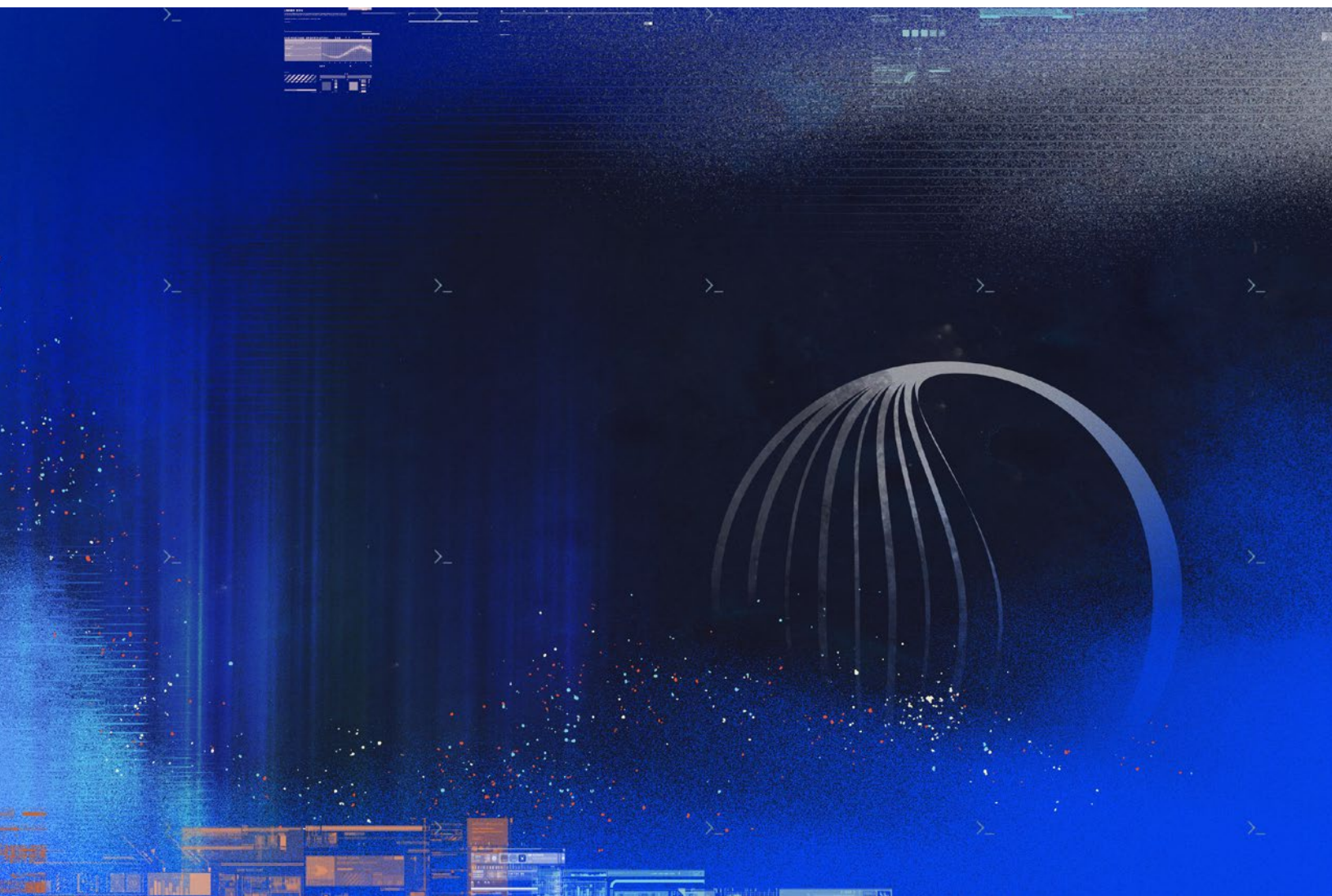




CoreWeave Delivers Higher Throughput and Lower Cost per Token on NVIDIA Blackwell: A Three-Year TCO and Tokenomics Analysis

AUTHOR

Mitch Lewis

Performance Analyst | Signal65

IN PARTNERSHIP WITH



SEPTEMBER 2026

Executive Summary

Enterprise AI programs increasingly run on rented infrastructure. Rather than absorbing the capital cost, lead times, and ongoing maintenance of standing up hardware on-premises, organizations can reserve dedicated GPU capacity on demand. They can also skip infrastructure management altogether and call hosted models through serverless APIs priced per token. Both paths are widely available across cloud providers. What differs substantially is what they cost.

Pricing for AI cloud services varies significantly between traditional hyperscaler clouds and specialized, AI-optimized clouds such as CoreWeave, and at production scale those differences compound into material sums. This study quantifies them. It models a 5,000-GPU NVIDIA Blackwell deployment over a three-year term against three anonymized hyperscaler clouds, using public on-demand pricing, MLPerf Inference v6.0 throughput results, and published serverless token rates. For inference workloads it accounts for performance as well as price, since how efficiently a provider runs a given model shapes the real cost per token as much as the published rate does.

The results show CoreWeave consistently delivers the lowest total cost of ownership for NVIDIA Blackwell GPU deployments, doing so in six of the seven configurations compared, and that advantage widens once inference performance is taken into account. CoreWeave's cost advantage additionally extends past GPU rental to storage, Kubernetes, and other supporting services, and again to Serverless Inference.



CoreWeave delivers lower total cost than hyperscaler clouds for 3-year AI deployments on NVIDIA Blackwell GPUs:

- Up to 41% lower cost on NVIDIA HGX B200
- Up to 65% lower cost on NVIDIA GB200 NVL72
- Up to 52% lower cost on NVIDIA HGX B300



CoreWeave also delivers lower cost per token for inference, driven by a combination of lower list pricing and industry-leading performance:

- CoreWeave achieves between 4% and 124% greater throughput than comparable references when running DeepSeek-R1 on NVIDIA GB200 NVL72.
- CoreWeave's performance and pricing advantages compound to increase per-token cost advantages. Compared to the most competitively priced hyperscaler competitors, CoreWeave's per-token cost advantage widens from 22% based on infrastructure pricing alone to 65% once performance is factored in.



CoreWeave leads on Serverless Inference pricing.

- Across 17 models with a direct hyperscaler comparison, CoreWeave Serverless Inference is lower priced on 11 and identically priced on 4.
- CoreWeave provides 26% lower median input token costs and 15% lower median output token costs across 17 model comparisons.

Introduction

As enterprises race to deploy AI models and infrastructure at scale, cloud platforms have become a core avenue for accessing the GPU capacity those workloads require, offering flexible, on-demand access to the latest accelerators. NVIDIA's Blackwell and Blackwell Ultra generation — including the NVIDIA HGX B200, NVIDIA HGX B300, and NVIDIA GB200 NVL72 systems — represent the current generation of AI infrastructure in large-scale production deployment. Rather than committing the capital and enduring the long lead times and ongoing maintenance required to deploy this latest generation of hardware on-premises, organizations can reserve dedicated Blackwell GPU instances on demand and run their own training or inference workloads on them.

For organizations that don't need or want to manage dedicated infrastructure, cloud providers also offer serverless inference: hosted access to popular AI models through a simple API, billed per token rather than per GPU-hour. This removes the need to provision, configure, or scale GPU infrastructure at all, trading some flexibility for a faster path to production and a pricing model that scales directly with usage.

Pricing for both approaches varies significantly across providers, and those differences can materially affect the economics of running AI workloads at scale. Traditional hyperscaler clouds and specialized, AI-optimized clouds such as CoreWeave often take meaningfully different approaches to pricing these services. For organizations planning cloud-based AI deployments, understanding these cost differences has become as important as evaluating the hardware itself.

For inference workloads in particular, the price of either option only tells part of the story. How efficiently a provider's infrastructure actually runs a given model plays a direct role in the real cost per token, alongside the sticker price itself. This study examines cloud costs across both paths, accounting for performance as well as price.

Methodology

This analysis evaluates cloud costs for CoreWeave and three anonymized hyperscaler cloud competitors, referenced as Hyperscaler Cloud A, B, and C. The analysis evaluates cloud costs across three distinct levels:

Total cost for a cloud AI deployment over 3 years

The total cost of ownership sets a baseline cost analysis for multi-year AI cloud deployments, inclusive of GPUs and additional supporting infrastructure. This analysis uses public, on-demand pricing for CoreWeave and three hyperscaler cloud competitors. All pricing data was collected as of August 2026. The cost model includes the following assumptions:

- 3-year duration
- 5,000 GPUs
- 2 PB of object storage
 - Storage split among tiers: 15% hot, 35% warm, 50% cold
 - Monthly egress fees: Data is replicated across 3 regions, with an assumed new data generation rate of 20%
- Additional costs:
 - Direct connect networking
 - Kubernetes service and support
 - Observability: Audit log and metric forwarding costs
 - Support: Calculated as a percent of total spend

Signal65 Comment: While this analysis models monthly egress fees, it does not account for all additional storage-related fees common among traditional hyperscaler clouds. These can include per-operation storage charges and early-deletion penalties on archive-tier storage. CoreWeave's storage pricing does not include comparable charges. In practice, these fees can increase the total cost of storage on hyperscaler clouds, further widening CoreWeave's cost advantage.

Inference Cost per Million Tokens

Inference cost builds on the same baseline assumptions as the three-year cloud cost, but incorporates inference performance data to translate a total contract cost into a per-million token price. This provides a realistic view into the cost efficiency to run inference workloads on each cloud.

- Each cloud is priced at a per-GPU-hour rate, derived from the fully loaded cost found in the three-year cost analysis.
- Performance data is sourced from MLPerf Inference v6.0 Closed Division results, running DeepSeek-R1 under the Server scenario. MLPerf is used rather than theoretical hardware specifications because it reflects audited, real-workload throughput rather than peak theoretical performance, and submitted results can showcase cloud-specific optimizations. Where possible, each vendor's own MLPerf performance data is used. However, not all cloud vendors included in this analysis submitted MLPerf results. In cases where vendor-specific data is missing, NVIDIA's own reference submissions are used as a substitution.
- Cost per token assumes 80% GPU utilization, applied uniformly across all providers.

Signal65 Comment: Real-world GPU utilization varies widely and is often lower than this figure, particularly in less-optimized environments. The 80% assumption used here reflects a well-orchestrated, dedicated production deployment, and is applied uniformly across all providers

Token costs for serverless model APIs

The serverless comparison provides an alternative analysis of inference costs, using cloud API token costs instead of deriving token costs from cloud instances. Serverless compute provides a convenient, flexible option for organizations that do not want to maintain their own GPU instances. Because serverless options are priced per token, the same model's cost can be directly compared across providers. Due to varying model support across cloud providers, comparisons in this analysis are limited to models with confirmed pricing on multiple platforms. Where multiple comparable models exist, median (rather than average) price differences are reported, since a small number of outlier models can otherwise distort a simple average.

Limitations

While this analysis uses publicly available pricing and performance data to create a fair and transparent analysis, not all pricing and performance data is available for all configurations. To maintain the breadth of this analysis across vendors, some values have been derived or substituted for the closest available data, and should be noted as limitations of this study.

- Not all cloud pricing is publicly available for every GPU configuration in this study. Where a provider does not publish on-demand pricing for a given GPU, that comparison is either narrowed to the providers with available data, or a price estimate is derived by applying an observed cross-generation price markup — calculated using providers with pricing available on both GPU generations — to that provider's known pricing on a related GPU. Any provider whose markup pattern differs substantially from its peers is treated as a statistical outlier and excluded from that calculation. All such estimates are clearly labeled throughout this report and should not be read as confirmed pricing.
- MLPerf submissions are incomplete across vendors and GPU generations. Where a vendor's own submission exists, it is used regardless of whether it's favorable or unfavorable to that vendor; where it doesn't, NVIDIA's own reference submission is used as a labeled stand-in.
- Serverless model pricing is not available for every model on every platform. Only models with confirmed, directly comparable public pricing on both CoreWeave and at least one hyperscaler are included in the serverless comparison.
- This analysis compares CoreWeave to traditional hyperscaler cloud providers rather than other specialized AI infrastructure ("neocloud") providers. Hyperscalers represent the default infrastructure choice for the majority of enterprise cloud workloads today, making them the most relevant baseline for organizations evaluating a shift toward specialized AI infrastructure.

Results

NVIDIA Blackwell Costs 22–65% Less on CoreWeave

A 5,000-GPU NVIDIA Blackwell cluster is a multi-billion-dollar, multi-year commitment, and choosing the right cloud provider matters. This cost analysis models what a production cluster actually requires: GPUs plus storage, networking, observability, and support — not GPU rental alone. This analysis was completed across three distinct systems based on the NVIDIA Blackwell architecture: NVIDIA HGX B200, NVIDIA GB200 NVL72, and NVIDIA HGX B300.

NVIDIA HGX B200

When evaluating NVIDIA HGX B200 deployments, CoreWeave achieves significant advantages against two of the three hyperscaler clouds. Compared to Hyperscaler A, CoreWeave is found to be 41% lower cost over 3 years. Compared to Hyperscaler B, CoreWeave is found to be 25% lower. Hyperscaler C configured with NVIDIA HGX B200s was found to be the one configuration in this study that achieved a lower three-year cost than CoreWeave. This advantage, however, is minimal at 3.7% lower contract cost over the three-year period.

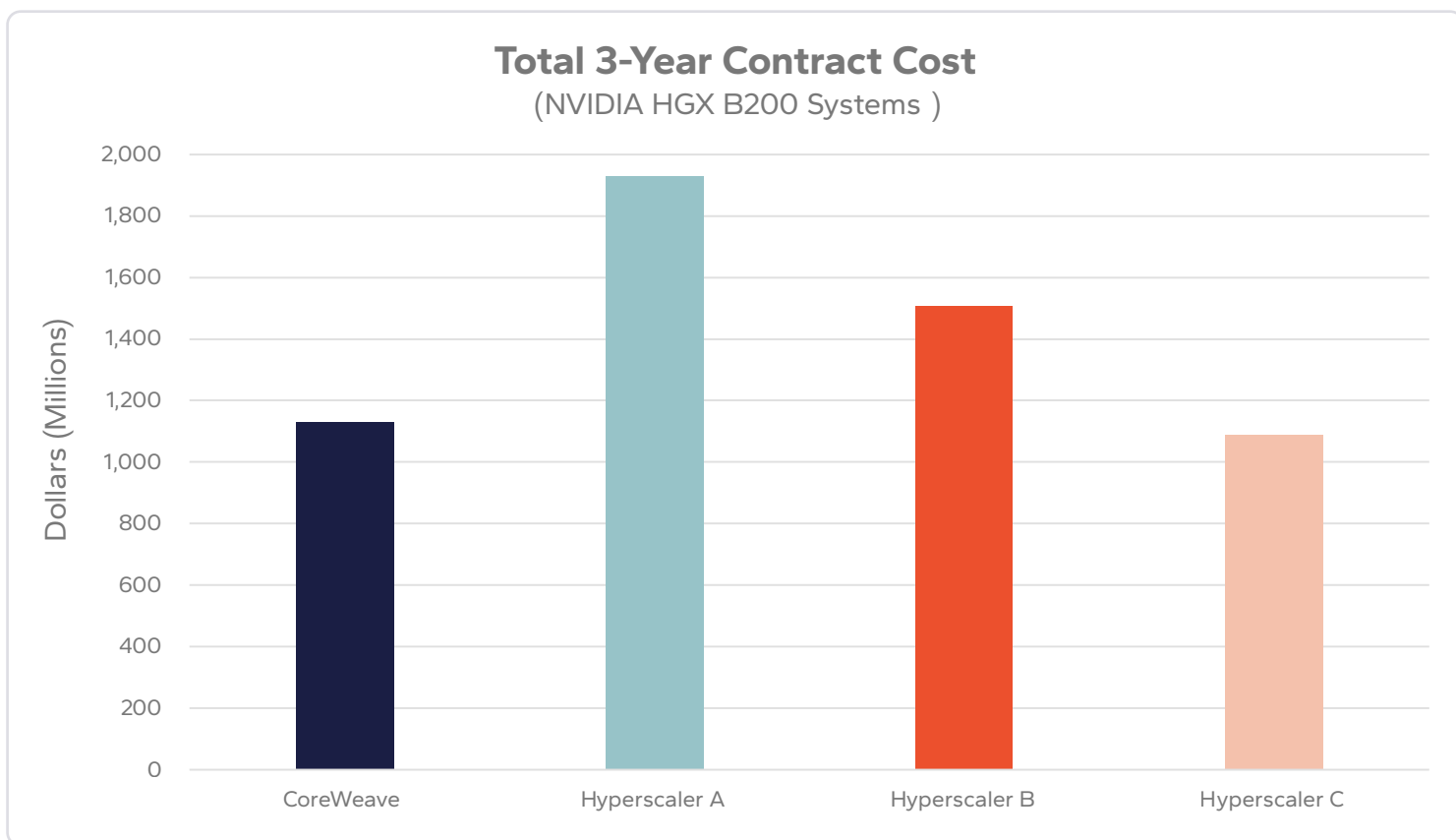


Figure 1: Total 3-Year Cloud Costs (NVIDIA HGX B200 Systems)

NVIDIA GB200 NVL72

When configured with NVIDIA GB200 NVL72 systems, CoreWeave achieves cost advantages compared to all three hyperscaler clouds. While Hyperscaler C achieved a slight cost advantage when configured with NVIDIA HGX B200, it is the most expensive solution when configured with NVIDIA GB200 NVL72, at \$3,912,073,679 over 3 years. Comparatively, an NVIDIA GB200 NVL72 deployment on CoreWeave is found to be 65% cheaper. It is additionally found to be 36% lower cost than Hyperscaler A, and projected to be approximately 22% lower cost than Hyperscaler B, using an estimated hourly cost for NVIDIA GB200 NVL72.

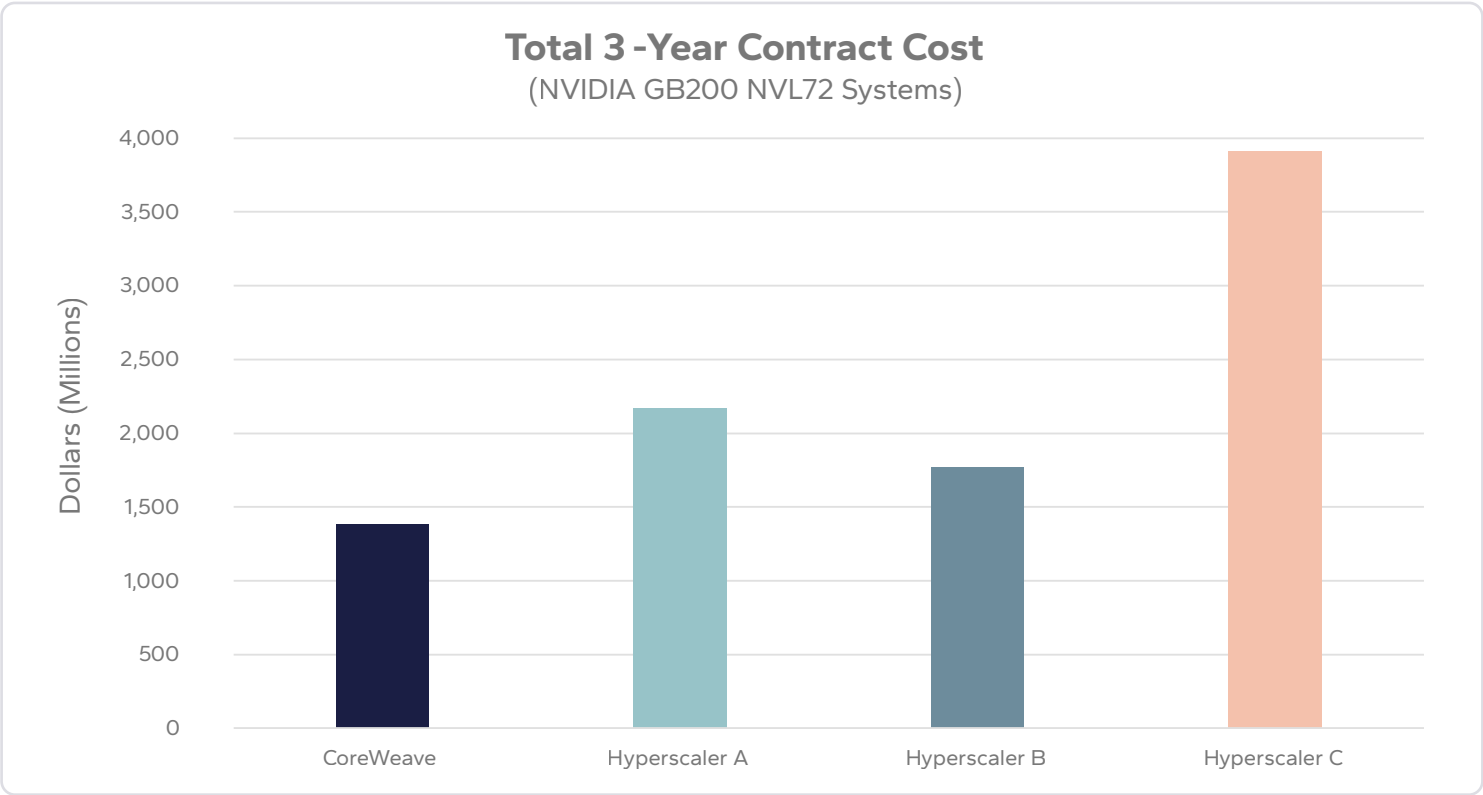


Figure 2: Total 3-Year Cloud Costs (NVIDIA GB200 NVL72 Systems)

NVIDIA HGX B300

Cost comparisons for NVIDIA HGX B300 deployments are limited, as only Hyperscaler A has publicly available pricing data. Despite limited comparisons, CoreWeave again demonstrates a significant overall cost savings in this configuration. While Hyperscaler A was found to have a three-year cost of \$2,410,986,386, CoreWeave’s three-year total cost was calculated to be \$1,165,645,080 – a total savings of 52%.

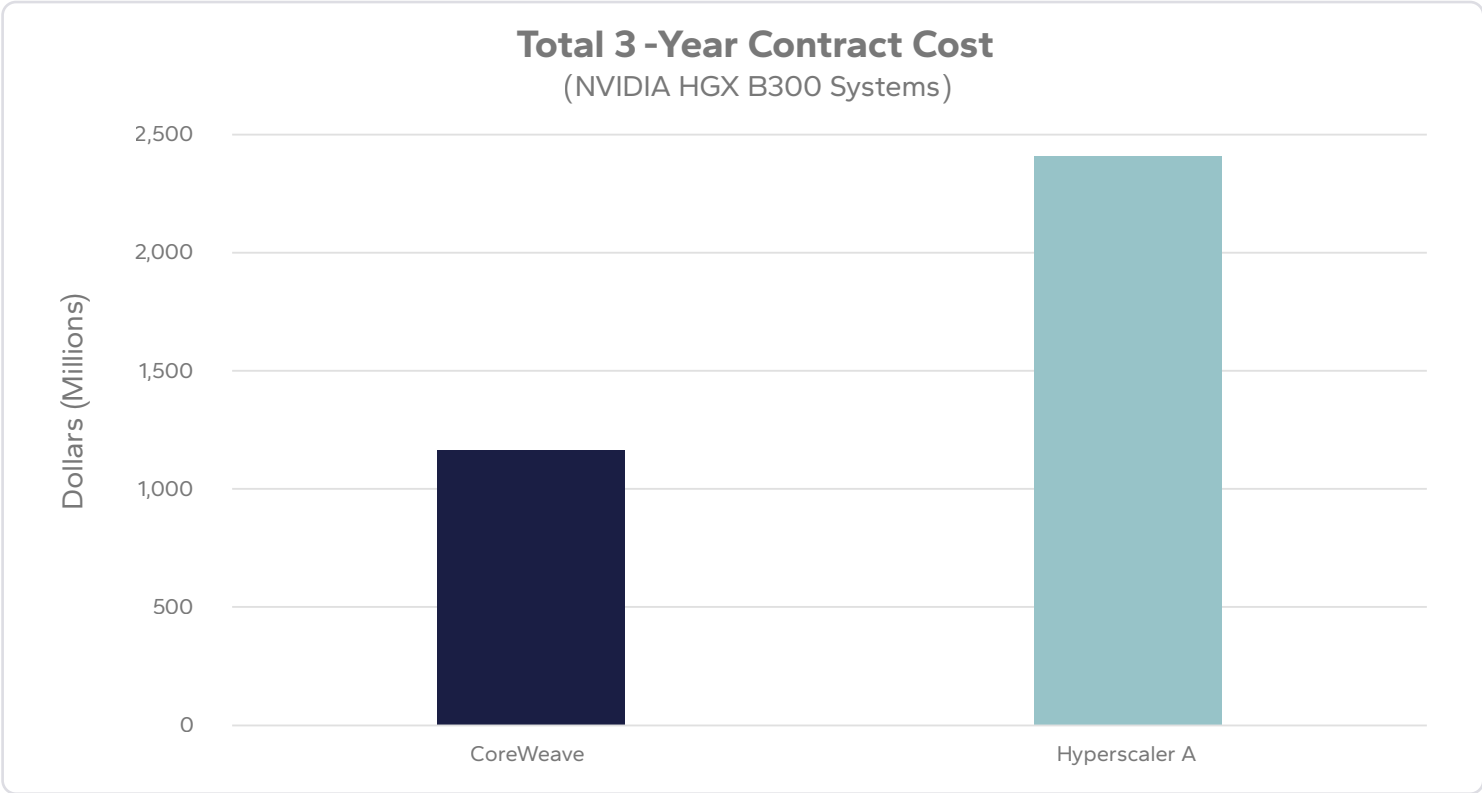


Figure 3: Total 3-Year Cloud Costs (NVIDIA HGX B300 Systems)

What Drives the Cost Gap

GPU Costs Dominate the Contract

Across all three GPU configurations, GPU costs consistently accounted for the overwhelming majority of the total three-year cloud cost, more than 90% for all configurations. While GPUs drove the majority of costs in all configurations, they notably accounted for more than 99% of all costs on CoreWeave, highlighting the low cost of additional infrastructure requirements. For enterprise customers deploying AI, this means that more than 99% of cloud spend is attributed directly to the most critical AI infrastructure.

	CoreWeave	Hyperscaler A	Hyperscaler B	Hyperscaler C
GPU % of Cloud Costs	99.8%	96.9% - 97%	96.9%	90.4% - 90.8%

Figure 4: GPU % of Cloud Costs

It is additionally notable that despite the high percentage of cost attributed to GPUs, CoreWeave consistently achieved overall cost advantages in all but one comparison. Because GPU cost represents 90%+ of total spend for every provider, CoreWeave's overall cost advantage cannot be explained by lower ancillary fees alone; it reflects genuinely more competitive GPU pricing itself.

Storage Costs Add Up Over Time

Although GPU costs dominate the overall cost of the 3-year deployment, the remaining costs still add up over time. When evaluating storage costs, CoreWeave again achieves significant savings compared to all three hyperscaler clouds. In total, CoreWeave achieved savings of between 28% and 55% on data storage. Against Hyperscaler C, the most expensive provider in the study, that is a 55% lower storage bill — \$2,386,800 over the three-year term. While storage represents a small share of total deployment cost, these savings compound over a three-year contract and reflect the same pattern found across every layer of CoreWeave's infrastructure stack.

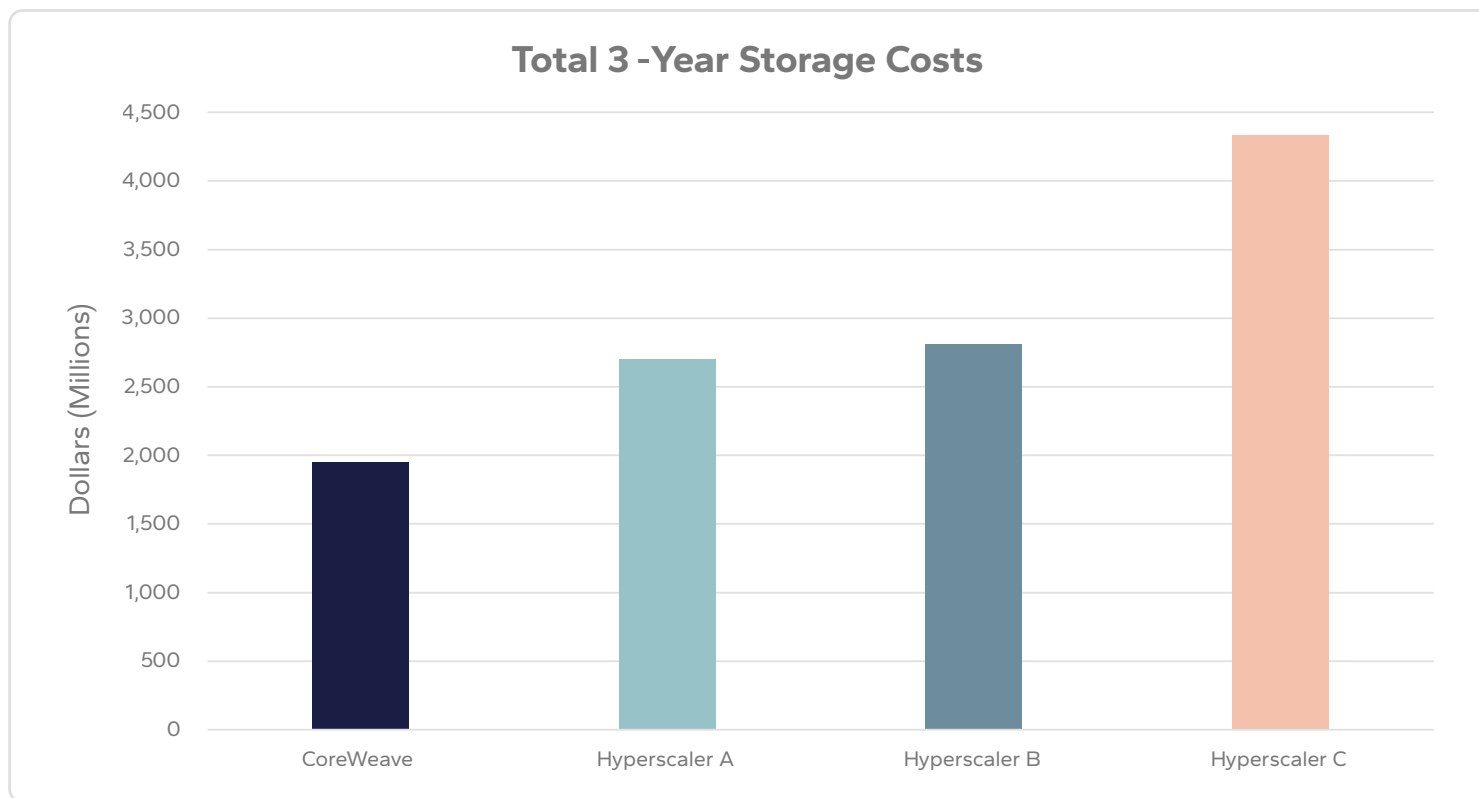


Figure 5: Total 3-Year Storage Costs

CoreWeave Eliminates Extra Fees

CoreWeave further extends its cost advantage by eliminating extra fees that are attached to supporting services and infrastructure. For data storage, CoreWeave does not charge egress fees, which ranged from \$576,000 to \$2,505,600 in the hyperscaler environments over the three-year period. It additionally does not charge support fees or observability fees. Removing observability costs was found to save between \$181,183 and \$828,000. The removal of support fees was even more impactful, with total hyperscaler support costs ranging from \$43,900,804 up to \$355,626,792. Because support costs are calculated as a percentage of total contract spend, they scale directly with a provider's overall pricing, meaning higher GPU and infrastructure costs compound into disproportionately larger support fees. For CoreWeave customers, the absence of these fees translates into direct cost savings.

Performance Widens the Gap

The comparison so far is based on what a cluster costs to run. Inference capacity, though, is bought for the tokens it produces, and throughput differs between providers even on identical NVIDIA GB200 NVL72 hardware. That gap reflects the full infrastructure stack surrounding the GPU, including software, networking, and system-level tuning. Combining the three-year cost with audited MLPerf throughput gives a cost per million tokens, which accounts for both price and performance.

This analysis uses the baseline 3-year contract costs, alongside MLPerf performance data to provide a realistic view into the token cost efficiency that can be achieved by each cloud provider. This evaluation uses MLPerf Inference results for NVIDIA GB200 NVL72 deployments running DeepSeek-R1 under the Server scenario. Both CoreWeave and Hyperscaler B submitted results for this scenario. While the other two hyperscaler clouds have not submitted MLPerf performance results for this scenario, NVIDIA's own reference is used as a substitute.

Notably, CoreWeave achieved the highest per-GPU token throughput of all three MLPerf submissions, with 4,867 tokens per second. NVIDIA's own results show highly competitive performance, at 4,668 tokens per second. While this is not a huge performance advantage, CoreWeave's ability to produce 4% more tokens per second demonstrates that CoreWeave's infrastructure matches or exceeds NVIDIA's own reference implementation. Hyperscaler B, which also submitted MLPerf results running DeepSeek-R1 on NVIDIA GB200 NVL72, achieved 2,177 tokens per second, 55% less than CoreWeave.

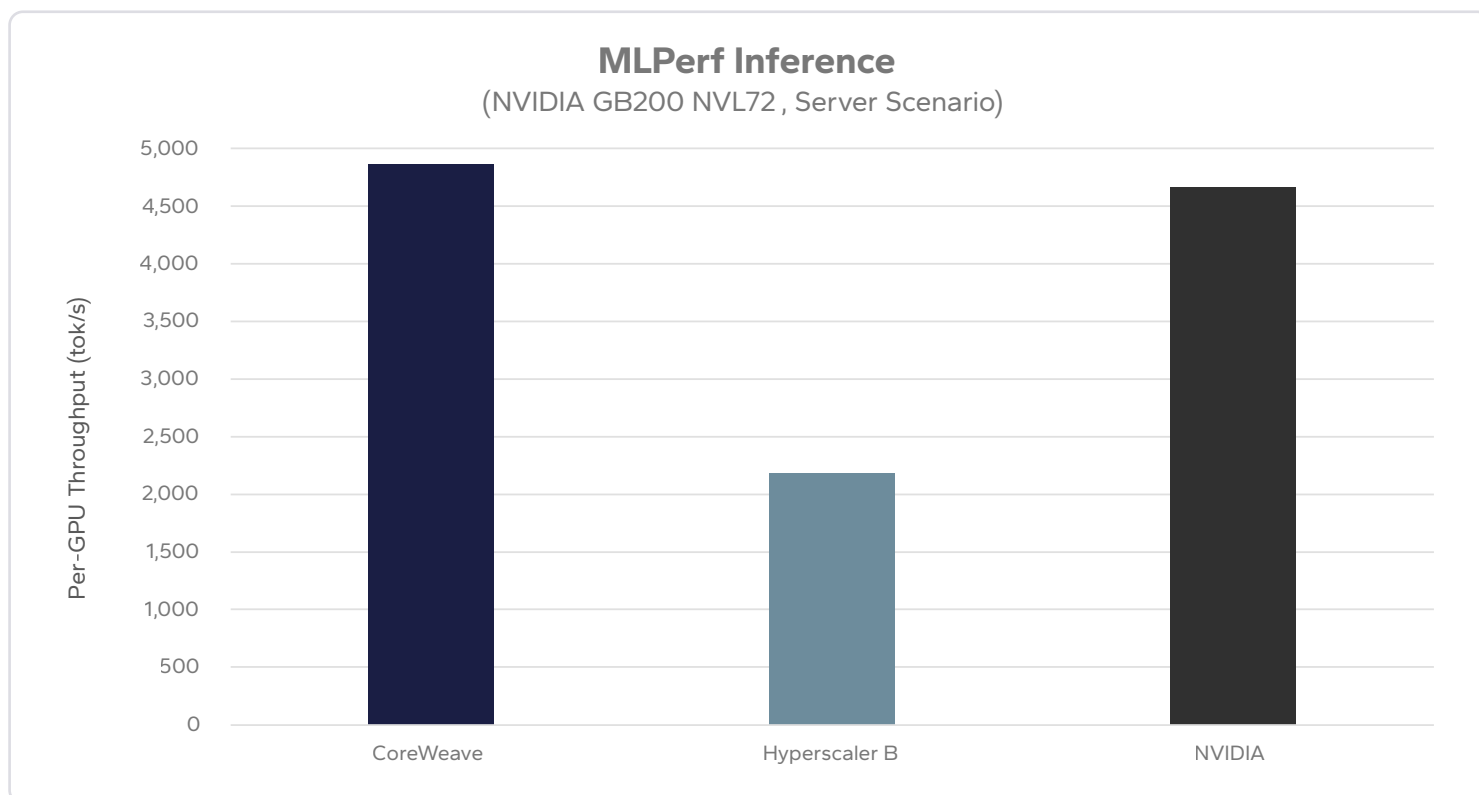


Figure 6: Per-GPU Throughput (tok/s)

In practice, CoreWeave’s performance advantages can further amplify its overall cost advantages, by creating a lower cost per million tokens. Compared to Hyperscaler A and Hyperscaler C, which are modeled with the competitive NVIDIA performance results, CoreWeave’s additional increase in savings is modest. CoreWeave achieved 36% and 65% total 3-year savings compared to Hyperscaler A and Hyperscaler C deployments of NVIDIA GB200 NVL72, while it achieves slightly higher per token savings at 39% and 66%. Although the increase is relatively small, any additional savings at this scale can still be meaningful in terms of raw cost.

Provider	Deployment \$/hr	Throughput Basis	Throughput (tok/s/GPU)	Cost per M Tokens	CoreWeave Savings %
CoreWeave NVIDIA GB200 NVL72	\$10.52	Own submission	4,867.7	\$0.7504	—
Hyperscaler A	\$16.52	NVIDIA reference	4,668.1	\$1.2286	38.93%
Hyperscaler B	\$13.46*	Own submission	2,177.5	\$2.1463	65.04%
Hyperscaler C	\$29.79	NVIDIA reference	4,668.1	\$2.2157	66.15%

Figure 7: Cost per Million Tokens Details

*Hyperscaler B uses estimated NVIDIA GB200 NVL72 pricing based on NVIDIA HGX B200 pricing

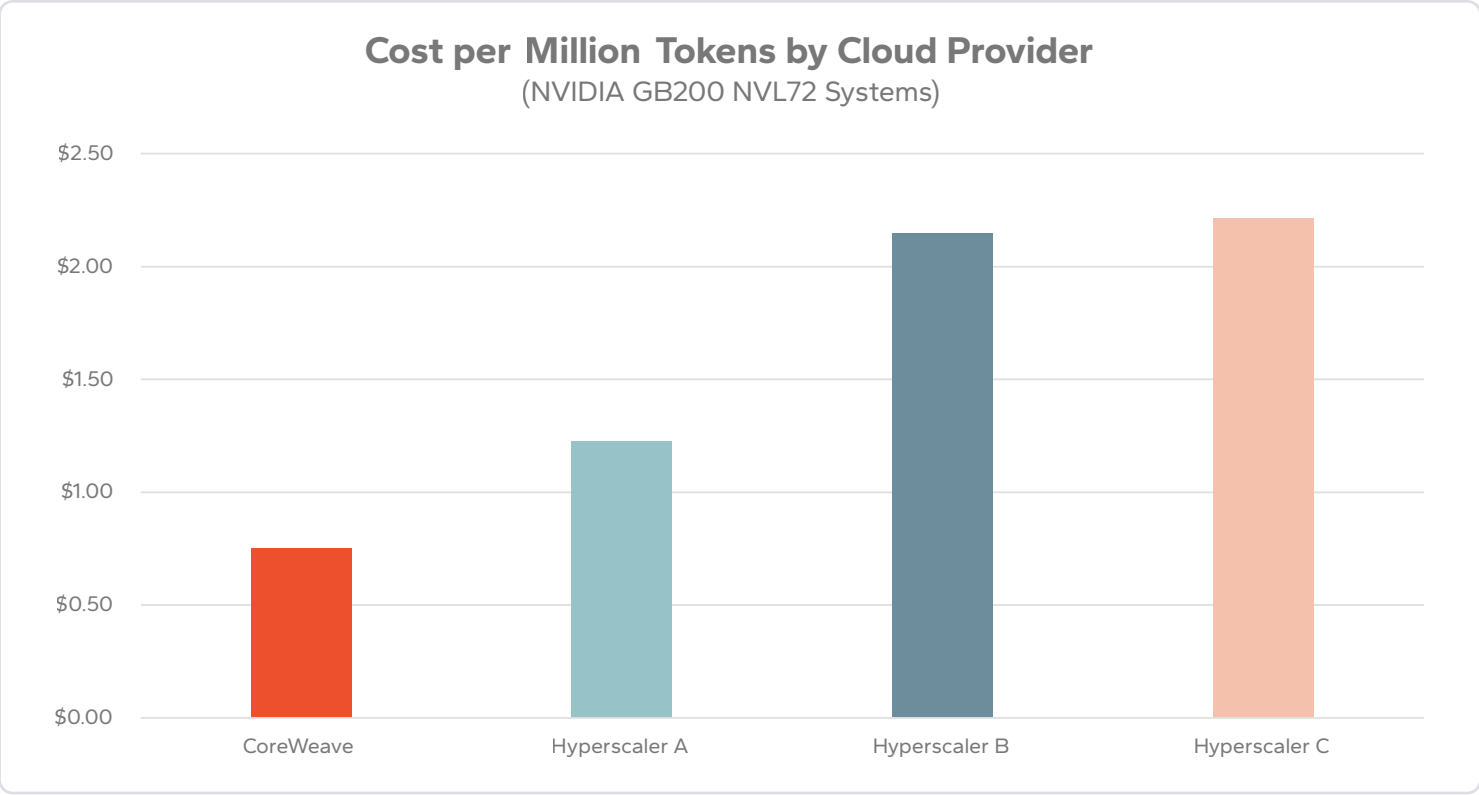


Figure 8: Cost per Million Tokens Comparison

The comparison to Hyperscaler B, which submitted its own MLPerf data, however, clearly shows how CoreWeave’s optimized inference performance can further amplify its cost efficiency. The three-year cost calculations found Hyperscaler B to be the second lowest cost option for NVIDIA GB200 NVL72 deployments, with CoreWeave achieving a 22% cost advantage. When factoring in CoreWeave’s significant performance advantage, however, this price advantage grows to 65% lower cost per million tokens.

For real-world inference deployments, this combination of lower hourly costs and higher token throughput compounds to generate real, tangible savings and more work per dollar. To illustrate the practical impact of these advantages: for the same \$1,000,000 annual infrastructure commitment on NVIDIA GB200 NVL72, an organization would be able to process approximately 1.33 trillion tokens on CoreWeave, compared to 814 billion tokens on Hyperscaler A, 466 billion tokens on Hyperscaler B*, and 451 billion tokens on Hyperscaler C. This means that CoreWeave delivers approximately 64% more token throughput than Hyperscaler A, 186% more than Hyperscaler B, and 195% more than Hyperscaler C for the same fixed spend. At this scale, the combination of CoreWeave’s pricing and performance advantages can translate into substantially greater inference capacity for an equivalent budget, or equivalently, hundreds of thousands to millions of dollars in avoided infrastructure spend for an equivalent workload.

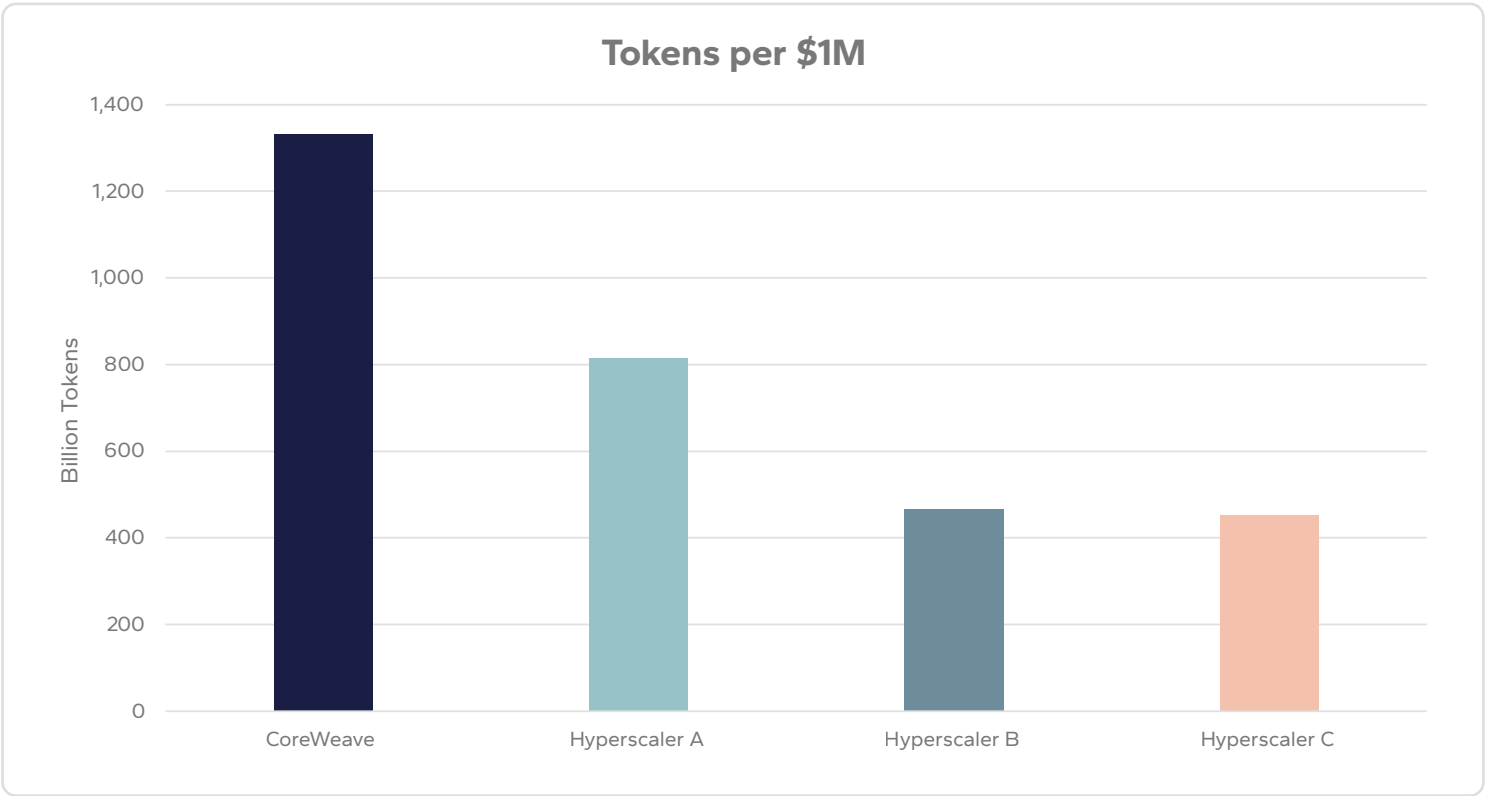


Figure 9: Tokens Processed per \$1 Million

Signal65 Comment: Tokens are the core unit of actual AI work, so achieving more tokens for the same dollar amount directly translates into more work completed. For enterprises, this can mean more customer interactions served, more workflows automated, or more ambitious AI initiatives pursued, without a proportional increase in infrastructure spend. Beyond greater AI productivity, this efficiency can free up budget to reinvest in future business initiatives.

*Hyperscaler B figure based on estimated NVIDIA GB200 NVL72 pricing (see note above).

Serverless Inference Costs

Beyond GPU rental, CoreWeave's inference service offers direct access to popular models via API endpoints. For organizations that need inference capacity but would rather not rent GPUs and operate their own model deployments, this is a simpler path, priced per token.

To fully evaluate how various cloud providers are supporting enterprise inference requirements, this analysis compared the cost of serverless API pricing as a separate lens alongside the GPU rental analysis above. While each cloud provides a different set of available models, CoreWeave was found to offer consistent cost advantages for models offered across services. Across 17 models with at least one direct hyperscaler comparison, CoreWeave provided lower pricing on 11 and identical pricing on 4. In total, CoreWeave achieves a median cost savings of 26% on input tokens, and 15% on output tokens.

A clear example of CoreWeave's advantage can be seen with gpt-oss-120b, a model that is available across all clouds included in this analysis. For this model, CoreWeave achieves significant cost savings compared to all three cloud competitors. CoreWeave input token costs are 80% lower than both Hyperscaler A and Hyperscaler C, and 67% lower than Hyperscaler B. CoreWeave additionally provides large cost savings for output tokens, at 72% lower than Hyperscaler A and Hyperscaler C, and 53% lower than Hyperscaler B.

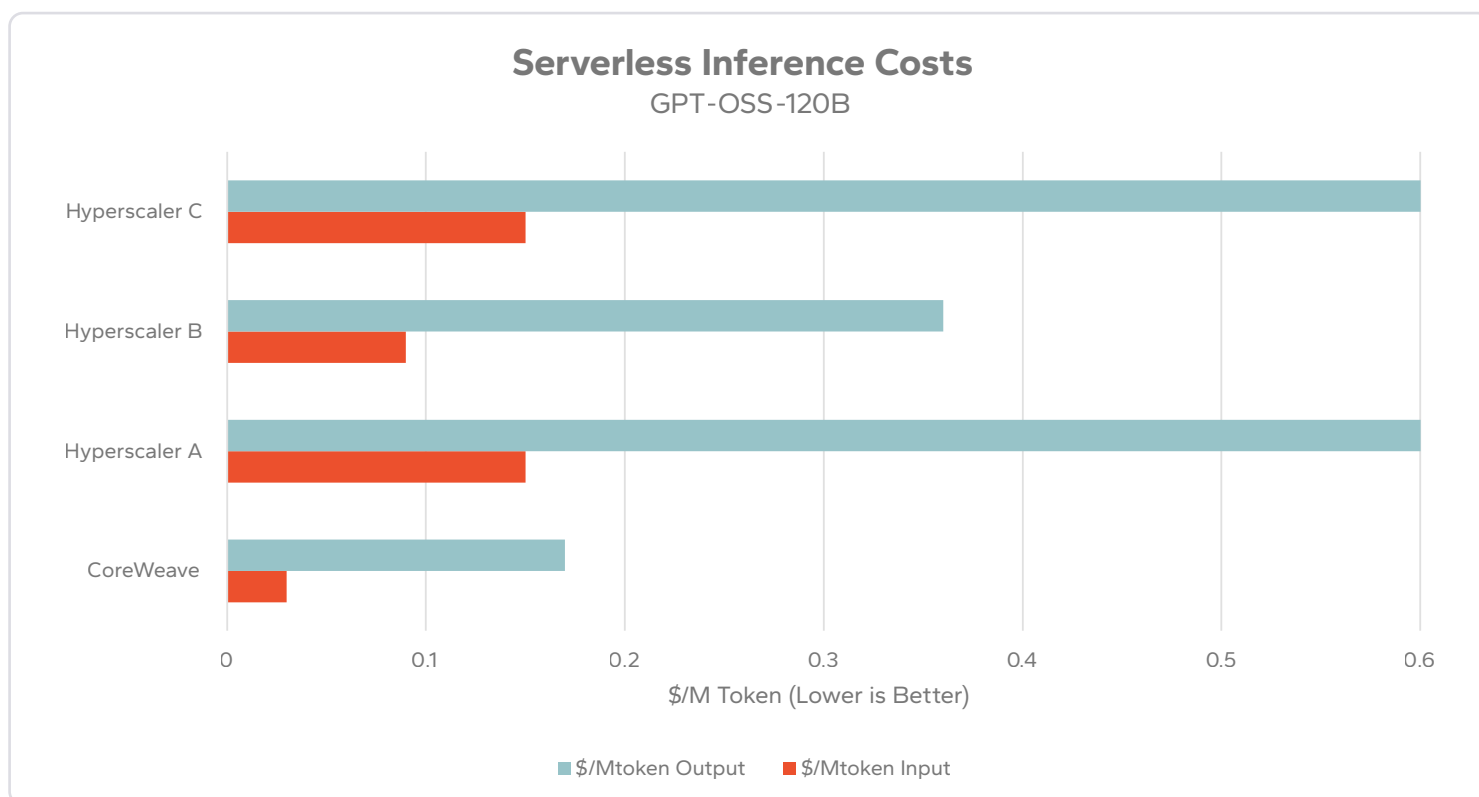


Figure 10: Serverless Inference Costs (gpt-oss-120b)

To illustrate the practical impact of these savings, this analysis models two representative enterprise workloads for gpt-oss-120b, sized using common enterprise token consumption patterns. Figure 11 summarizes the assumed workload characteristics for each scenario.

Metric	Enterprise Search Assistant (RAG)	Enterprise Document Processing
Workload type	Internal/customer-facing knowledge assistant	Claims, contracts, or records processing
Monthly volume	200,000 queries	500,000 documents
Input tokens/month	850M	4.5B
Output tokens/month	150M	500M
Total tokens/month	1B	5B
Input:output ratio	85:15	90:10

Figure 11: Example Workload Scenarios

At the smaller, department-level scale, an organization running the Enterprise Search Assistant workload would spend approximately \$612 annually on CoreWeave's platform, compared to \$2,610 on Hyperscaler A or C, and \$1,566 on Hyperscaler B — savings of \$1,998 and \$954 per year, respectively. At the larger, enterprise-wide scale, the Enterprise Document Processing workload would cost approximately \$2,640 annually on CoreWeave, compared to \$11,700 on Hyperscaler A or C, and \$7,020 on Hyperscaler B — savings of \$9,060 and \$4,380 per year, respectively. Because serverless pricing scales directly with usage, CoreWeave's cost advantage holds steady across both scales, at approximately 77% lower than Hyperscaler A and C, and 61–62% lower than Hyperscaler B.

Key Takeaways

Consistent Cost Advantages for AI Deployments

This study evaluated CoreWeave and three competitive cloud providers and found that CoreWeave consistently achieves cost advantages for various AI deployments. CoreWeave achieved notable cost advantages across the three distinct NVIDIA Blackwell and Blackwell Ultra systems – NVIDIA HGX B200, NVIDIA GB200 NVL72, and NVIDIA HGX B300 – showcasing its ability to support organizations with varying infrastructure needs. Beyond GPU costs, CoreWeave was found to achieve significant savings on storage, support, and observability by eliminating fees commonly found with other cloud platforms. When evaluating NVIDIA Blackwell GPU deployments, CoreWeave achieved a greater than 20% cost advantage in all but one comparison. These consistent cost advantages enable organizations to leverage the latest NVIDIA GPUs to achieve their critical AI deployments at far lower costs than other cloud offerings.

These findings build on Signal65's prior analysis of CoreWeave's NVIDIA H100 deployments, in which CoreWeave was found to deliver up to 47% lower three-year TCO against hyperscaler clouds. When factoring in GPU efficiency, the gap widened to 54%. This analysis finds the same underlying pattern on NVIDIA's Blackwell generation: cost advantages driven by lower-cost GPUs, lower-cost storage, and eliminated fees, with performance further increasing cost advantages.

Compared to the three competitive clouds, CoreWeave was found to offer median cost advantages of 26% on input tokens, and 15% on output tokens across 17 model comparisons – savings which become increasingly impactful with long-running inference workloads.

The consistency of these savings – across various NVIDIA Blackwell GPUs, additional services such as storage, and serverless offerings – provides enterprise organizations a cost-effective route to deploying AI, regardless of their specific workload or infrastructure requirements.

Price and Performance Compound

Beyond the consistent price advantages, CoreWeave also achieves leading performance, and the most consequential finding is that the two advantages multiply. On audited MLPerf results, CoreWeave edged NVIDIA's own reference submission by 4% on per-GPU throughput and beat the one hyperscaler that submitted for this benchmark by 124%. Applied to identical NVIDIA GB200 NVL72 hardware, a 22% three-year price advantage over the closest competitor becomes a 65% advantage in cost per million tokens. In capacity terms, a fixed \$1,000,000 annual commitment buys roughly 1.33 trillion tokens on CoreWeave against 466 billion on that competitor – which means an evaluation run on hourly rate alone systematically understates the gap.

Looking Ahead: Accelerating Further with NVIDIA Vera Rubin

While this report focused on NVIDIA's Blackwell and Blackwell Ultra GPUs, the Blackwell generation is not the end of the roadmap. NVIDIA has introduced NVIDIA Vera Rubin as Blackwell's successor, claiming substantial generational gains — up to a 10x reduction in cost per token, according to NVIDIA's own published figures.

The NVIDIA Vera Rubin platform is already moving from announcement to production, and CoreWeave was the first cloud provider to bring it online — deploying NVIDIA Vera Rubin NVL72 capacity alongside purpose-built Spectrum-X Ethernet networking designed specifically for its rack-scale architecture.

While other clouds, such as the competitors included in this study, have likewise announced plans to bring Vera Rubin capacity to their own platforms, the advantages found in the Blackwell generation indicate that CoreWeave is well-suited to achieve competitive advantages once again. Across the NVIDIA Blackwell generation, CoreWeave demonstrated its ability to offer low-cost GPU access while achieving industry-leading performance results. With NVIDIA Vera Rubin, CoreWeave is bringing the same optimization and purpose-built approach, positioning it well to extend these advantages into the next generation of NVIDIA GPUs.

Appendix: Pricing and Assumptions

This appendix provides the underlying pricing, performance, and modeling assumptions used throughout this report. Hyperscaler Cloud A, B, and C refer to the same three anonymized competitors used consistently in the report body. All pricing reflects publicly available, on-demand rates collected as of August 2026, unless otherwise noted as an estimate.

A. GPU On-Demand Pricing

Published on-demand hourly pricing per GPU, by provider and GPU generation.

GPU	CoreWeave	Hyperscaler A	Hyperscaler B	Hyperscaler C
NVIDIA GB200 NVL72	\$10.50	\$16.01	\$13.04 (estimated)*	\$27.04
NVIDIA HGX B200	\$8.60	\$14.24	\$11.12	\$7.50
NVIDIA HGX B300	\$8.86**	\$17.80	Not publicly available	Not publicly available

*Hyperscaler B does not publish an on-demand NVIDIA GB200 NVL72 rate. This figure is estimated by applying the average GB200-over-B200 price markup observed at CoreWeave and Hyperscaler A (1.17x) to Hyperscaler B's real, published B200 rate. Hyperscaler C is excluded from this markup calculation as a statistical outlier on this specific ratio (its own GB200-over-B200 markup is approximately 3.6x, versus roughly 1.1-1.2x at CoreWeave and Hyperscaler A).

**CoreWeave does not publish NVIDIA HGX B300 pricing. The pricing used in this report is estimated based on Signal65 analysis.

B. Serverless Model Pricing

Published per-token, on-demand serverless pricing for all models with confirmed pricing on CoreWeave's platform and at least one hyperscaler competitor. All figures are US dollars per 1 million tokens. Where a provider does not offer a given model, or pricing could not be confirmed, the cell is marked N/A.

Model	CoreWeave (in/out)	Hyperscaler A (in/out)	Hyperscaler B (in/out)	Hyperscaler C (in/out)
gpt-oss-120b	\$0.03 / \$0.17	\$0.15 / \$0.60	\$0.09 / \$0.36	\$0.15 / \$0.60
gpt-oss-20b	\$0.03 / \$0.13	\$0.07 / \$0.30	\$0.07 / \$0.25	N/A
DeepSeek-V3.1	\$0.55 / \$1.65	\$0.58 / \$1.68	\$0.60 / \$1.70	\$1.23 / \$4.94
DeepSeek V4-Pro	\$1.15 / \$2.55	N/A	N/A	\$1.74 / \$3.48
DeepSeek V4-Flash	\$0.14 / \$0.28	N/A	N/A	\$0.19 / \$0.51
DeepSeek V4-Flash-0731	\$0.13 / \$0.28	N/A	N/A	\$0.44 / \$1.32
Kimi K2.6	\$0.65 / \$3.41	N/A	N/A	\$0.95 / \$4.00

Model	CoreWeave (in/out)	Hyperscaler A (in/out)	Hyperscaler B (in/out)	Hyperscaler C (in/out)
Kimi K2.7 Code	\$0.71 / \$3.50	N/A	N/A	\$0.95 / \$4.00
MiniMax M2.5	\$0.30 / \$1.20	\$0.30 / \$1.20	\$0.30 / \$1.20	N/A
GLM-5	\$1.00 / \$3.20	\$1.00 / \$3.20	\$1.00 / \$3.20	N/A
Qwen3 Coder 480B A35B	\$1.00 / \$1.50	\$0.45 / \$1.80	\$0.22 / \$1.80	N/A
Gemma 4 31B	\$0.10 / \$0.34	\$0.14 / \$0.40	N/A	N/A
NVIDIA Nemotron 3 Super 120B	\$0.20 / \$0.80	\$0.15 / \$0.65	N/A	N/A
Llama 3.3 70B	\$0.71 / \$0.71	N/A	\$0.72 / \$0.72	\$0.71 / \$0.71
GLM-5.1	\$1.40 / \$4.40	N/A	N/A	\$1.54 / \$4.84
GLM-5.2	\$0.76 / \$2.42	N/A	N/A	\$1.54 / \$4.84
MiniMax M3	\$0.23 / \$0.96	N/A	N/A	\$0.33 / \$1.32

C. MLPerf Performance Data

DeepSeek-R1 throughput, Server scenario, NVIDIA GB200 NVL72, MLPerf Inference v6.0 Closed Division. All results are sourced from official MLCommons submissions except where noted as an NVIDIA reference substitution.

Source	GPU Count	Aggregate Throughput (tok/s)	Per-GPU Throughput (tok/s)	Basis
CoreWeave	64	311,531	4,867.7	Own submission
Hyperscaler B	72	156,781	2,177.5	Own submission
NVIDIA	72	336,106	4,668.1	Used for Hyperscaler A and C, which did not submit

Source: MLCommons, mlcommons.org/benchmarks/inference-datacenter/. Hyperscaler A and Hyperscaler C did not submit MLPerf Inference v6.0 results for this benchmark and are modeled using NVIDIA's own reference submission, per the substitution methodology described in the Methodology section.

D. Total Cost of Ownership Modeling Assumptions

Assumptions underlying the 3-year TCO and per-token cost calculations in this report.

Assumption	Value
Contract term	3 years
GPU count	5,000
Object storage	2 PB

Assumption	Value
Storage tier split	15% hot, 35% warm, 50% cold
Regions (data replication)	3
New data generation rate	20% per month
Additional cost categories	Direct Connect networking, Kubernetes service and support, observability (audit log and metric forwarding), support (percent of total spend)
Inference GPU utilization	80%, applied uniformly across all providers
Pricing date	August 2026

Important Information About this Report

CONTRIBUTORS

Mitch Lewis

Performance Analyst | Signal65

PUBLISHER

Ryan Shrout

President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and lab services to many high-tech companies, including those mentioned in this paper.

ABOUT SIGNAL65

Signal65 exists to be a source of data in a world where technology markets and product landscapes create complex and distorted views of product truth. We strive to provide honest and comprehensive feedback and analysis for our clients in order for them to better understand their own competitive positioning and create optimal opportunities to market and message their devices and services.



IN PARTNERSHIP WITH



CONTACT INFORMATION

Signal65 | signal65.com