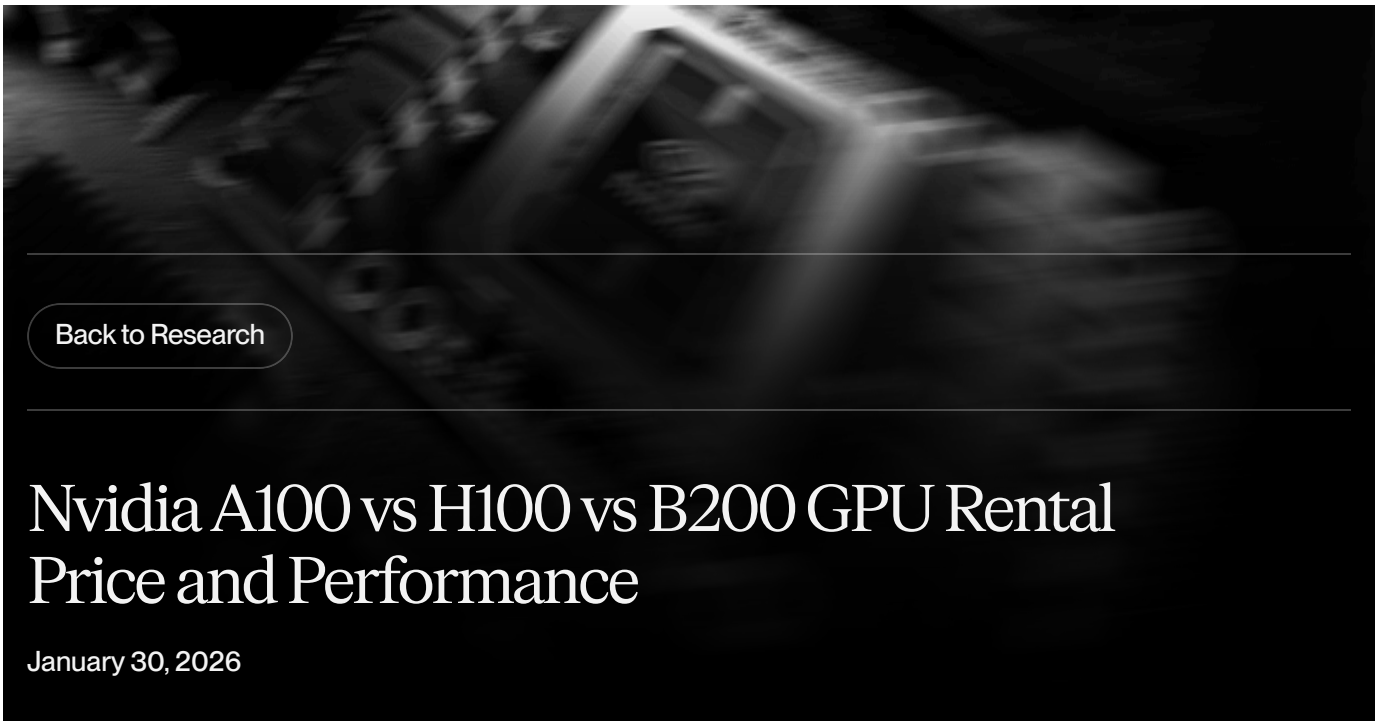

Back to Research

Nvidia A100 vs H100 vs B200 GPU Rental Price and Performance

January 30, 2026

Evaluating Cost-Effectiveness and Workload Suitability in Today's AI Compute Market

In 2025 the landscape of AI compute continues to evolve rapidly as data center operators and cloud providers wrestle with ai compute demand, ai compute spending, and shifting hardware economics. Three GPUs dominate discussions for high-performance workloads: the NVIDIA A100, the NVIDIA H100, and the NVIDIA Blackwell-based B200.

Each of these accelerators presents a distinct balance of price, performance, and workload fit when you consider gpu rental prices and real-world performance characteristics. In this

article we compare on-demand rental costs alongside key architectural differences and performance profiles so you can make informed deployment decisions for training, inference, or mixed AI workloads.

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:

- A100: $0.80 per hour
- H100: $1.75 per hour
- B200: $3.05 per hour

These figures reflect prevailing market-level pricing for single-GPU, on-demand access in 2026. Actual realized prices may vary by provider, region, contract structure, and workload characteristics. Alternative pricing models, including spot usage, volume discounts, or committed capacity agreements, can materially affect effective hourly costs, but the benchmark provides a consistent

reference point for comparing relative price levels across GPU generations.

Architectural Generations and Performance Differentials

NVIDIA A100

The NVIDIA A100, based on the Ampere architecture, remains a cost-effective option for many AI workloads, particularly where software maturity and memory capacity are priorities. In commonly deployed configurations, the A100 delivers approximately:

- 312 TFLOPS FP16 tensor performance
- 19.5 TFLOPS FP32
- 1.6 TB/s memory bandwidth with HBM2e
- 40 GB or 80 GB HBM capacity, depending on SKU

For transformer training and general deep learning workloads, these specifications translate into solid baseline performance, especially for models that fit comfortably within memory and do not fully benefit from newer low-precision formats. While the A100 lags newer generations in throughput, it remains widely used due to its lower GPU rental price and broad framework support.

NVIDIA H100

The NVIDIA H100, built on the Hopper architecture, represents a substantial performance increase over A100, particularly for transformer-heavy workloads. Published and observed performance characteristics include:

- ~1,000 TFLOPS FP16 tensor performance with sparsity
- ~500 TFLOPS FP16 without sparsity
- 60 TFLOPS FP32
- Up to 3.35 TB/s memory bandwidth with HBM3
- 80 GB HBM capacity

In practice, these improvements often translate into 2x to 3x higher training throughput compared to A100 on large language models, depending on batch size and model architecture. H100 also introduces native FP8 support, which significantly improves efficiency for modern training pipelines and inference workloads.

As a result, H100 frequently delivers better performance per dollar than A100, even at a higher hourly rental price, particularly for large models and sustained training runs.

NVIDIA B200

The NVIDIA B200, part of the Blackwell generation, delivers the highest performance of the three GPUs and is designed explicitly for next-generation AI workloads. Indicative performance figures for B200 include:

- Up to ~2,000 TFLOPS FP16 equivalent tensor performance
- Native FP4 and FP8 acceleration

- Significantly higher effective throughput for transformer workloads
- Up to 8 TB/s memory bandwidth with next-generation HBM
- 192 GB HBM capacity per GPU

Across published benchmarks and early production deployments, B200 frequently demonstrates 1.8x to 2.5x performance improvements over H100 for large-scale training and high-throughput inference. For some workloads, particularly memory-bound or communication-heavy models, B200 can reduce wall-clock training time by roughly half relative to H100.

While B200 carries a higher on-demand GPU rental price, its substantially higher throughput can result in lower total cost per training job or per inference unit for sufficiently large and performance-sensitive workloads.

Performance Profiles by Workload Types

When evaluating GPUs it is important to match hardware to workload. Across real-world tests and industry comparisons:

Training Large Models

For frontier training tasks, especially where model size and memory bandwidth are dominant constraints, the B200 generally outpaces the H100 and A100. In head-to-head benchmarks for large-scale training, B200 has exhibited

roughly 2x the speed of previous generation GPUs and often shortens overall training wall-clock time.

H100 delivers strong performance for a wide range of training sizes and remains a reliable middle ground with powerful tensor core acceleration compared to A100. It produces significant speedups relative to A100, particularly for FP8 workloads, and is a common choice for teams requiring stable performance across varied model sizes.

A100 continues to serve as a cost-effective platform for models that fit its memory footprint and do not benefit as much from the latest architectural features. Its lower price point and mature ecosystem make it viable for many established workflows.

Inference and Throughput-Intensive Tasks

For inference, performance gaps widen once lower-precision formats and batching are applied. On transformer models such as LLaMA-2 70B, operators typically observe H100 delivering roughly 2x the tokens per second of A100 when using FP8, while B200 can deliver an additional ~1.5x to 2x improvement over H100 when FP4 or FP8 is enabled and batch sizes are large.

In practical terms, this means a single B200 can often replace multiple A100s for high-throughput inference workloads, in exchange a higher hourly rental price. H100 provides a balanced middle ground for mixed training and inference environments, while A100 remains viable

for steady, lower-throughput inference where cost per hour is the primary constraint.

Cost-Performance Considerations in Practice

Comparing raw rental price rates alone does not capture the full picture of cost effectiveness. When normalized for performance, the relative effective compute pricing can shift:

- At $0.80 per hour, A100 offers the lowest entry cost but may require more GPU hours to complete equivalent workloads relative to H100 or B200.
- H100 at $1.75 per hour sits in the middle of the cost and performance spectrum, often delivering stronger performance per dollar for many training benchmarks compared to A100.
- B200 at $3.05 per hour has the highest on-demand hourly cost, but due to its greater throughput and memory bandwidth it may complete jobs in substantially fewer hours, reducing total budget required for large or complex AI jobs.

For example, if B200 training requires half the time of an H100 for a given workload, the effective cost per training job can be competitive despite higher hourly rental rates.

Effective cost comparisons should always consider throughput per GPU, memory and bandwidth bottlenecks, job wall-clock time, and aggregate resource utilization as essential factors in calculating full economic costs.

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.

