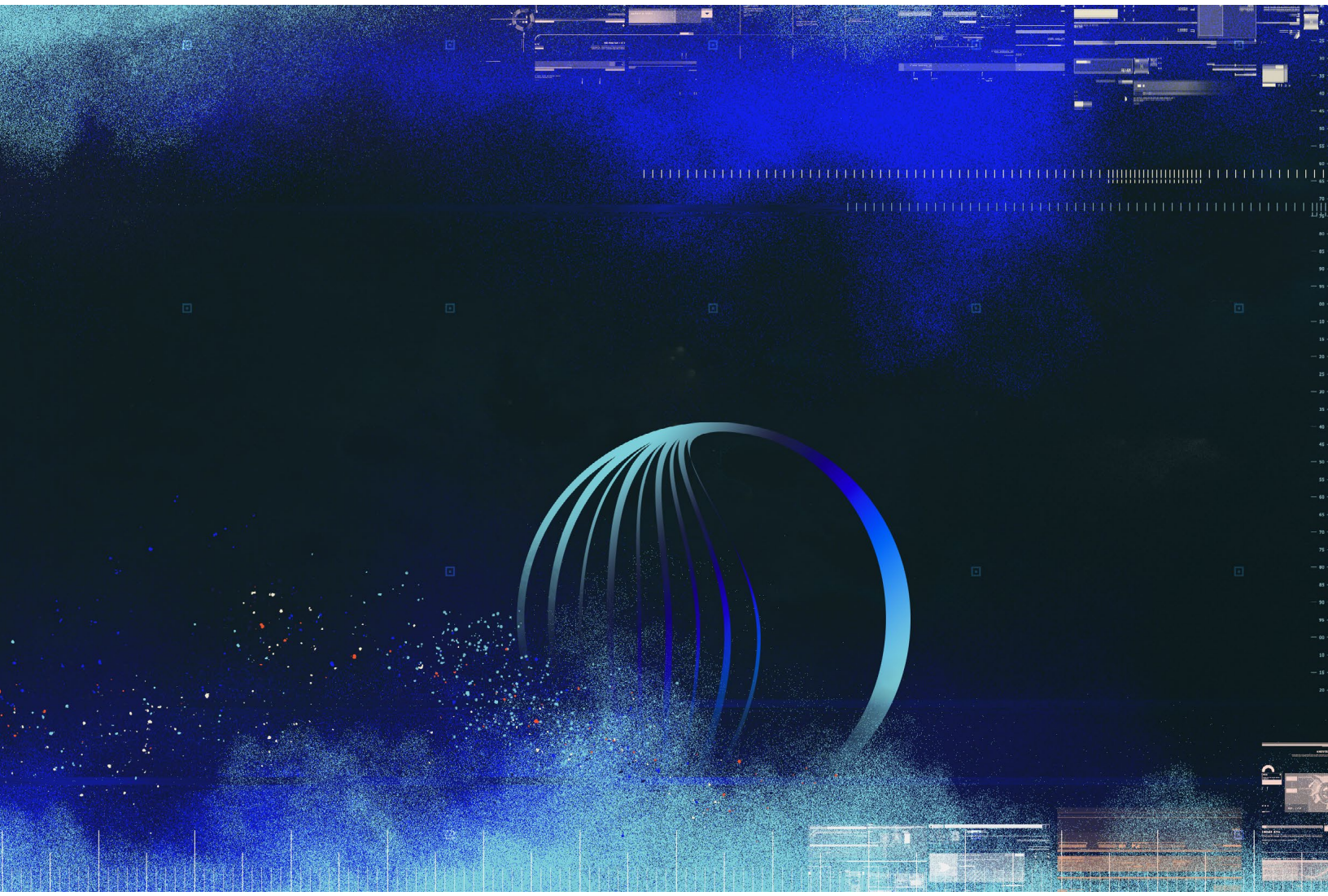
signal65



COREWEAVE PROVIDES LEADING COST AND EFFICIENCY FOR AI:

A 3-Year TCO Analysis of AI Cloud Deployments

AUTHORS

Mitch Lewis
Performance Analyst | Signal65

Russ Fellows
VP, Labs | Signal65

IN PARTNERSHIP WITH

CoreWeave

APRIL 2026

Executive Summary

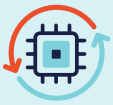
For organizations looking to leverage AI infrastructure in the cloud, several options exist ranging from traditional hyperscalers to newer, AI-optimized clouds such as CoreWeave. Given the high performance infrastructure requirements and potential scale of AI deployments, organizations must carefully consider the financial impact of their various options.

This study presents a Total Cost of Ownership (TCO) analysis of AI cloud deployments, comparing CoreWeave to average hyperscaler pricing. This report evaluates the impact of specific costs, including GPUs and storage, and additionally evaluates the impact of CoreWeave's GPU efficiency.

The right AI infrastructure strategy doesn't just support scale, it fundamentally changes the cost equation. Key findings include:

Up to 47% **lower** TCO over 3 years

Up to 96% **more** TFLOPs per dollar



Up to 54% **lower** TCO when normalized for GPU efficiency

Cloud Infrastructure Options for AI Workloads

AI workloads, both model training and inference, are among the most infrastructure-intensive workloads in modern IT. Training and serving AI models typically requires significant computational resources, particularly GPU acceleration, along with high-performance networking and large-scale storage capable of supporting massive datasets. While AI technologies offer transformative capabilities across industries, the infrastructure required to support these workloads can represent a substantial cost and operational barrier for many organizations, particularly at scale.

Cloud infrastructure provides a practical path for organizations to access the compute, storage, and networking resources required for AI without the capital investment and operational complexity associated with building on-premises GPU clusters. Cloud service providers offer on-demand

access to high-performance GPUs, distributed storage systems, and scalable networking, enabling organizations to rapidly provision infrastructure and scale AI workloads as needed. However, while cloud platforms improve accessibility to advanced AI infrastructure, the cost of operating large-scale AI workloads remains a critical consideration for IT and financial decision-makers.

The Cloud Divide: General-Purpose vs AI-Native

Organizations evaluating cloud infrastructure for AI workloads today typically need to choose between two primary categories of providers: large hyperscale cloud platforms and specialized AI-focused cloud providers. Hyperscalers—including providers such as AWS, Microsoft Azure, and Google Cloud—offer broad, general-purpose cloud platforms that support a wide range of enterprise workloads. These platforms provide GPU-accelerated compute instances and AI services alongside extensive ecosystems of storage, networking, data, and platform services. While hyperscalers deliver mature infrastructure and integrated cloud services, their pricing and infrastructure models are generally designed for multi-purpose enterprise workloads rather than being optimized specifically for large-scale AI training and inference.

In contrast, CoreWeave, as a specialized AI cloud provider, designed its infrastructure specifically to support AI workloads. CoreWeave focuses on delivering GPU-dense compute clusters, high-throughput networking, and storage systems optimized for large-scale model training and inference pipelines. By concentrating on AI-centric infrastructure rather than general-purpose cloud services, specialized AI cloud providers often deliver simplified architectures, optimized performance for GPU workloads, and unique pricing models designed specifically around AI compute utilization.

As organizations scale AI initiatives, understanding the cost and performance implications of these different cloud infrastructure approaches becomes increasingly important. This paper examines the total cost of ownership (TCO) of running AI workloads on traditional hyperscale cloud platforms compared to CoreWeave's AI-optimized cloud.

TCO Overview

To evaluate the TCO of cloud-based AI infrastructure, Signal65 modeled costs for both CoreWeave and the large hyperscale cloud providers. This modeling included costs for GPUs, storage, networking, observability, and support. For this analysis, cloud costs were modeled using published on-demand pricing. In practice, customer pricing often varies based on contractual agreements, including factors such as committed usage levels, contract duration, and negotiated discounts.

To accommodate the varying needs of AI customers, three distinct configurations were modeled, each with different GPU and storage requirements. These configurations were built to approximate various mixtures of training and inference, as well as demonstrate varying levels of scale. Storage capacities were selected based on extensive modeling and lab testing conducted by Signal65. Additional details on the storage modeled in this analysis can be found in the appendix. The financial models utilize NVIDIA H100 GPUs which are widely deployed today. An overview of the three configurations modeled can be seen in Figure 1.

Configuration	Total NVIDIA H100 GPUs	Workload Mix (Inference / Training)	Total Storage Capacity
Small	72	85% / 15%	1.785 PB
Medium	576	70% / 30%	11.76 PB
Large	4,608	50% / 50%	67.2 PB

Figure 1: *Configuration Overview*

All models calculated the total costs for a 3-year period, utilizing pricing for NVIDIA H100 GPUs, the required storage capacity, and any additional costs for networking, observability, and support. To create a balanced comparison between CoreWeave and various competitive cloud offerings, the hyperscaler comparison utilized a blended average of hyperscale cloud costs.

Signal65 Comment *– The average hyperscaler costs used in this study represent a generalization of leading cloud providers and utilize a mixture of pricing from various clouds. To accommodate large price differentiations between cloud providers, costs were weighted by the approximate market share of each cloud. It should be noted that this approach most heavily weighted the lowest cost hyperscale solutions and that a direct analysis of individual clouds may find costs that exceed those shown in this study.*

Results

Overview

In total, CoreWeave was found to be far more cost effective across all three configurations modeled, ranging from 44% to 47% less expensive than the hyperscale alternative. Cost breakdowns for each configuration can be seen in Figure 2, Figure 3, and Figure 4.

Signal65 Comment *– A key difference between CoreWeave and hyperscaler clouds is how observability and support are priced. CoreWeave includes these capabilities at no additional cost, while hyperscaler clouds price them as add-ons that increase TCO.*

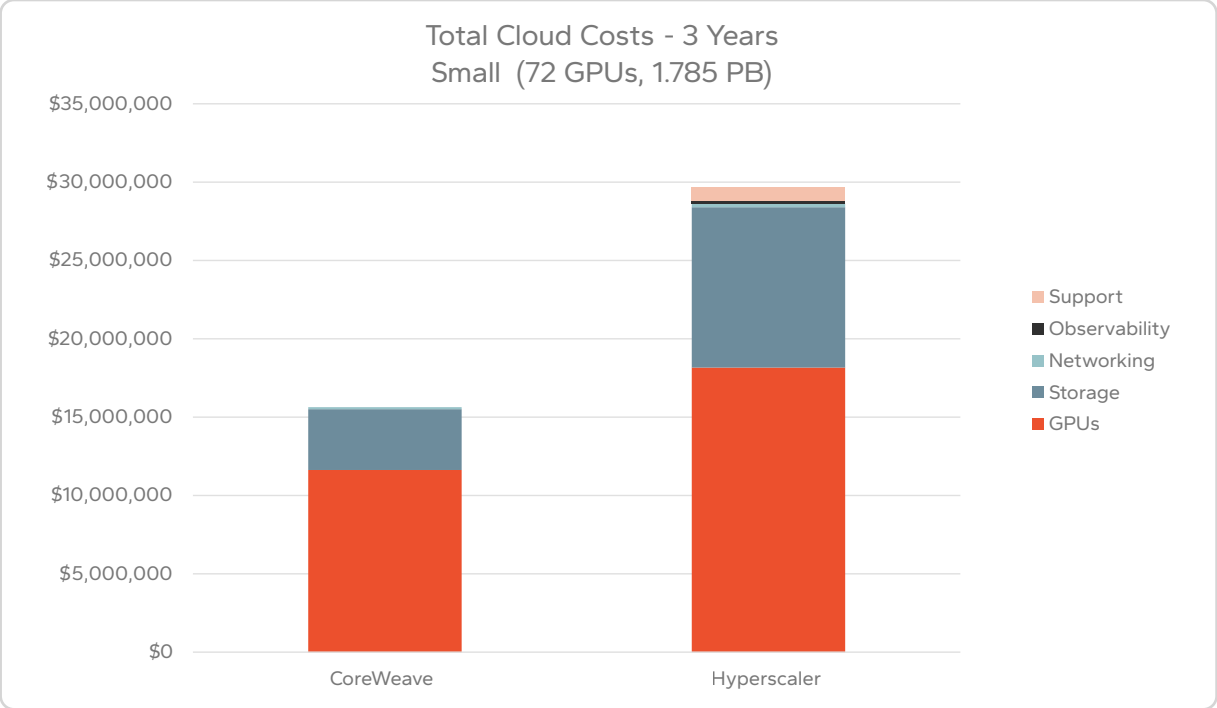


Figure 2: *Total Cloud Costs Over 3 Years – Small Configuration*

In total, the cost for the small configuration amounted to $15,630,463 for CoreWeave, while totaling $29,676,383 for the composite hyperscaler. This configuration demonstrated the largest total price differential, with CoreWeave achieving a 47% price advantage.

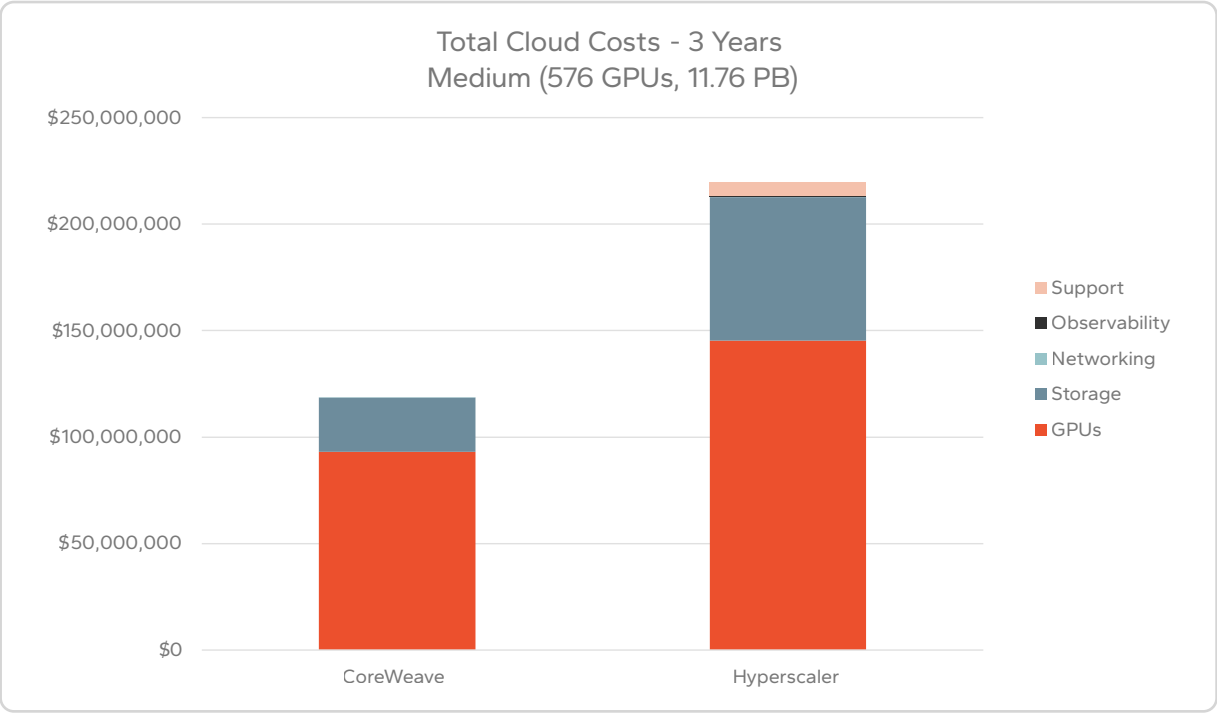


Figure 3: *Total Cloud Costs Over 3 Years – Medium Configuration*

In the medium configuration, CoreWeave achieved a similar overall cost advantage of 46%. The total cost of this configuration was calculated to be $118,655,506 in CoreWeave and $219,889,021 in the composite hyperscaler.

Figure 4: *Total Cloud Costs Over 3 Years – Large Configuration*

The large configuration represents a very significant AI deployment, with substantial cloud infrastructure investments. For CoreWeave, the total costs in this configuration amounted to $890,238,251. Comparatively, the same configuration cost $1,597,841,349 in the hyperscaler. For this configuration, CoreWeave was found to achieve 44% lower costs over three years.

At this scale, this type of cost difference results in dramatic cost savings. While 44% was the smallest advantage achieved by CoreWeave, this amounts to a difference of $707,603,098 over three years – more than the total costs of the small and medium configurations in either cloud.

GPU Costs

GPUs are the core driver of AI computing and the primary focus for most organizations deploying AI workloads—they also drive a significant portion of the total cost. Throughout this modeling exercise, GPUs accounted for the largest portion of the overall AI cloud costs in all three environments.

This analysis modeled the cost of NVIDIA H100 GPUs, utilizing the hourly on-demand cost per GPU in each cloud evaluated. Overall, CoreWeave was found to offer a lower hourly cost than the hyperscale clouds, at $6.155/GPU/hour. In comparison, the average hyperscaler cost was calculated to be $9.606/GPU/hour. In total, this comes to a 36% cost savings on GPUs alone. The total GPU cost savings ranged from $6,526,800 in the smallest configuration, to $417,715,226 in the largest.

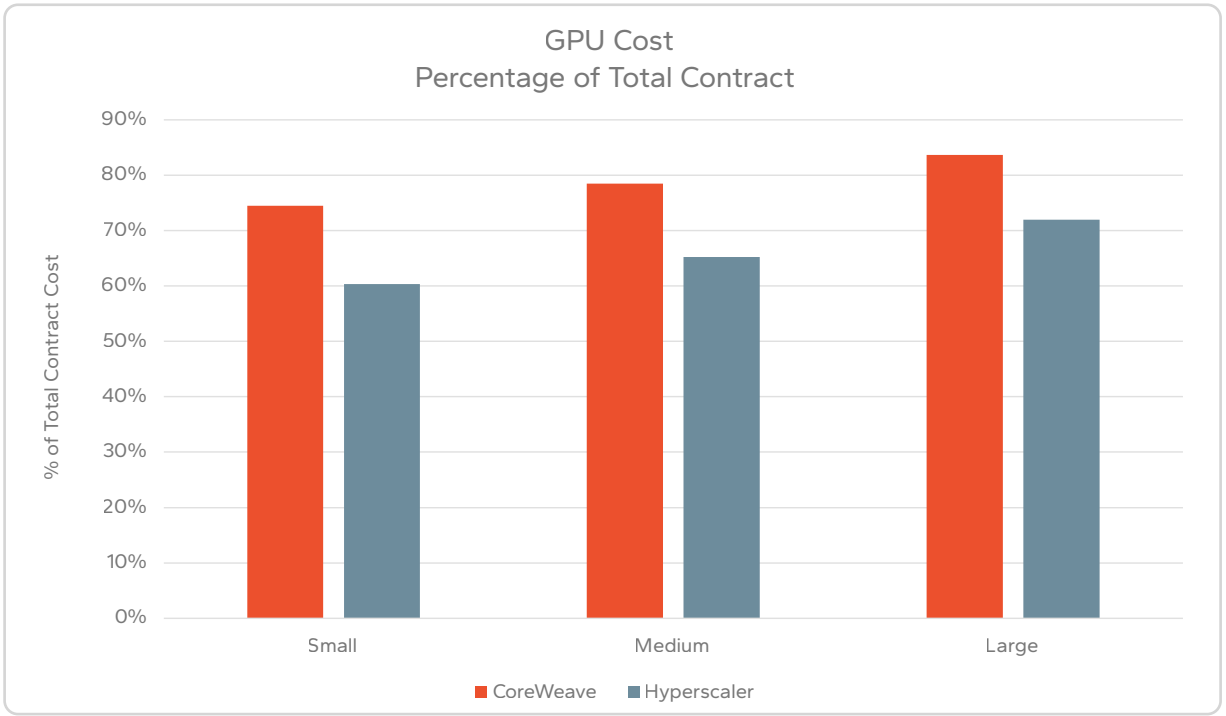


Figure 5: *GPU Costs as Percentage of Total Contract*

On average, GPU costs alone were found to account for 78.88% of the total contract costs in CoreWeave and 65.85% in the hyperscaler. This demonstrates that in hyperscale clouds, a larger portion of the total cost is consumed by ancillary services, such as data storage, networking, data transfer, support, and observability. CoreWeave achieves a higher percentage of GPU costs while maintaining a lower hourly rate, which is particularly notable. For customers, this means paying less overall, with more of their money being attributed directly towards the most valuable component in the AI infrastructure stack.

GPU Efficiency

Beyond offering lower raw costs than hyperscale alternatives, CoreWeave has additionally differentiated itself with its infrastructure efficiency. For large scale AI deployments, GPU efficiency becomes a significant factor to consider alongside traditional cost analysis. GPU efficiency is a measurement of how effectively GPUs are used to perform actual training or inference computations relative to their maximum theoretical capability. This efficiency is commonly measured by two key metrics:

- **Model FLOPs Utilization (MFU)** – MFU measures how much of the theoretical GPU compute capacity is actually used for model training. This is achieved by calculating a ratio of the actual model flops achieved to the maximum theoretical FLOPs (TFLOPs) rate of the GPU.
- **Goodput** – Goodput measures the percentage of time that a system is doing useful training work compared to other interruptions, such as checkpointing or node crashes.

Utilizing these two metrics gives a clear view into infrastructure efficiency – which when considered alongside TCO, ultimately translates into the value provided by a cloud service.

In practice, achieving 100% GPU utilization is unrealistic, with MFU values often falling in the 35–45% range for most real-world workloads. Communication overhead, memory access constraints, and synchronization across distributed systems all limit how much of the theoretical peak performance can be used for model computation.

CoreWeave has notably achieved both MFU and goodput metrics that exceed industry averages. While industry averages for MFU are typically in the range of 35% - 45%, CoreWeave has achieved MFUs exceeding 50%[1]. Additionally, CoreWeave has achieved goodput of 96%, outperforming an industry average of approximately 90%[2].

Utilizing the peak TFLOPs for an NVIDIA H100 cluster, alongside MFU benchmark data from MosaicML, GPU efficiency can be calculated for both CoreWeave and industry benchmarks. It should be noted that the benchmarks used in this calculation are not directly sourced from any hyperscale cloud, and instead represent a broad industry average.

[1] https://www.coreweave.com/blog/coreweave-leads-the-charge-in-ai-infrastructure-efficiency-with-up-to-20-higher-gpu-cluster-performance-than-alternative-solutions

[2] https://www.coreweave.com/blog/achieve-ai-infrastructure-goodput-of-up-to-96-with-3-key-strategies

Delivered Performance	CoreWeave	Industry Average
MFU	49.2%[3]	41.85%
Goodput	96%[4]	90%
NVIDIA H100 Peak Theoretical TFLOPs (BF16)	1,979	1,979
Realized TFLOPs (accounting for MFU)	974	828
Effective TFLOPs (accounting for goodput)	935	745
Efficiency	**47%**	**38%**
Relative efficiency difference in TFLOPs	25.4%	-20.3%

Figure 6: *GPU Efficiency*

While no detailed GPU efficiency information is publicly available for specific hyperscale cloud vendors, these industry metrics can be used to approximate how GPU efficiency can impact TCO.

Figure 7: *Effective TFLOPs per Dollar Comparison*

Utilizing the effective TFLOPs calculated and the hourly GPU costs in each cloud, CoreWeave is found to achieve 96% more TFLOPs per dollar. In practice, this means more GPU processing per dollar, further boosting the TCO value of CoreWeave for large scale AI processing.

[3] MFU can vary depending on cluster size as well as other factors. The MFUs used in this comparison were measured from clusters of 128 NVIDIA H100 GPUs during training runs of a 30B MPT model with a context window of 2k tokens.

[4] Based on goodput measurements of 96% or more observed across CoreWeave customer clusters ranging between 4K and 15K GPUs.

The financial impact of this efficiency advantage becomes clear when normalizing the 3-year TCO model for equivalent AI output. The base TCO model assumes a fixed number of GPUs for each environment size over a 3-year period. However, higher effective TFLOPs enable more work to be completed per GPU.

As a result, environments with higher efficiency can achieve the same total output using fewer GPUs. To reflect this, the configurations below scale GPU counts to deliver equivalent compute output across providers.

Environment	CoreWeave GPUs (performance-adjusted)	3 Year GPU Cost	Hyperscaler GPUs	3 Year GPU Cost
Small	59	$9,538,221	72	$18,166,663
Medium	466	$75,335,782	576	$145,333,309
Large	3,726	$602,362,926	4,608	$1,162,666,477

Figure 8: *Performance-adjusted GPU Requirements*

This performance-adjusted view highlights the infrastructure efficiency advantage. While CoreWeave already delivers significant TCO savings without accounting for GPU efficiency, normalizing for efficiency further amplifies these benefits. By requiring fewer GPUs to achieve the same result, CoreWeave increases its GPU cost advantage from 36% to 48%, and its total 3-year TCO advantage from 47% to 54%.

Signal65 Comment *– It is important to note that the efficiency metrics utilized for the hyperscaler are not specific to any particular cloud. Hyperscale cloud vendors do not typically publish these efficiency metrics, and therefore industry average numbers were substituted as an approximation. Actual hyperscale GPU efficiency may vary; however, this comparison demonstrates CoreWeave's significant advantage over the efficiency that is seen in AI datacenters from comparable industry examples.*

Storage Costs

After GPUs, data storage accounts for the second most substantial portion of total contract costs. The deployment of high-density AI infrastructure creates unique data center requirements. At rack-scale and larger, high-speed storage is required in order to maintain high utilization rates of the GPUs to maximize their economic value.

Modeling storage for AI presents a unique challenge, as the specific performance and capacity requirements can vary drastically depending on specific workload mixtures and overall environment size. The storage in this TCO comparison was modeled to meet the scale and workload mixtures defined in the small, medium, and large configurations. Each environment was modeled with two storage tiers: a high performance buffer, and an object storage capacity tier. The specific capacity

points chosen were based on in-depth analysis and performance testing conducted by Signal65. An overview of the storage for each environment can be found in Figure 9 and more details on the performance justifications are provided in the appendix.

Configuration	Total GPUs	Workload Mix (Inference / Training)	High Performance Tier	Object Storage Tier	Total Capacity
Small	72	85% / 15%	85 TB	1.7 PB	1.785 PB
Medium	576	70% / 30%	560 TB	11.2 PB	11.76 PB
Large	4,608	50% / 50%	3.2 PB	64 PB	67.2 PB

Figure 9: *Storage Configuration Overview*

A key distinction between CoreWeave and hyperscaler clouds is in the storage services offered to meet AI-specific needs. In hyperscaler clouds, object storage can be utilized, but typically needs to be paired with a higher performance file system offering, such as Lustre, to meet high performance requirements. Unlike hyperscalers, CoreWeave delivers high performance directly from the object storage layer, eliminating the need for a separate, high-cost Lustre tier. Built into CoreWeave AI Object Storage, is a functionality called Local Object Transport Accelerator (LOTA), which acts as a high-performance proxy that automatically caches frequently accessed or pre-staged data onto local NVMe disks within each GPU node to accelerate reads. An overview of the storage services modeled and their prices, can be found in Figure 10.

	CoreWeave	Hyperscaler
High Performance Tier	LOTA Storage — Included (no additional cost)	Hyperscaler Managed Lustre - $0.60 ($/TB/Month)
Object Storage Capacity Tier	CoreWeave AI Optimized Storage - $0.06 ($/TB/Month)	Object Storage - $0.13 ($/TB/Month)
API Requests (PUT/GET)	No Additional Fees ($0 per request)	Charged per request
Data Retrieval	No Additional Fees ($0 per request)	Charged per GB
Data Egress	No Additional Fees ($0 per request)	Charged per GB

Figure 10: *Storage Cost and Fee Comparison*

Signal65 Comment *– Our financial models assume all data resides in the "hot" tier of AI Object Store. However, if access patterns change, the lower rates will automatically be applied, making the line item for storage costs for CoreWeave the "up to" value. That is, costs could be lower, but never greater than our model.*

Notably, CoreWeave offers a significant price advantage for both storage tiers. When compared to CoreWeave's AI Object Store, hyperscaler object storage is over 2x more expensive. Even more impactful is the cost difference for high-performance storage. CoreWeave's LOTA is included at no additional cost, effectively pricing high-performance storage at the same $0.06/TB/month as its object storage tier — representing a 10x cost advantage compared to hyperscaler managed Lustre offerings.

Additionally, CoreWeave does not charge for API requests (GET/PUT) or data egress. This is critical for both model training and modern inference workloads. In particular, KV-cache offload and retrieval for long-context inference is a major driver of storage access and data movement, requiring frequent reads and writes of intermediate state to reduce compute requirements. Training workflows similarly involve frequent movement of large datasets and checkpoints. Over a 36-month period, the elimination of these fees provides greater budget predictability.

Signal65 Comment *– The elimination of API request and data egress fees is a significant differentiator for CoreWeave. At petabyte scale, these charges can amount to millions of dollars over three years. Beyond the raw costs, the greater challenge is their unpredictability – introducing ongoing complexity in cost forecasting. By removing these fees entirely, CoreWeave delivers not only lower total storage costs but also a simpler, more transparent pricing model without hidden variables.*

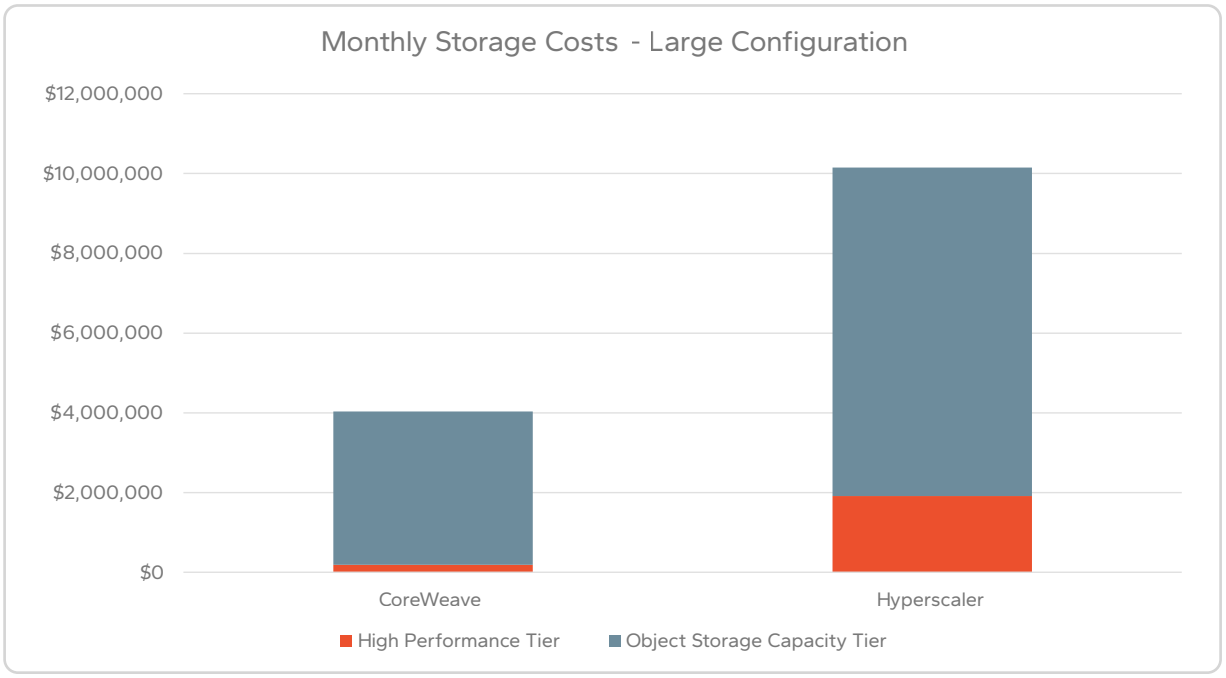


Figure 11: *Monthly Storage Costs – Large Configuration*

The significance of the storage cost savings becomes increasingly apparent when evaluating the monthly cost. Depending on the configuration, CoreWeave reduced storage costs between $162,690 and $6,124,800 per month compared to hyperscalers. These savings are most dramatic in the large configuration – with a total 67 PB of storage – in which CoreWeave reduces the high performance storage costs by $1,728,000/month and reduces the object storage costs by $4,396,800/month.

Configuration	CoreWeave 3-Year Storage Cost	Hyperscaler 3-Year Storage Cost	Total Savings
Small	$3,855,600	$10,226,520	$6,370,920
Medium	$25,401,600	$67,374,720	$41,973,120
Large	$145,152,000	$384,998,400	$239,846,400

Figure 12: *3-Year Storage Costs*

In total, data storage costs contributed between 16% and 34% of the total 3 year costs, depending on the configuration. Over a 3 year period, the lower storage costs available in CoreWeave were found to save between $6,370,920 and $239,846,400.

Additional Costs: Networking, Orchestration, Support, and Observability

The remaining costs factored into the model are networking, orchestration, support, and observability. In total, these costs represent only a small fraction of the total 3 year cost, however, there are some notable differences between CoreWeave and hyperscale clouds.

Networking costs were included with 10Gbps direct connect networking for three regions. This cost was static across the three configuration sizes modeled. Over the 3 years modeled, these networking costs totaled $135,000 in CoreWeave and $195,523 in the hyperscaler.

For the remainder of the additional costs, CoreWeave and hyperscaler pricing differs significantly, as these components are all included in CoreWeave at no additional cost.

- Orchestration: CoreWeave includes a managed Kubernetes service called CoreWeave Kubernetes Service (CKS) that is purpose built for AI workloads at no additional cost. To include similar functionality in the hyperscaler alternative, pricing for a hyperscale managed Kubernetes service and extended support for Kubernetes versions was included.
- Observability: Observability is an additional area that CoreWeave offers at no additional cost. Hyperscaler observability costs were calculated using an estimation of audit log costs and metrics costs. For audit log costs, it was assumed that the audit log rate was 10 TB/month for each of the three configurations. Metrics were assumed to be recorded at a rate of 250 samples / GPU / second with 1% forwarded to the observability platform.
- Support: Like orchestration and observability, CoreWeave does not charge additional fees for support. Hyperscaler support plans are typically priced as a percentage of total cloud spend, with effective rates generally ranging from 3% to 10% depending on support tier and overall usage. Given the large overall costs of the AI infrastructure included in this model, a conservative support fee of 3% was included in the total cost.

While these additional costs ultimately make up a small portion of the total 3-year TCO, CoreWeave's inclusion of these components for no additional cost enables notable savings on top of the already lower GPU and storage costs.

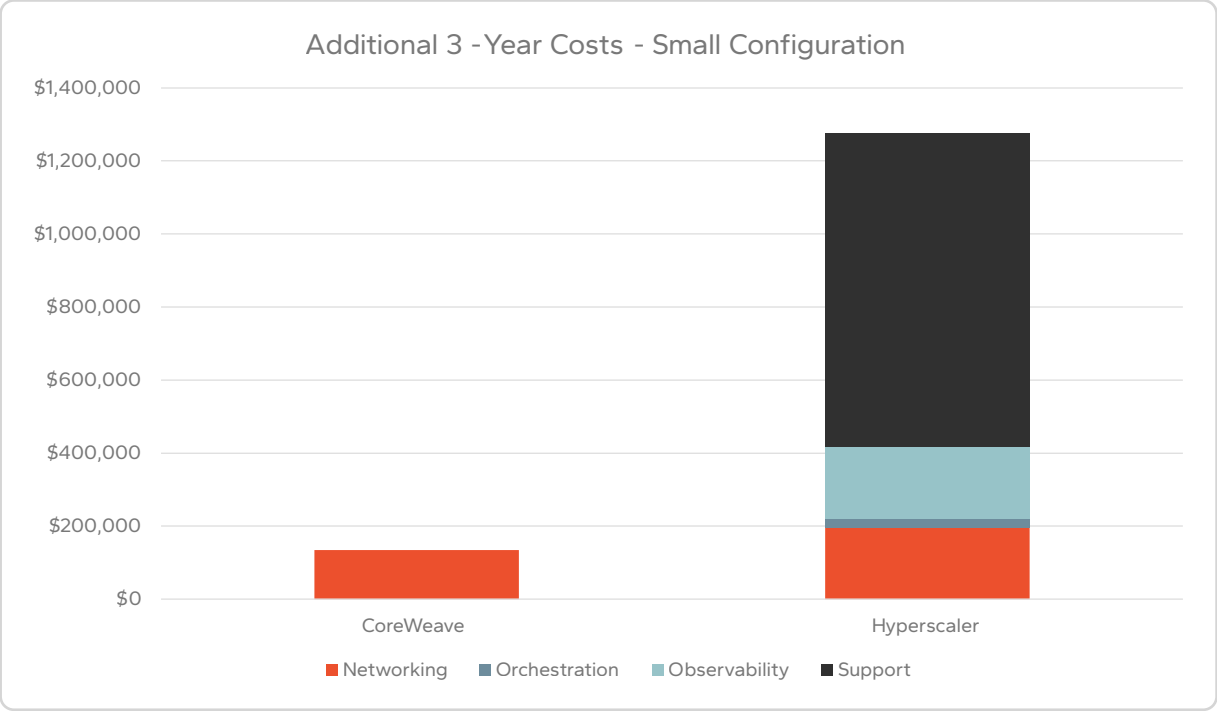


Figure 13: *Additional 3-Year Costs – Small Configuration*

At the smallest configuration, CoreWeave totaled $135,000 for these costs – attributed entirely to networking. Meanwhile, the Hyperscaler totaled $1,275,316 resulting in $1,140,316 in savings. As the configurations scale larger, CoreWeave maintains a cost of $135,000 for networking, while the hyperscaler costs grow additionally large, primarily due to the relative growth of support costs. In the large configuration, the additional hyperscaler costs reach $50,168,588 with $46,532,442 attributed to support.

Final Thoughts – CoreWeave Provides Leading Cost and Efficiency for AI

Given the requirements of high performance GPUs and petabyte-scale storage, AI workloads of all sizes require significant IT investment. While cloud services provide a flexible, scalable solution for organizations to leverage AI infrastructure, not all clouds are the same when considering TCO. Unlike traditional hyperscalers, CoreWeave provides a fully AI-focused cloud, with offerings purpose-built to meet the requirements of large-scale AI workloads.

This study evaluated the TCO of running AI workloads in CoreWeave compared to hyperscalers, and identified several significant advantages. In total, CoreWeave was found to provide up to 47% lower costs over three years, with significant cost advantages found across all core components modeled. When evaluating GPUs, the most significant portion of the overall TCO, CoreWeave achieved up to 36% lower hourly costs, while offering up to 96% more TFLOPs per dollar. CoreWeave was additionally found to be far more efficient when considering data storage, with its AI Object Storage and high performance LOTA tier – offering up to 62% lower storage costs. Costs for ancillary services were also far lower in CoreWeave, with no additional cost required for Kubernetes orchestration, observability, or support.

Beyond lower total cost of ownership, CoreWeave has invested significant engineering effort to deliver industry-leading efficiency for AI workloads. CoreWeave achieves MFU and goodput compared to typical industry benchmarks, driven by optimizations across the full stack. This includes built-in GPU telemetry and automated mitigation mechanisms, such as detection of underperforming or faulty GPUs and intelligent scheduling that replaces impacted resources without interrupting workloads. As a result, organizations can achieve more consistent performance and extract greater value from their infrastructure investments. When normalizing the 3 year TCO model for GPU efficiency, CoreWeave's cost advantage grew up to 54%.

While hyperscale clouds have long provided flexible IT services to meet various enterprise workloads, AI presents an entirely unique set of cost and infrastructure challenges. As organizations of all sizes increasingly leverage AI, CoreWeave provides an AI-optimized cloud to meet various workload demands, lower total cost of ownership, and increase efficiency. These benefits extend beyond training into inference, where consistent performance, efficient resource utilization, and predictable costs are critical. As workloads evolve toward more dynamic, agentic AI systems with continuous demand, CoreWeave's optimized infrastructure enables organizations to deliver responsive, cost-efficient AI services in production.

Appendix

Detailed Storage Analysis

The financial model is grounded in a set of operational parameters essential for large-scale AI deployments supporting inferencing, training and testing. The financial model utilizes a rack scale building block, where a rack of GPU nodes such as NVIDIA H100 or an NVIDIA NVL-72 system serves as the primary economic and technical unit.

Configuration	Total GPUs	Workload Mix (Inference / Training)	High Performance Tier	Object Storage Tier	Total Capacity
Small	72	85% / 15%	85 TB	1.7 PB	1.785 PB
Medium	576	70% / 30%	560 TB	11.2 PB	11.76 PB
Large	4,608	50% / 50%	3.2 PB	64 PB	67.2 PB

Storage utilization is determined by the specific mixture of training and inference. Specifically, inferencing places the greatest burden on storage performance, and capacity utilization, due to the use of KV-Cache offloading to maintain peak GPU efficiency.

In order to be effective, the KV-Cache must also meet specific performance requirements, of up to 50 GB/s read and write rates per rack, with peak rates possibly exceeding 100 GB/s. While the use of high-speed shared storage can meet these requirements, its cost is excessive.

Moreover, our modeling uses two-tiers of storage, a high-speed buffer to absorb the majority of KV-cache I/O, coupled with a much larger object storage capacity tier. Additionally, our modeling indicates that up to 2 PB of data may be generated per day for KV-Cache offloading to storage. As such, we specify this as the minimum storage capacity per rack, with capacities scaled appropriately for the larger 8 and 64 rack deployment sizes.

Signal65 Comment *– Research and analysis by Signal65 into KV-Cache offloading to storage has found that it has the potential to increase the effective GPU utilization rates by up to 20x. Although this 20x is the maximum potential benefit, in practice the use of KV-Cache offload can increase the efficiency of output token generation significantly. Moreover, our models have demonstrated that a single NVIDIA GB200 NVL72 rack of GPUs can produce up to 2 PB of storage per day.*

Technical Performance Requirements

To avoid "IO-wait" bottlenecks where valuable GPU resources sit idle, the storage backend must deliver performance that matches the compute density of the GPU architecture.

- **High-Speed Buffer:** To minimize data movement costs and avoid re-hydration fees, the model assumes a baseline of 100 TB of high-speed capacity per rack for 100% inference. This is then scaled by the inference percentage of each configuration (e.g., 85 TB for the 1-rack config). In practice, CoreWeave LOTA may enable even larger cache sizes than the 100 TB per rack modeled, depending on the infrastructure utilized.
- **Object Storage Requirement:** Total storage required is calculated at 2 PB per rack for 100% inference, similarly adjusted for each configuration (e.g., 64 PB of object storage for the 64-rack config).

Each head node requires a dedicated 20 GB/s (160 Gbps) sequential read data stream. This is driven by the requirements of the inference cycle and the intensity of training epochs together with KV-Cache load and unloading. Additionally, the write performance bandwidth is equally important, although the latency requirements are significantly reduced. These requirements are based upon several factors:

1. **Training:** Fine-tuning workloads can compose up to 50% of usage. For a 2 PB rack environment, this equates to 1 PB read with an average object size of 64 MB, the system must process over 15,000 GET operations per training epoch. As a result, the storage must support millions of requests without
latency spikes.
2. **KV-Cache Offload:** During inference, the primary driver for storage is Key-Value (KV) Cache offload. The KV-cache stores states and tokens to speed up inference. Context lengths grow in multi-turn conversations or long-form document extraction. For example, one session can be 43 GB for a session using the Llama 3.1-70B model.
3. **Latency Budget:** To maintain a "Time to First Token" (TTFT) budget of less than 200ms, the storage tier must deliver sequential reads at 20 GB/s for rehydration. For modern models like Llama 3.5 70B, which generate approximately 0.32 MB of KV-cache per 1,024 tokens, a multi-turn coding session (32,768 tokens) generates a 10.2 GB rehydration request.

Lustre Performance

Standard object storage (S3 Standard, Azure Blob Hot, or GCS Standard) cannot natively deliver the consistent 20+ GB/s per GPU head node required. To bridge this gap, active training data and KV-caches must be moved to a high-performance parallel file system like Lustre.

- **Performance-Bound Sizing:** Services like AWS FSx for Lustre and GCP Parallelstore offer a performance tier of 1,000 MB/s per TiB. Azure Managed Lustre provides 500 MB/s per TiB.
- **Provisioning Inefficiency:** To achieve 20 GB/s (approx. 20,480 MB/s) per node on AWS or GCP, a user must provision 20 TiB of Lustre per node. At a 64-rack scale, reaching the aggregate cluster throughput would require over-provisioning to 12.5 PB of capacity. This creates "phantom capacity" where users pay for millions of gigabytes of empty storage solely to secure the network bandwidth required for the GPUs.

High-Performance Object Tier

To minimize the time it takes to move data from persistent storage to the Lustre buffer (re-hydration), the hyperscaler model utilizes the highest-performance object tiers currently available.

Hyperscalers utilize a "low storage, high access" fee structure. While standard tiers may look inexpensive, the act of reading 600 TB daily during fine-tuning triggers significant retrieval fees and API request charges. Over 36 months, these transactional costs for high-epoch training become a massive OpEx driver that is often underestimated during initial budgeting.

CoreWeave AI Object Store Performance

- **Performance Scaling:** LOTA provides up to 7 GB/s of throughput per individual GPU, which exceeds the 20 GB/s requirement per 8 GPUs for fine-tuning and KV-cache rehydration.
- **Cost Avoidance:** There is zero line-item cost for a performance tier. 100% of the storage is consumed via the CoreWeave AI Object Hot tier at a transparent rate of $0.06/GB.
- **Usage-Based Automated Tiering:** CoreWeave uses real-time tracking to adjust billing across three levels—Hot ($0.06), Warm ($0.03), and Cold ($0.015)—based on the last access date. When a "Paused Project" resumes, reading the cold data automatically promotes it to the Hot rate for 7 days, making it performant immediately without retrieval "penalties."

For modern AI workloads, storage is a primary driver of both cost and performance, particularly due to high-throughput patterns such as KV-cache offload. CoreWeave avoids the overprovisioning and request-based cost structures of hyperscalers through its LOTA architecture and zero API/egress fees, enabling more efficient scaling and greater cost predictability.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Performance Analyst | Signal65

Russ Fellows
VP, Labs | 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

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION

Signal65 I signal65.com