Futurum RESEARCH

CoreWeave



Specialized AI Clouds Belong on the Shortlist for Multi-Year AI Commitments

OCTOBER 2026



Executive Summary

Enterprises are committing to multi-year AI infrastructure contracts before they have a clear picture of the options. The stakes are rising quickly: the largest US cloud providers plan to raise capital expenditure by nearly 80% in 2026, and power is now as scarce as capital. So whichever provider secures an enterprise's GPU capacity today is likely to host its AI workloads for years. Yet Futurum's buyer data show that most enterprises have barely evaluated specialized AI clouds competing with hyperscalers for that business.

A new Signal65 total cost of ownership (TCO) and tokenomics study gives buyers firmer numbers. Modeling a 5,000-GPU NVIDIA Blackwell deployment over three years, Signal65 found that CoreWeave cost less than three anonymized hyperscalers in six of seven configurations, with the gap widening once inference throughput is accounted for. We believe enterprises should price these commitments on cost per token rather than hourly GPU rates, and should place GPU-intensive work with a specialized cloud before today's contracts harden into defaults.

THE STAKES

~80%

planned capex increase by the largest US cloud providers in 2026

Power is now as scarce as capital

Most enterprises have barely evaluated specialized AI clouds

SIGNAL65 TCO STUDY

6 of 7

configurations where CoreWeave cost less than three anonymized

5,000 NVIDIA Blackwell GPUs · 3 years

Gap widens once inference throughput is counted

THE RECOMMENDATION

Cost per token

Price commitments on cost per token rather than hourly GPU rates.

Place GPU-intensive work with a specialized cloud

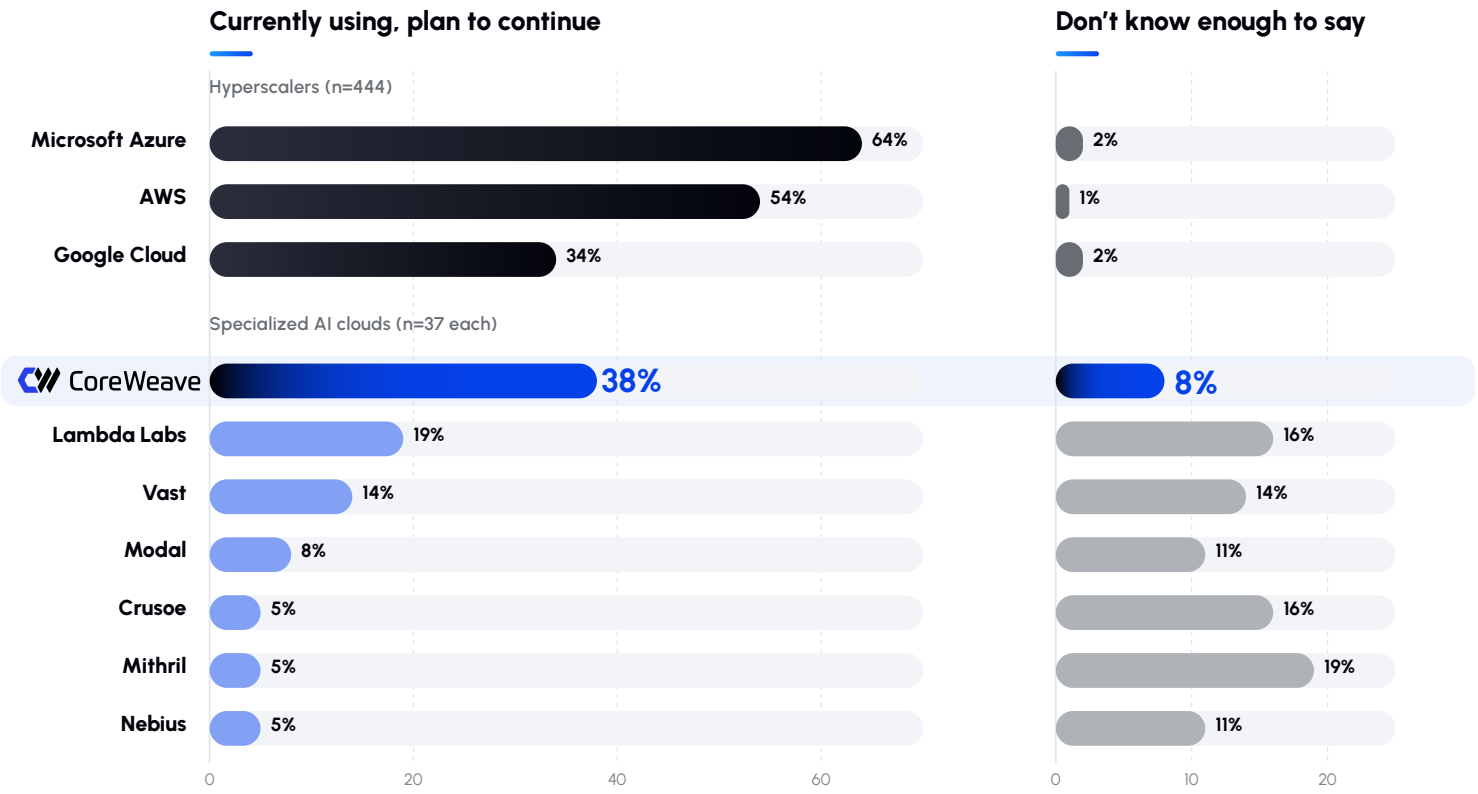
Act before today's contracts harden into defaults

Capacity Is Scarce, and Buyers Are Still Undecided

The scale of investment now underway raises the stakes of every hosting decision. Amazon, Alphabet, Microsoft, and Meta are together guiding well over $700 billion of capital expenditure in 2026, up from roughly $410 billion in 2025. Even that is not enough. Amazon said in July that its spending would still not buy enough capacity to meet demand this year, with constraints expected to run into 2027, and power and data center space, rather than chips, increasingly govern how quickly new capacity comes online. When capacity is this scarce, a three-year GPU reservation signed today shapes where an enterprise's AI workloads run well into the next hardware generation.

Buyers, however, are far less settled than suppliers. In Futurum's September 2026 AI Product Series survey of organizations building AI internally, 64% of respondents asked about Microsoft Azure and 54% of those asked about AWS currently using the platform for AI application development and planning to continue (n=444). For specialized AI clouds, the equivalent figures range from 5% to 38% (n=37 each), and 8% to 19% of respondents did not know enough about a given specialized provider to answer, compared with 1% to 2% for the largest hyperscalers (see Figure 1). CoreWeave is furthest along at 38%, up from 18% in March 2026, and its Net Score in Futurum's spending-intentions data rose from 19 in October 2025 to 34 in October 2026 (n=50). Most enterprises have yet to form a view of the category, at which point hard cost evidence counts most.

Figure 1. Current Use of Cloud Platforms for AI Application Development: Hyperscalers Versus Specialized AI Clouds

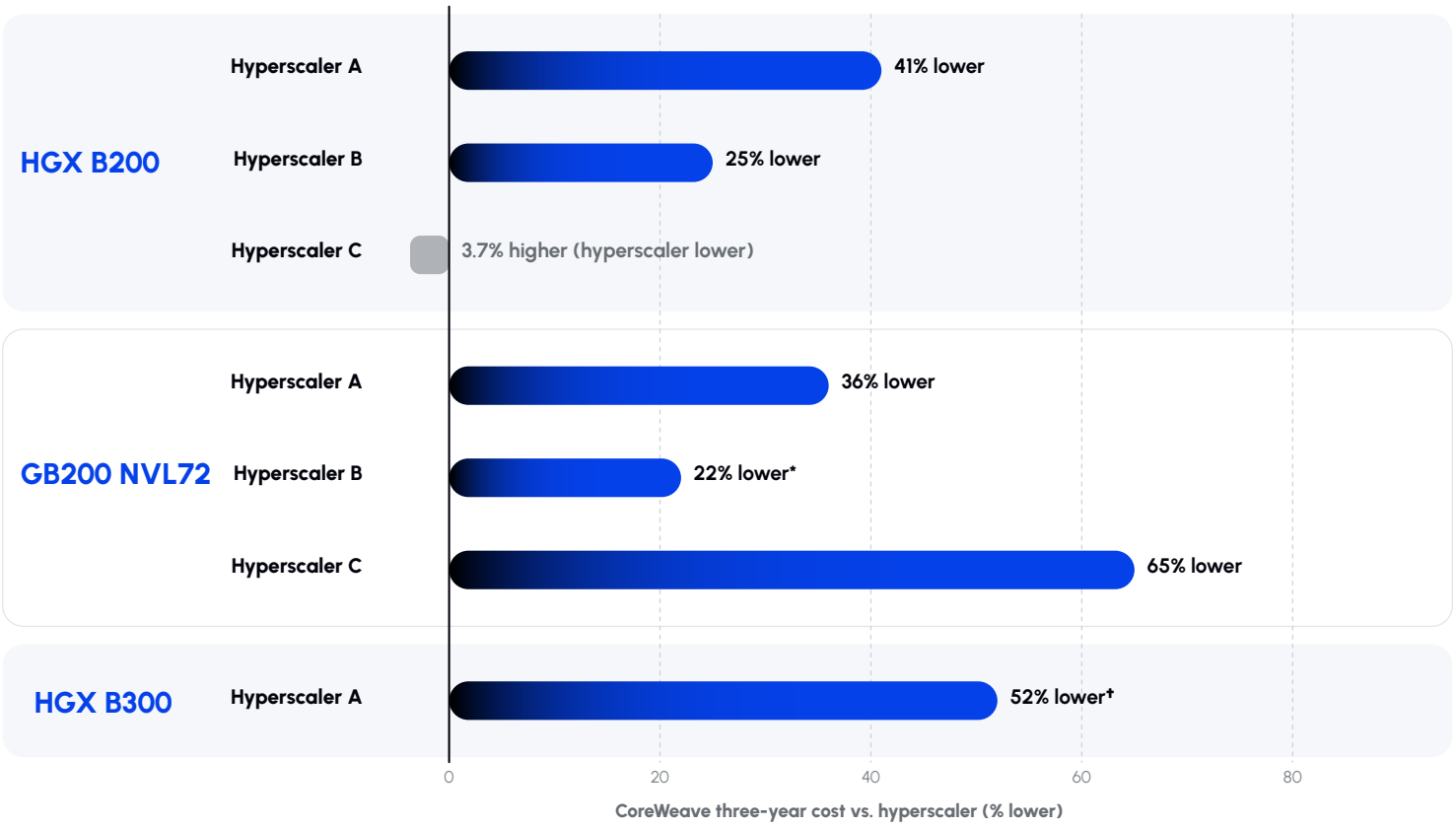


Note. Responses to "What best describes your company's plans with the following hosting and deployment platforms for its AI application development?" The left panel shows the share answering "Currently using, plan to continue to use"; the right panel shows the share answering "Unsure/Do not know enough to say." Bases are n = 444 for each hyperscaler and n = 37 for each specialized AI cloud. Source: Data from the Futurum AI Product Series, Theme 3 (Building AI Internally), September 2026 survey.

Hourly GPU Rates Understate the Real Cost Gap

Signal65's study, CoreWeave Delivers Higher Throughput and Lower Cost per Token on NVIDIA Blackwell: A Three-Year TCO and Tokenomics Analysis by Mitch Lewis, Performance Analyst at Signal65, models a 5,000-GPU NVIDIA Blackwell deployment over three years, including storage, networking, Kubernetes, observability, and support, at August 2026 public on-demand prices for CoreWeave and three anonymized hyperscalers. Signal65 found CoreWeave's three-year cost up to 41% lower on NVIDIA HGX B200, up to 65% lower on NVIDIA GB200 NVL72, and 52% lower on NVIDIA HGX B300, where only one hyperscaler publishes pricing (see Figure 2). The single exception was one hyperscaler's B200 configuration, which came in 3.7% below CoreWeave.

Figure 2. Modeled Three-Year Cost Advantage for CoreWeave by NVIDIA Blackwell Configuration



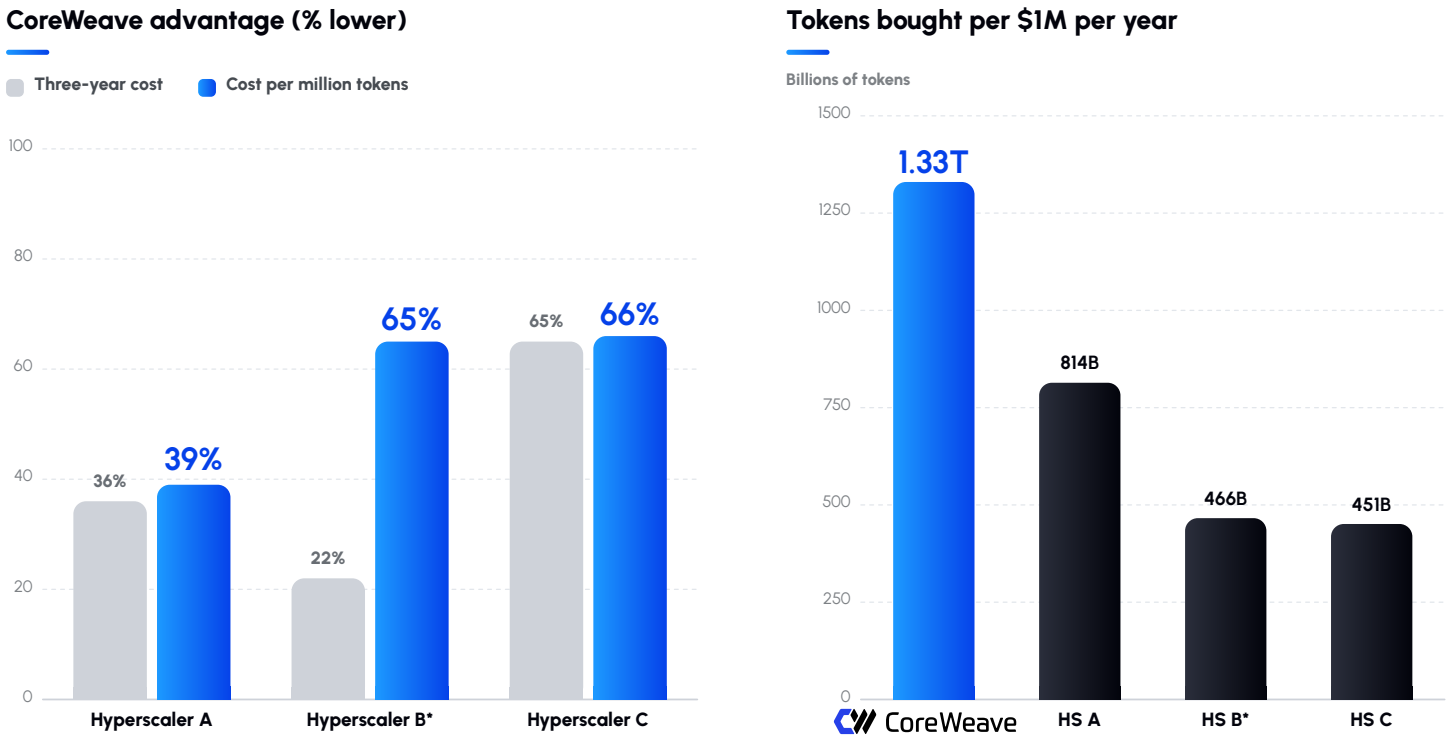
*Hyperscaler B GB200 rate estimated by Signal65. †CoreWeave B300 rate estimated by Signal65; only Hyperscaler A publishes B300 pricing.

Note. Bars show the percentage by which CoreWeave's modeled three-year cost is lower than each anonymized hyperscaler; the negative value marks the one configuration in which the hyperscaler is lower. The model assumes a 5,000-GPU deployment, 2 PB of object storage, and August 2026 public on-demand pricing. Hyperscaler B's GB200 NVL72 rate and CoreWeave's HGX B300 rate are Signal65 estimates; only Hyperscaler A publishes HGX B300 pricing. Source: Data from CoreWeave Delivers Higher Throughput and Lower Cost per Token on NVIDIA Blackwell: A Three-Year TCO and Tokenomics Analysis, by M. Lewis, 2026, Signal65.

The gap widens once throughput is factored in, because a provider's real cost per token depends on both how many tokens its infrastructure produces and its hourly rate. Using MLPerf Inference v6.0 results for DeepSeek-R1 on GB200 NVL72, Signal65 found CoreWeave delivered 4,868 tokens per second per GPU, against 4,668 for NVIDIA's reference submission and 2,178 for the only hyperscaler that submitted its own result. Against that hyperscaler, a 22% three-year cost advantage became a 65%

cost-per-million-tokens advantage, so that $1 million a year buys roughly 1.33 trillion tokens on CoreWeave versus 466 billion (see Figure 3). Signal65 also found CoreWeave's serverless inference pricing lower for 11 of 17 comparable models and identical for 4.

Figure 3. Three-Year Cost Advantage Versus Cost-per-Token Advantage for CoreWeave on NVIDIA GB200 NVL72



Note: The left panel compares CoreWeave's advantage on modeled three-year cost with its advantage on cost per million tokens; the right panel shows tokens processed per $1 million of annual spend. Throughput is per-GPU MLPerf Inference v6.0 (Closed Division, DeepSeek-R1, Server scenario); CoreWeave and Hyperscaler B use their own submissions, and Hyperscalers A and C use NVIDIA's reference submission. Cost per token assumes 80% GPU utilization. Hyperscaler B's GB200 NVL72 rate is a Signal65 estimate. Source: Data from Signal65 (Lewis, 2026).

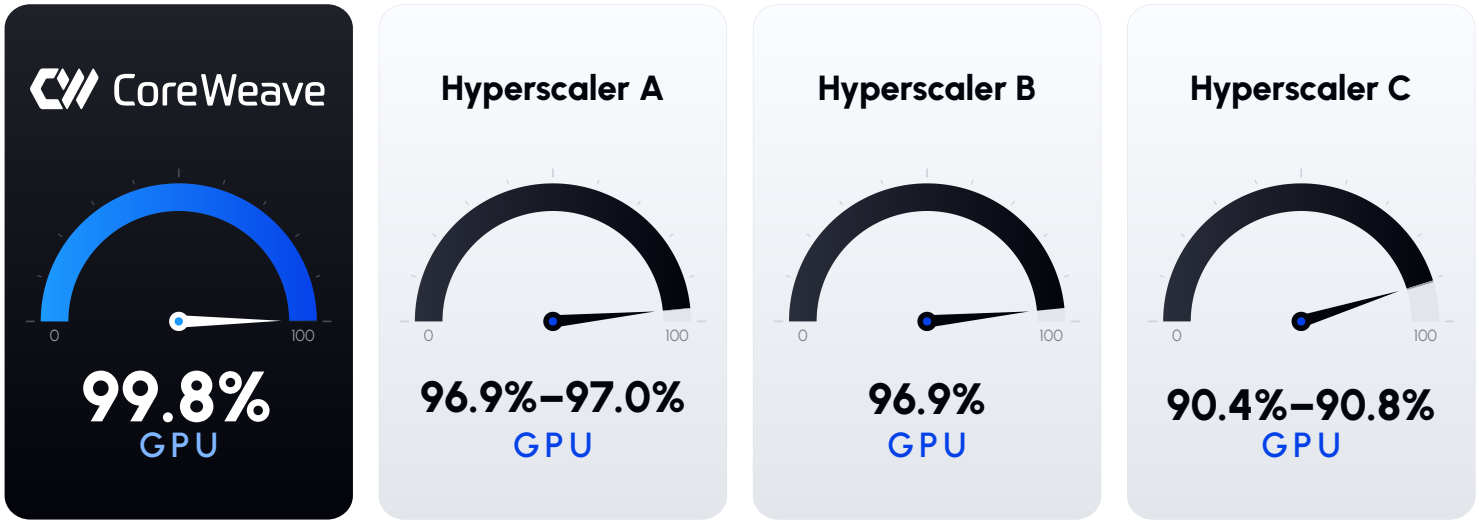
The model relies on list prices rather than negotiated discounts, assumes 80% GPU utilization throughout, and estimates a single hyperscaler's GB200 rate. It also compares CoreWeave only with hyperscalers, so it says nothing about how other specialized clouds would fare on the same basis. Its methodological point stands regardless: hourly GPU rates alone can badly understate the differences between providers.



CoreWeave's Fit

The Signal65 results describe what CoreWeave's specialization delivers in practice, namely lower GPU rates, no egress, observability, or support fees, and MLPerf throughput above NVIDIA's own reference. In Signal65's model, more than 99% of CoreWeave's three-year spend went to GPUs, against 90% to 97% at the hyperscalers (see Figure 4), so almost every dollar of the commitment buys compute. Signal65 also notes that CoreWeave was the first cloud provider to bring NVIDIA Vera Rubin NVL72 capacity online, which matters for contracts that will span a hardware transition.

Figure 4. Share of Modeled Three-Year Cost Allocated to GPUs, by Provider



Note: Ranges span the NVIDIA Blackwell configurations modeled for each provider. Non-GPU cost covers storage, networking, Kubernetes, observability, and support. Source: Data from Signal65 (Lewis, 2026).

CoreWeave's footprint across the wider enterprise base remains small; however, with Pervasion in Futurum's spending-intentions data below 4%, its fit is workload-specific. The case is strongest for large-scale training and high-volume inference, where GPU costs dominate the bill, and cost per token sets unit economics. For workloads tightly bound to a hyperscaler's data services, identity, and security tooling, the savings must be weighed against the integration effort.



Recommendations

Enterprises signing multi-year commitments should evaluate inference on cost per token and training on fully loaded three-year cost, using audited throughput data such as MLPerf, and should put more than one specialized provider alongside the hyperscalers in the comparison, requiring each to price the same workload model, including storage, egress, observability, and support.

They should also avoid committing their entire AI estate to a single provider class while the market takes shape. Keeping data-bound workloads with the incumbent hyperscaler while moving GPU-intensive training and production inference to a specialized cloud preserves bargaining power and gives the organization its own cost data at scale.

Our verdict is that specialized AI clouds now belong on every shortlist for GPU-intensive work, alongside the hyperscalers rather than in their place. Signal65's data makes a strong cost case for CoreWeave relative to the hyperscalers, and enterprises should hold every provider to the same cost-per-token test before signing, rather than discovering the gap at renewal.

Important Information About This Report

AUTHORS

Nick Patience
Vice President & Practice Lead,
AI Platforms | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group 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 "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

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

DISCLOSURES

The Futurum Group 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.

CoreWeave

ABOUT COREWEAVE

CoreWeave is a specialized AI cloud provider built from the ground up to power large-scale artificial intelligence, machine learning, and compute-intensive workloads. By delivering high-performance GPU infrastructure optimized for training and inference, CoreWeave provides enterprise organizations and AI innovators with cost-effective, high-throughput compute solutions. The company's purpose-built platform eliminates extra costs like egress and observability fees, ensuring maximum budget allocation directly toward high-performance infrastructure. As an industry pioneer, CoreWeave consistently brings next-generation accelerator platforms to market to support complex AI deployments across the enterprise ecosystem. To learn more about CoreWeave's dedicated AI cloud solutions and infrastructure performance, visit www.coreweave.com.

Futurum®

ABOUT THE FUTURUM GROUP

The Futurum Group 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.

Futurum®
CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337

© 2026 The Futurum Group. All rights reserved.