

Amazon/AWS runs dual-track procurement: NVIDIA GPUs for customer-facing cloud instances (P6-B200 and P6-B300 both live) alongside custom Trainium chips for internal workloads.

xAI is scaling the fastest by absolute GPU count, with plans to integrate

550,000 GB200 and GB300 GPUs into its Colossus 2 cluster targeting 1 million total GPUs.

The pattern across all of these buyers is consistent: B300 for new training clusters, B200 for inference fleets and cost-sensitive deployments. As inference grows to represent an estimated two-thirds of all AI compute demand in 2026 (up from one-third in 2023), the B200's economics become increasingly compelling for the fastest-growing segment of the market.

Conclusion

The B200 and B300 each offer distinct trade-offs that reflect a broader structural shift in AI compute:

- B200 delivers strong inference economics at lower hourly rates, with 192 GB memory sufficient for the majority of production serving workloads. Its price is compressing as B300 supply ramps, making it increasingly attractive for cost-sensitive deployments.
- B300 leads on training throughput with 50% more memory, 50% more FP4 compute, and doubled networking bandwidth. It commands a justified premium for frontier training and workloads that require 288 GB per GPU.

The market is bifurcating. Training fleets are consolidating around B300 and GB300 NVL72 racks. Inference fleets

are finding that B200 delivers comparable per-token economics at lower capital and operating cost. With NVIDIA's Rubin platform arriving in H2 2026 promising another order-of-magnitude reduction in inference cost, the window to extract maximum ROI from Blackwell deployments is compressed. Matching GPU type to workload has never mattered more.

Abundance needs an exchange.

Find Compute

See the data



X
LinkedIn
Substack

Careers
About
Research
Docs
Trust Center
Regulatory

Privacy Policy
Terms of Use

Brand + Website by Haptic

©2026 Ornn AI, Inc. All rights reserved.

