William Blair

AI INSIGHTS

Equity Research
Technology, Media, Communications | Infrastructure Software

September 30, 2026
Industry Report

Jason Ader, CFA
+1 617 235 7519
jader@williamblair.com

Arjan Kochar
+1 312 364 8864
akochar@williamblair.com

Frederick Gooding III
+1 312 551 4588
fgooding@williamblair.com

Rise of the Neoclouds:

From GPU Specialists to Long-Term Hyperscaler Challengers



This report is intended for AGraber@williamblair.com. Unauthorized distribution prohibited.

Please refer to important disclosures on pages 46-48. Analyst certification is on page 46.
William Blair or an affiliate does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. This report is not intended to provide personal investment advice. The opinions and recommendations herein do not take into account individual client circumstances, objectives, or needs and are not intended as recommendations of particular securities, financial instruments, or strategies to particular clients. The recipient of this report must make its own independent decisions regarding any securities or financial instruments mentioned herein.

Contents

Summary 3

Key Findings 5

What Is a Neocloud? 7

AI Market Dynamics 10

Neocloud Infrastructure Stack 16

Neocloud Customer Segments 18

Neocloud Approaches to Owning the Physical Infrastructure 21

Neocloud Contract Types 22

The Neocloud Business Model 24

Utilization Is the Key Driver of Neocloud Returns 28

Neocloud Competitive Landscape 30

Key Risks 42

Summary

The neocloud market has emerged as one of the clearest beneficiaries of the AI infrastructure boom, with specialized cloud providers such as CoreWeave, Nebius, IREN, Nscale, Crusoe, and Lambda using access to GPUs, power, data-center capacity, and project financing to bridge the gap between current levels of AI compute demand and supply. What began as a response to acute capacity shortages is quickly evolving into a distinct cloud infrastructure category, with the largest neoclouds increasingly positioning themselves as long-term alternatives to general-purpose hyperscalers.

We estimate global neocloud revenues will grow from roughly $25 billion in 2026 to $200 billion-$300 billion by 2031, with a base case of about $250 billion, fueled by continued demand for inference, agentic workloads, enterprise AI, hyperscaler offload, and sovereign compute. This forecast assumes roughly 20 GW of powered and online neocloud capacity by 2031, versus more than 18.5 GW of disclosed contracted or secured power capacity already in place across top neoclouds, and an effective revenue yield of about $12.5 million per active MW. Four of the top neoclouds—CoreWeave, Nebius, IREN, and Nscale—have already amassed roughly $275 billion of customer commitments, while long-term contracts and prepayments are increasingly helping finance the infrastructure required to fulfill them.

At the infrastructure layer, neoclouds rely largely on standardized Nvidia hardware, limiting product-level differentiation. Competitive advantage instead comes from power and GPU access, cost of capital, deployment speed and reliability, customer relationships, and up-the-stack software capabilities. These advantages may prove less durable as AI computing supply expands.

The business model also carries meaningful risk. Neoclouds are capital-intensive businesses built around assets subject to technological obsolescence and uncertain residual value, concentrated customer bases, and large fixed debt and lease obligations, with economics that depend heavily on sustained AI demand, utilization, and pricing. If any of those variables weaken, leverage and negative free cash flow could amplify the downside—particularly as hyperscalers add their own purpose-built AI capacity and become less dependent on third-party providers for overflow capacity.

Notwithstanding the numerous risks, our core view is that neoclouds are more than temporary beneficiaries of GPU scarcity, but that many subscale players will likely be consolidated or squeezed out as supply normalizes. The best-positioned players are using today's supply imbalance to secure customers, capital, power, and scale while broadening into managed inference, storage, security, databases, and other core cloud services. Over time, this should allow a handful of scaled neoclouds to become durable AI infrastructure platforms—serving not just hyperscalers, frontier labs, and AI-native start-ups, but also many enterprises and government entities.

With this report, we aim to provide investors with a framework for understanding the rise of neoclouds, the economics of the neocloud business model, the competitive positioning of key public and private players, and the principal risks that could determine whether neoclouds become long-term AI infrastructure winners or prove to be cyclical beneficiaries of a temporary capacity shortage—and potentially among the first to come under pressure if the current investment cycle ultimately proves unsustainable.

From a stock perspective, we believe that CoreWeave and Nebius are the best-positioned players as both have demonstrated their expertise in securing power and data center capacity, financing that capacity efficiently, and delivering large-scale compute to the world's most demanding customers. Looking ahead, both are also moving up the stack and positioning themselves to become durable challengers to the hyperscalers for AI-centric workloads.

Our overarching conclusions are:

- ***Neoclouds are filling a structural AI capacity gap.*** Exploding compute demand, constrained power and semiconductor availability, and longer data-center development timelines have created a multiyear opening for specialized AI infrastructure providers.

- ***The market opportunity is large and scaling quickly.*** We estimate the global neocloud market will grow tenfold from roughly $25 billion in 2026 to about $250 billion in 2031, with neoclouds capturing a rising share of AI-centric cloud revenue driven by training, inference, agentic, and sovereign compute demand.

- ***Sources of differentiation exist, but durable moats remain questionable.*** Because neoclouds rely largely on standardized Nvidia hardware, product-level differentiation is limited. Long-term winners will need to turn current infrastructure advantages into more defensible differentiation through scale, software, and deeper customer integration.

- ***Backlog and customer prepayments are becoming a key financing engine.*** Large, long-duration commitments from hyperscalers and AI labs are underwriting the neocloud buildout by supporting project-level debt and reducing upfront equity requirements, with customer prepayments covering a growing portion of capex in recent contracts. Growing lender acceptance of contracted AI infrastructure has also broadened access to institutional capital and lowered financing costs for the strongest projects.

- ***The largest neoclouds are moving up the stack.*** Leading providers are expanding beyond bare metal GPU access into storage, networking, orchestration, security, and managed inference services, allowing them to capture more aggregate spending and support a broader set of customers and AI workloads.

- ***Current economics look attractive, but leverage magnifies the downside risk.*** Shrinking payback periods, rising contribution margins, resilient GPU pricing, and customer prepayments support attractive returns today. However, those returns depend on high utilization, strong revenue per MW, durable pricing, long hardware lives, and timely capacity delivery. Because depreciation, debt service, and much of the cost base are fixed—and rising interest rates potentially increase the cost of financing—even modest shortfalls in pricing, utilization, or deployment timing could disproportionately pressure EBIT, free cash flow, and ROIC.

- ***Consolidation is likely as the market matures.*** As hyperscalers add more of their own AI capacity and supply constraints ease, subscale neoclouds may struggle to sustain attractive utilization, pricing, and funding. We expect the market to consolidate around a small number of well-capitalized neocloud platforms with access to power and capital, broad customer bases, proven execution at scale, and product-level differentiation.

- ***Neocloud stocks are not for the faint of heart.*** Neoclouds, in many ways, are at the "tip of the spear" of the AI infrastructure buildout, operating in one of the fastest-moving and most capital-intensive markets in technology history. It is difficult to have high conviction in any long-term outcome given how quickly demand, supply, technology, pricing, and financing conditions are evolving, and numerous things still need to go right. Even so, the long-term opportunity is massive, and the near-term demand signals continue to point toward sustained AI infrastructure investment, tight capacity, and growing customer willingness to commit capital well ahead of deployment.

- ***CoreWeave and Nebius are first movers.*** We are initiating coverage of CoreWeave and Nebius with Outperform ratings as we believe both have proved they can rapidly secure power and data center capacity, efficiently raise substantial amounts of capital, and deliver large-scale

compute to the world's most demanding customers. The next step is moving up the stack, expanding their software and services capabilities, and establishing themselves as durable challengers to the hyperscalers for AI-centric workloads.

Key Findings

Right Place, Right Time
The confluence of exploding AI compute demand, constrained GPU availability, and access to power/capital has fueled the rise of a new generation of specialized, AI-focused cloud providers (aka "neoclouds"), which includes publicly traded companies such as CoreWeave, Nebius, and IREN, along with private players such as Nscale (recently filed for IPO), Crusoe, and Lambda. We estimate that the global neocloud market, driven by the transition from experimental AI deployments to production-scale training and inference workloads, will grow tenfold from roughly $25 billion in 2026 to $250 billion by 2031.

Insatiable AI Demand, Constrained Supply
AI compute demand is accelerating across multiple fronts—including larger and more capable models, increasingly compute-intensive agentic workloads, rising inference volumes, broadening enterprise deployment, and physical AI—and continues to outpace global supply. We expect this imbalance to persist for several years given constraints around power, advanced chips, and skilled labor, as well as growing political and community resistance that could slow data center approvals and construction.

Bridging the AI Capacity Gap
Neoclouds have stepped into the computing supply vacuum by offering GPU-centric data center capacity for frontier AI labs, hyperscalers, AI-native start-ups, government entities, and select large enterprises. Five of the top six leading neoclouds have already secured over 18.5 GW of power capacity (Lambda has not disclosed any of its contracted capacity commitments), while only a fraction is online and supporting AI workloads today, highlighting the enormous infrastructure buildout still ahead.

Leveraging GPU Expertise
Neoclouds have carved out a distinct position in the cloud services market by building infrastructure specifically tailored to high-performance, GPU-dense computing for AI training and inference, versus the broader mix of general-purpose workloads handled by traditional hyperscalers. Several early neoclouds also drew on their crypto-mining roots, bringing expertise in securing low-cost power, operating high-density facilities, and managing large-scale compute infrastructure—capabilities that enabled them to pivot rapidly toward AI.

Aggregate Neocloud Backlog Is More Than $275 Billion
Four of the top neoclouds have amassed roughly $275 billion of combined customer commitments, including large contracts signed early in the third quarter. This massive backlog reflects a willingness among hyperscalers and AI labs—and increasingly well-funded AI-native start-ups—to lock up scarce compute capacity years in advance, effectively underwriting the neoclouds' rapid infrastructure buildout (see our next finding).

Backlog and Prepayments as Financing Engine
Neoclouds have turned their massive backlogs into a powerful financing mechanism: long-dated hyperscaler and AI-lab commitments provide the cash-flow visibility needed to raise project-level debt against the GPUs and infrastructure required to fulfill those contracts, reducing the upfront equity burden of scaling up capacity. For example, CoreWeave's cost of secured capital has fallen

dramatically as its business has matured, with spreads on major contract-backed financings compressing from nearly 1,000 basis points over SOFR in 2023 to just 225 basis points on its $8.5 billion DDTL financing in 2026. In addition, all the major neoclouds are using customer prepayments to fund a substantial portion of the associated capex.

Attractive Businesses Today—With Shrinking Payback Periods
In the most recent quarter, Nebius management disclosed that its four $1 billion-plus deals signed in the second quarter are expected to generate $20 million to $25 million of annual contract value (ACV) per megawatt, with upfront customer payments covering 50% to 60% of associated capex. Those economics have helped shorten the company's expected payback period on new deployments from a historical 2-3 years to just 1 year and 10 months.

GPU Pricing Proving Stickier Than Expected
Contrary to expectations that rapid product cycles would quickly erode the value of older GPUs, prior-generation chips are retaining both their economic usefulness and pricing power longer than anticipated. Nvidia recently noted that pricing and margins for Blackwell and next-generation Vera Rubin GPUs are reaching new highs, while pricing for older-generation chips remains at or above levels seen several years ago. Of note, CoreWeave raised prices by 25% in July in response to strong demand, while Nebius disclosed that its first compute-capacity auction in the second quarter cleared Blackwell pricing roughly 15% above the highest level it had previously charged.

Eyes on the Hyperscaler Prize
Recognizing the risks of customer concentration and an overly narrow service offering, neoclouds are broadening their capabilities beyond GPU rental while diversifying their customer bases beyond the large AI labs and hyperscalers. Over time, this will position neoclouds to compete more directly with the hyperscalers that today represent some of their largest customers. Ironically, by turning to neoclouds to alleviate near-term capacity constraints, hyperscalers such as Microsoft may effectively be financing the development of future competitors.

Not All Neoclouds Are Created Equal; Market Consolidation Inevitable
As competition heats up and supply ultimately catches up to demand, not all neoclouds will survive as standalone companies. The enormous capital requirements, scarcity of power, benefits of less expensive financing, and growing need for a full cloud software stack all favor scale economies, likely leaving subscale providers increasingly disadvantaged. We expect the largest neoclouds to use M&A activity both horizontally to acquire customers and capacity and vertically to secure power, data centers, and software capabilities.

Credit Is Central to the Buildout, but Balance Sheet Risk Varies
CoreWeave, Nebius, and IREN carry roughly $29.5 billion, $0.5 billion, and $1.5 billion of net debt, respectively, equal to respective net leverage ratios of approximately 6.7x, 1.3x, and 6.2x (debt/TTM adjusted EBITDA). CoreWeave is the most aggressively leveraged, Nebius remains relatively conservatively financed given its large cash balance and focus on customer prepayments, and IREN sits in between while still transitioning from Bitcoin mining toward AI cloud. Rapid EBITDA growth can drive deleveraging, but underutilized capacity could leave operators servicing substantial fixed obligations against rapidly depreciating assets.

Nvidia Is an Important Strategic Investor Across the Neocloud Ecosystem
Nvidia has taken equity positions across many of the leading public and private neoclouds—including CoreWeave, Nebius, IREN, Nscale, Lambda, Crusoe, and Firmus—providing growth capital to companies that in turn deploy billions of dollars of Nvidia infrastructure. Supporting the neoclouds also helps Nvidia diversify its revenue across more cloud customers, boosting its pricing

power and reducing its dependence on the major hyperscalers. Still, the growing overlap between supplier, investor, and customer relationships adds to concerns about the circularity of AI infrastructure spending.

Biggest Risk: The Model Is Highly Sensitive to Utilization and Pricing
Neoclouds would likely be among the first and hardest hit in an AI infrastructure downturn given their large upfront commitments to GPUs, data centers, and power, often financed with debt, and predicated on sustained long-term demand. While take-or-pay structures provide meaningful near-term revenue protection, downside could be amplified by heavy customer concentration, negative free cash flow, fixed debt obligations, circular financing, and potentially depreciating GPU assets if future demand, renewal activity, utilization, or pricing weakens.

Neocloud Stocks Are Heavily Shorted
Despite improving fundamentals across the neocloud group, investors retain a significant degree of skepticism toward the category, driven by high stock-price volatility, uncertain valuation frameworks, capital intensity, high debt levels, customer concentration, negative free cash flow, and concerns about normalized GPU economics. A key short thesis is that today's scarcity-driven economics will normalize as hyperscalers add capacity and AI infrastructure supply expands. Importantly, some reported short interest is likely mechanical, reflecting convertible-arbitrage strategies in which investors buy neocloud convertible debt while shorting the underlying shares. As of July 31, 2026, CRWV, NBIS, and IREN carry short positions of 18%, 24%, and 29%, respectively.

Initiating Coverage of CoreWeave and Nebius
We are initiating coverage of CoreWeave and Nebius with Outperform ratings. CoreWeave offers unmatched scale, backlog, and demonstrated financing capability, while Nebius combines a stronger balance sheet with a differentiated cloud architecture and greater near-term upside from software and higher-layer AI services. In both cases, we believe current investor skepticism understates the earnings power that could emerge as contracted capacity comes online, utilization matures, and returns on invested capital improve.

What Is a Neocloud?

We define a *neocloud* as a specialized cloud infrastructure provider whose core business is delivering large-scale accelerated compute environments optimized for AI workloads, rather than a broad portfolio of general-purpose cloud services. Under this definition, a neocloud must secure access to power, commit capital to GPUs and data-center capacity, and exercise meaningful control over the underlying physical infrastructure (even if it does not own the data center facilities itself). Notably, some companies that fit this definition, including CoreWeave and Nebius, do not use the term "neocloud," instead positioning themselves as "AI hyperscalers" or "AI cloud providers."

Neoclouds Versus Hyperscalers
Several characteristics distinguish neoclouds from traditional hyperscalers. Most importantly, neocloud infrastructure is purpose-built around accelerated compute, with large, high-density GPU clusters at the center of their platforms. Traditional hyperscale data centers were built to support a broad range of workloads and shift capacity dynamically as demand changed, whereas AI workloads carry more rigid constraints—particularly around parallelism, locality, and compute density. Neoclouds represent an architectural response to those constraints.

Neocloud service portfolios are also considerably narrower than those of the hyperscalers. AWS, Azure, and Google Cloud offer broad product menus spanning databases, storage, analytics, application development, identity, security, and hundreds of other platform services. Neoclouds, by

contrast, remain primarily focused on accelerated computing and the software, networking, and orchestration required to train and run AI models at scale. Historically, customers using neoclouds for GPU compute have had to supply much of the surrounding infrastructure themselves, including databases, security, and orchestration. That is beginning to change as providers such as CoreWeave and Nebius broaden their platforms and move further up the stack (see page 14).

The neoclouds' narrower service portfolios and infrastructure optimized specifically for AI workloads have also enabled them to offer GPU capacity at substantial discounts to traditional hyperscalers, with H100 rental rates often 40%-60% below comparable hyperscaler on-demand pricing. According to data intelligence platform Silicon Data, neocloud H100 pricing averaged roughly $2.53 per GPU-hour in spring 2026, versus about $7.40-$7.55 per GPU-hour for hyperscalers, implying a discount of roughly 65%.

That advantage remains visible with Nvidia's latest Blackwell-generation accelerators. Nebius currently lists B200 capacity at $7.15 per GPU-hour and B300s at $7.85, while CoreWeave lists B200 capacity at roughly $8.60 per GPU-hour. By comparison, AWS currently prices B200 Capacity Blocks at roughly $12.36 per GPU-hour and B300s at $14.04, implying a neocloud discount of roughly 30%-45% on comparable current-generation hardware.

Speed and access to scarce capacity are also central to the neocloud model. Many neoclouds emerged as AI demand began outstripping the ability of hyperscalers to bring new GPU capacity online, creating an opportunity for specialized providers that could secure power, procure GPUs, and deploy high-density infrastructure more quickly. In this sense, neoclouds initially filled a capacity gap, offering customers faster access to large-scale AI infrastructure than they could get from traditional cloud providers.

The line between neoclouds and other specialty cloud providers is not always clear. DigitalOcean, Vultr, and Akamai Cloud, for example, all offer GPU capacity and compete for some of the same workloads, but their core businesses are more firmly rooted in broader developer-focused cloud platforms rather than purpose-built AI infrastructure. We therefore consider them adjacent competitors, not true neoclouds.

Competition is also beginning to emerge from a new category of provider: large technology companies that have built substantial AI infrastructure primarily for their own workloads but are increasingly willing to monetize excess capacity externally. Meta, for example, is developing a cloud infrastructure business that would sell third-party access to its AI compute and models, while SpaceXAI has begun leasing portions of its AI infrastructure to external customers and has indicated that it expects to enter additional compute-services agreements.

While these businesses are not neoclouds in the traditional sense, their enormous infrastructure footprints, access to capital, and ability to treat third-party cloud revenue as incremental monetization of capacity could create another source of competition for AI workloads. More broadly, this development suggests that the boundaries between captive AI infrastructure, hyperscale cloud, and specialized GPU cloud are becoming increasingly blurred.

Neoclouds Versus AI Inference Providers

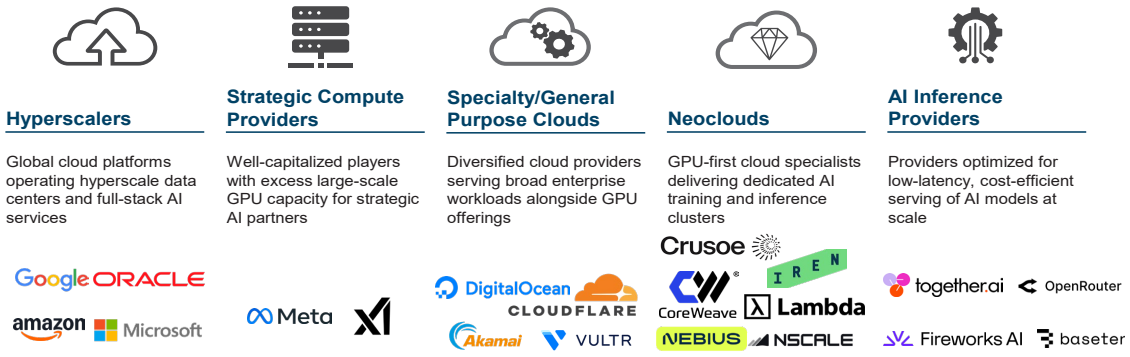
The term "neocloud" is increasingly being applied to a wide range of AI infrastructure companies, but we believe an important distinction should be made between asset-heavy neoclouds and asset-light AI inference providers. In our narrower definition, neoclouds such as CoreWeave, Nebius, IREN, Lambda, Nscale, and Crusoe make long-term commitments to power, GPUs, and data-center capacity and assume much of the capital and utilization risk associated with those assets. The

data centers themselves do not necessarily need to be owned—CoreWeave, for example, operates extensively through leased and colocation facilities—but these companies control the underlying compute capacity and commit substantial capital to building it out.

In contrast, dedicated AI inference providers—including Together AI, Fireworks AI, OpenRouter, and Baseten—sit further up the stack. These companies primarily differentiate through inference software, model serving, orchestration, and GPU and kernel-layer optimization, while sourcing their underlying compute capacity from hyperscalers, neoclouds, and other infrastructure partners.

Baseten, for example, aggregates GPU capacity across 20-plus clouds and regions into a fungible pool, while Fireworks AI relies on third-party cloud providers for its underlying infrastructure. Together AI is somewhat more infrastructure-intensive, offering dedicated GPU clusters alongside inference and training services, but its differentiation still sits primarily in the software and optimization layers rather than in owning power or data-center assets. At the most asset-light end is OpenRouter, which functions more like an inference marketplace and routing layer, directing requests across more than 70 providers based on price, latency, throughput, and availability rather than supplying the compute itself.

Exhibit 1
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Neocloud Competitive Landscape



Source: William Blair Equity Research

Market Sizing

Neoclouds are emerging as a distinct and rapidly scaling category of cloud infrastructure provider, specializing in GPU-as-a-service (GPUaaS), generative AI platforms, and high-density data center capacity. We estimate global neocloud revenues will grow from roughly $25 billion in 2026 to $200 billion-$300 billion by 2031, with a base case of about $250 billion, fueled by continued demand for inference, agentic workloads, enterprise AI, hyperscaler outsourcing, and sovereign compute.

This forecast assumes roughly 20 GW of powered and online neocloud capacity by 2031 (versus the more than 18.5 GW currently under contract with five of the six largest neoclouds) and an effective revenue yield of about $12.5 million per MW, well below the $20 million-$25 million per MW being achieved on some recent contracts. This per-MW assumption allows for lower utilization and some normalization in GPU pricing as supply expands. At the high end, if the industry can sustain revenue closer to $15 million per MW, the same 20 GW footprint would support roughly $300 billion of annual revenue.

The biggest risk to this forecast is likely power rather than demand (see next section). U.S. data center electricity consumption is expected to more than double by 2030, while major power markets are already showing signs that new dispatchable generation is not keeping pace with load

growth. For example, PJM Interconnection (regional power-grid operator for a large part of the Mid-Atlantic and Midwest) recently fell roughly 6.8 GW short of its reliability requirement, meaning that it did not secure enough dependable power supply to meet its own target for reliably serving expected peak demand plus a safety margin. As a result, PJM is considering measures that could require new data centers to provide or secure their own electricity generation.

Bringing 20 GW of neocloud capacity online will therefore require not only data-center construction, but substantial additions of gas, nuclear, and renewables, plus storage, behind-the-meter generation, and transmission infrastructure. If sufficient power cannot be developed on time, capacity additions—and therefore the revenue opportunity—could fall materially below our base case. Globally, the IEA expects data center electricity consumption to roughly double to 945 TWh by 2030, underscoring the scale of the required power buildout.

AI Market Dynamics

Physical Constraints: Power Grid Bottlenecks and Local Pushback

Neoclouds face the same physical constraints limiting AI infrastructure growth across the broader market, with power availability—not land or chips—emerging as the primary bottleneck. Supplying the enormous electricity loads required by dense GPU clusters is increasingly difficult within existing transmission networks, materially extending development timelines. U.S. projects completed in 2025 took a median of 61 months from initial interconnection request to commercial operation, up from 36 months in 2015.

According to our colleague, William Blair analyst Jed Dorsheimer, planned U.S. AI data-center capacity reached 109 GW by mid-2026. That surge in demand is projected to create a cumulative 28 GW shortfall between dispatchable power supply and data-center load by 2030. Equipment shortages are compounding the problem, with lead times for critical grid components such as transformers rising from roughly 50 weeks in 2021 to more than 160 weeks in 2026.

At the same time, the environmental and grid impact of high-density data centers is generating increasing resistance at the state and local levels, turning community opposition from a permitting hassle into a meaningful project risk. Policy responses have broadened from local zoning disputes to statewide moratoria, tighter permitting standards, grid-interconnection controls, and reduced tax incentives.

In July 2026, New York imposed a one-year moratorium on new or expanded data centers consuming at least 50 MW that require discretionary state permits while it develops a statewide environmental framework. Pennsylvania followed in August with binding requirements that require developers to secure local approval, bear the cost of incremental energy infrastructure, meet water and environmental standards, and comply with community-engagement requirements; the state also removed data centers from its fast-track permitting program and conditioned existing tax incentives on compliance.

Meanwhile, Illinois stopped accepting new applications for its data-center incentive program in July, while Texas has temporarily prevented data-center projects from advancing through ERCOT's interconnection process pending statewide verification and audit. Not to be outdone, Virginia has enacted new ratepayer protections and a first-of-its-kind tax on data-center electricity consumption.

According to Data Center Watch, at least $156 billion of planned U.S. data-center investment was blocked or delayed amid opposition in 2025, with another roughly $130 billion affected in the first quarter of 2026 alone. Collectively, these developments suggest that community acceptance, permitting, and grid access are becoming increasingly important constraints on the pace and location of new capacity.

Shifting AI Architectures and Use-Cases: From Training to Inference

AI workloads are undergoing a major transition from large-scale model training to production-level inference. AMD CEO Lisa Su recently predicted that 2026 will be the first year in which inference consumes more compute than training, accounting for roughly 60% of global AI compute capacity. Unlike training, which is typically concentrated in large centralized clusters, inference increasingly requires distributed, low-latency infrastructure closer to end-users. The rise of agentic AI is accelerating this shift, as autonomous systems execute multistep workflows, invoke tools, retrieve data, and generate longer reasoning chains.

Agentic workloads are also changing the mix of compute inside the data center. Traditional AI clusters often paired one CPU with four to eight GPUs, but increasingly complex agentic workflows require substantially more CPU capacity for orchestration, code execution, API calls, and other serial-processing tasks. As a result, some architectures are moving toward a much higher CPU-to-GPU ratio, potentially approaching 1:1 for certain workloads. This shift could expand the data-center CPU TAM to roughly $120 billion-$200 billion by 2030, with companies such as Arm expected to benefit from growing demand for high-core-count host processors.

At the same time, reasoning-intensive workloads are placing greater pressure on memory capacity and bandwidth. Multistep inference generates large key-value (KV) caches and requires frequent movement of data between processors and memory, increasing the risk that expensive GPUs sit idle while waiting for data. This is driving demand for higher-bandwidth memory, disaggregated memory architectures, and advanced optical interconnects designed to keep compute resources continuously in use.

For neoclouds, the shift toward inference should make AI demand more persistent and broaden the customer opportunity, but it also shifts the competitive advantage from simply securing GPUs and power toward operating a distributed, heterogeneous AI infrastructure stack spanning GPUs, CPUs, memory, networking, and software.

AI Efficiency Gains May Drive More Compute Demand, Not Less

Rapid improvements in model efficiency initially raised concerns that falling cost per token would reduce infrastructure requirements. There is growing evidence, however, that as models become less expensive to run, developers will use more reasoning steps, more agent loops, and more model calls and will pursue use-cases that were previously uneconomical. This could create a Jevons-paradox effect, where lower inference costs ultimately drive greater overall compute consumption. For neoclouds, that would mean more advanced chips and more efficient models may actually expand the AI infrastructure market rather than shrink it.

GPU Pricing Has Been Stickier Than Expected

One of the original bear cases for neocloud economics was that rapid improvements in accelerator performance would drive sharp price deflation for prior-generation GPUs, shortening their economic lives and pressuring returns. So far, the opposite has happened. SemiAnalysis's H100 rental index shows one-year contract pricing rebounding nearly 40%, from $1.70 per GPU-hour in October 2025 to $2.35 in March 2026, while CoreWeave raised pricing by roughly 25% across multiple GPU SKUs in July amid continued capacity tightness.

Importantly, that pricing strength is also extending to older hardware. Nebius said second-quarter pricing on prior-generation GPUs increased more than 30% sequentially, while CoreWeave recently signed a new contract for A100s—the architecture launched in 2020—extending through 2029 at what management described as an attractive price.

The reason appears to be that supply constraints have migrated beyond the GPU itself to power, data-center capacity, memory, and other infrastructure, while rapidly growing inference and agentic workloads can efficiently absorb older GPUs that are no longer optimal for frontier training. New GPU generations also require substantially denser power and cooling infrastructure, limiting how quickly they can displace installed fleets.

For neoclouds, this matters enormously: slower-than-expected GPU price depreciation extends useful asset lives, supports residual values and utilization, shortens effective payback periods, and materially increases lifetime returns on deployed capital. The key question is whether this pricing durability persists once power and infrastructure supply eventually catch up with demand.

Sovereign Compute and Decentralized Workloads Expand the Addressable Market
Growing concerns about data residency, regulatory control, and strategic dependence on foreign cloud providers are driving governments and regulated industries to build AI capacity within their own jurisdictions, creating a meaningful opportunity for neoclouds. Unlike the highly centralized models of most major hyperscalers, neoclouds can often deploy smaller regional clusters, partner with local infrastructure providers, and tailor environments around specific sovereignty, compliance, and security requirements.

Nscale, for example, is building Stargate UK and Stargate Norway with an explicit focus on sovereign workloads, while Crusoe's new Edge Zones are designed to bring low-latency, sovereign AI capacity into targeted geographies. This opportunity should grow as inference becomes more distributed and workloads increasingly need to run closer to users, data sources, and regulated environments. For neoclouds, sovereign and decentralized compute therefore represents both an incremental source of demand and a potential point of differentiation versus hyperscalers—particularly for operators that can combine leading-edge GPUs with flexible deployment models, local infrastructure partnerships, and jurisdiction-specific governance.

The Accelerator Market Is Becoming More Heterogeneous
Nvidia GPUs remain the dominant chips powering AI infrastructure, but alternatives are becoming increasingly credible for large-scale workloads. AMD is gaining traction with its Instinct accelerator roadmap, with the MI350 series targeting both training and inference and the next-generation MI400 family beginning to roll out in late 2026. AMD has also emphasized a more open software ecosystem through ROCm, giving cloud providers another merchant-silicon option alongside Nvidia.

On the custom silicon side, AWS is pushing Trainium and Inferentia, Google is expanding its TPU portfolio with separate architectures optimized for training and inference, and Microsoft is making progress with its Maia chips. Google's newest TPU 8i, for example, is specifically optimized for low-latency inference and claims 80% better performance per dollar than its prior generation, while AWS positions Trainium around full-stack training and inference economics.

This creates both a risk and an opportunity for neoclouds. The risk is that supplier diversity in the accelerator market could reduce the scarcity premium on Nvidia GPUs, historically an important advantage of neoclouds. The opportunity lies in supporting multiple chip architectures, which could become a new source of differentiation for neoclouds.

Access to Capital Is Emerging as a Competitive Advantage
The large upfront cost of deploying AI infrastructure means capital availability is increasingly as important as technical capability to neocloud success. For example, CoreWeave projects $35 billion-$39 billion of 2026 capex and has funded its buildout primarily with asset-level debt backed by take-or-pay customer contracts, supplemented by unsecured debt, convertibles, and equity.

Importantly, its cost of financing has declined materially as the business has scaled. While CoreWeave's corporate credit remains below investment grade at B+ with a positive outlook from S&P, its latest $8.5 billion delayed draw term loan (DDTL) facility (priced at SOFR + 225 basis points) received an investment-grade rating of A3 from Moody's because the financing is secured by GPU infrastructure and an associated customer contract. This illustrates an important evolution in the neocloud model: strong hyperscaler contracts can effectively transform otherwise sub-investment-grade corporate credit into investment-grade project-level financing.

Nebius is pursuing a similar path. The company expects $20 billion-$25 billion of 2026 capex and is increasingly using customer prepayments and contract-backed financing to fund expansion. In July, Nebius completed its first senior secured asset-level financing, a $775 million facility backed by deployed GPUs and contracted cash flows from an investment-grade customer, priced at SOFR + 250 basis points. Management views the structure as repeatable against more than $40 billion of additional contracted customer commitments. Nebius subsequently raised $5.75 billion of convertible debt in August, including $3.45 billion of 2030 notes carrying only a 0.5% coupon, further demonstrating its rapidly expanding access to institutional capital.

As the industry scales, operators able to convert long-duration investment-grade customer contracts into low-cost asset-level financing—while also tapping customer prepayments, infrastructure partners, and public capital markets—should be able to secure power and deploy capacity faster, potentially creating a meaningful financing moat around the largest players.

Hyperscalers Are Important Customers as Well as Competitors to Neoclouds
One of the most important developments in the neocloud market has been the willingness—and, increasingly, necessity—of hyperscalers to outsource portions of their AI infrastructure demand to companies that could ultimately compete with them. Microsoft, Meta, and other large technology companies have signed multibillion-dollar capacity agreements with CoreWeave, Nebius, Nscale, Crusoe, and IREN, validating the neocloud value proposition around speed of deployment, GPU access, and specialized infrastructure.

Nebius, for example, has signed major agreements with both Microsoft and Meta, including a new Meta commitment of up to $27 billion, while CoreWeave's backlog has surpassed $100 billion across hyperscalers, frontier labs, and other customers. While this allows neoclouds to access a major source of long-term compute demand, it also creates an unusual "coopetition" dynamic in which hyperscalers can simultaneously be important customers and formidable long-term competitors.

In addition, while hyperscaler insourcing remains a risk as initial long-term agreements expire, we believe neoclouds can remain important strategic partners where hyperscalers face constraints on self-built capacity. Power availability, permitting delays, transmission bottlenecks, and local opposition can all limit the pace at which hyperscalers add owned data center capacity. In that environment, neoclouds provide a flexible way to secure incremental AI infrastructure more quickly and across a broader set of locations, supporting continued partnerships even as hyperscalers expand their own internal capacity.

Neoclouds Are Moving Up-Stack

The leading neoclouds are increasingly evolving beyond their original role as specialized providers of GPU capacity and AI infrastructure. While training, inference, and model orchestration remain at the center of their platforms, companies such as CoreWeave, Nebius, Nscale, and Crusoe are beginning to add many of the foundational services traditionally associated with full-scale public clouds, including CPU compute, object and block storage, virtual networking, Kubernetes, identity and access management, security, observability, and, in some cases, managed databases. Their ambition is increasingly to own not only the AI workload, but also a larger portion of the cloud environment surrounding it.

The strategic rationale is straightforward. As GPU supply expands, raw compute should become more competitive and potentially more commoditized, making a broader software and services stack increasingly important for differentiation, customer retention, and margins. Just as important, enterprises generally do not run AI workloads in isolation: production applications require data storage, networking, databases, identity, security, governance, and application infrastructure. By supplying more of these surrounding services, neoclouds can capture a greater share of customer spending, deepen switching costs, and reduce their dependence on what could eventually become a lower-margin GPU-rental business.

This up-stack evolution increasingly positions neoclouds as potential competitors to the hyperscalers that are currently among their largest customers. Ironically, by turning to neoclouds to alleviate near-term capacity constraints, Microsoft, Meta, and others may also be helping fund the development of future competitors. The key question is whether neoclouds can build enough of the surrounding cloud stack to become credible alternatives for a meaningful share of workloads without recreating the complexity and cost structure of the hyperscalers themselves.

CoreWeave Is Building a Broader AI Cloud Platform

CoreWeave is rapidly expanding beyond GPU compute into storage, networking, Kubernetes, identity, observability, and other control-plane services, while CoreWeave Omni extends the full CoreWeave cloud software stack into customer-owned data centers through an asset-light model. The company has also used acquisitions to move meaningfully up the software stack. Its roughly $1.7 billion acquisition of Weights & Biases added model development, experiment tracking, evaluation, monitoring, and inference capabilities; subsequent acquisitions included OpenPipe, which added reinforcement learning; Marimo, which expanded its developer environment and notebook tooling; and Monolith, which extended the platform into industrial and physical AI use-cases.

Together, these moves position CoreWeave to capture more of the AI development lifecycle—from infrastructure and training through model optimization, inference, and application development—rather than simply renting GPU capacity. The distinction from Nebius is largely one of starting point: CoreWeave is building a broader AI cloud outward and upward from a GPU-infrastructure base, while Nebius is applying a multitenant cloud engineering heritage to an AI-first platform.

Moving up-stack does not require neoclouds to build every service themselves. CoreWeave's $335 million, five-year agreement with Backblaze illustrates a buy-versus-build approach, with Backblaze providing multi-exabyte, low-cost storage infrastructure that CoreWeave can integrate into its AI Object Storage offering. Such partnerships can accelerate the addition of foundational cloud services without requiring neoclouds to recreate every layer of the hyperscaler stack internally.

Nebius Comes From a Different Cloud Heritage Than Most of Its Peers
Following the separation from its predecessor organization (Yandex, aka the Russian Google), Nebius rebuilt its cloud platform from scratch with the help of hundreds of cloud and AI engineers who had experience building large-scale cloud services and infrastructure. This heritage matters because Nebius was conceived from the outset as a multitenant cloud platform, rather than primarily as a GPU farm layered on top of leased data-center capacity.

That gives Nebius a somewhat different starting point from peers such as CoreWeave, which began as a crypto-mining operation before pivoting into GPU cloud infrastructure, or IREN, whose roots are in vertically integrated Bitcoin mining. Today, Nebius offers a broad set of cloud services spanning CPU and GPU compute, object/block/file storage, Kubernetes, networking, IAM, security, and managed PostgreSQL. Its historical experience building general-purpose cloud infrastructure also positions the company well to deliver on its asset-light infrastructure strategy, in which Nebius aims to separate ownership of the underlying physical infrastructure from control of the higher-value cloud platform, operating system, and customer relationship.

Other Neoclouds Are Broadening Their Offerings as Well
Crusoe provides another example of how the leading neoclouds are expanding beyond raw accelerated compute toward something that increasingly resembles a specialized public cloud. Its platform now includes CPU and GPU virtual machines, persistent block and shared-file storage, object storage, VPC networking and firewalls, managed Kubernetes, and integrated monitoring and logging through its Command Center. These services matter because they allow customers to run more of the infrastructure surrounding an AI application on Crusoe rather than simply renting GPUs for an isolated training job. In effect, Crusoe is beginning to provide the compute, storage, networking, and orchestration elements needed to host an AI application throughout its lifecycle.

Meanwhile, Nscale has added CPU and GPU virtual machines, storage, VPC networking, Kubernetes, and enterprise identity and security capabilities. The company's pending acquisition of Anyscale extends its platform further into distributed data processing, training, inference, and agent deployment. Nscale explicitly describes the combination as extending its offering "from power to production AI."

The Enterprise Distribution Challenge
The bigger obstacle to neoclouds becoming true hyperscaler competitors may ultimately be distribution rather than technology. Adding storage, databases, networking, identity, and security is relatively straightforward compared with building the go-to-market infrastructure required to sell an integrated cloud platform to thousands of large companies. Cloud sales are highly consultative and require large direct-sales and solution-engineering teams, CIO and procurement relationships, global systems integrators and channel partners, enterprise-grade support, marketplaces, migration services, compliance expertise, and the ability to structure complex multiyear agreements. The neoclouds' early success selling scarce GPU capacity to a relatively small number of AI labs and hyperscalers does not necessarily translate into an ability to sell a broad cloud platform across the *Fortune* 1000.

Google Cloud provides a useful cautionary example for CoreWeave, Nebius, and other neoclouds. Despite possessing world-class compute, networking, storage, and distributed-systems technology, Google Cloud initially lacked the sales culture and customer relationships enjoyed by Microsoft and AWS. It took almost a decade and several leadership changes for Google to crack the code. By 2025, Google Cloud had surpassed a $50 billion revenue run-rate, in part because of its shift to a more customer-focused enterprise sales model under Thomas Kurian.

Neoclouds may therefore be able to assemble the technical pieces of a broader cloud stack relatively quickly, but building an effective sales and distribution engine will take longer. Large enterprises care about more than pure technology, including integration, governance, support, migration risk, and whether trusted partners such as Accenture, Deloitte, or Capgemini can implement and support the platform. Neoclouds will need to invest heavily in sales, solution architects, customer success, GSIs, channel partnerships, and vertical expertise—potentially adding significant operating expense well before those investments produce meaningful revenue.

Neocloud Infrastructure Stack

Unlike general-purpose clouds, neoclouds are purpose-built around the performance requirements of large-scale AI workloads. Rather than optimizing infrastructure for a broad mix of enterprise applications, their architectures are designed to maximize GPU utilization, minimize communication bottlenecks, and move enormous volumes of data quickly among compute, networking, and storage. The typical neocloud stack consists of several highly optimized layers:

- **GPU-as-a-Service (GPUaaS).** At the foundation is access to large clusters of high-performance accelerators such as Nvidia H100/H200 and Blackwell B200/B300 GPUs, with the roadmap extending to next-generation Rubin GPUs and the broader Vera Rubin platform. Neoclouds generally emphasize dedicated or tightly isolated GPU capacity and increasingly offer bare-metal configurations that eliminate or minimize virtualization overhead. CoreWeave, for example, runs its Kubernetes service directly on bare-metal GPU nodes, while other providers such as Nebius expose GPU clusters through virtual machines. The objective is the same: provide customers with predictable access to large pools of accelerators without the resource contention common in more generalized cloud environments.

- **High-Performance Networking.** GPUs become dramatically less useful at scale if they cannot communicate with one another quickly, making networking a critical element of the neocloud architecture. Leading providers deploy dedicated InfiniBand or AI-optimized Ethernet fabrics using technologies such as GPUDirect RDMA, which allows GPUs on different servers to exchange data with minimal CPU involvement. CoreWeave supports both Nvidia Quantum InfiniBand and Spectrum-X Ethernet, while Nebius connects its GPU clusters over InfiniBand with 400Gb/s interfaces per GPU; newer Nvidia Quantum-X800 infrastructure supports speeds up to 800Gb/s. This high-bandwidth, low-latency fabric is particularly important for distributed training, where thousands of GPUs may need to operate as a single system.

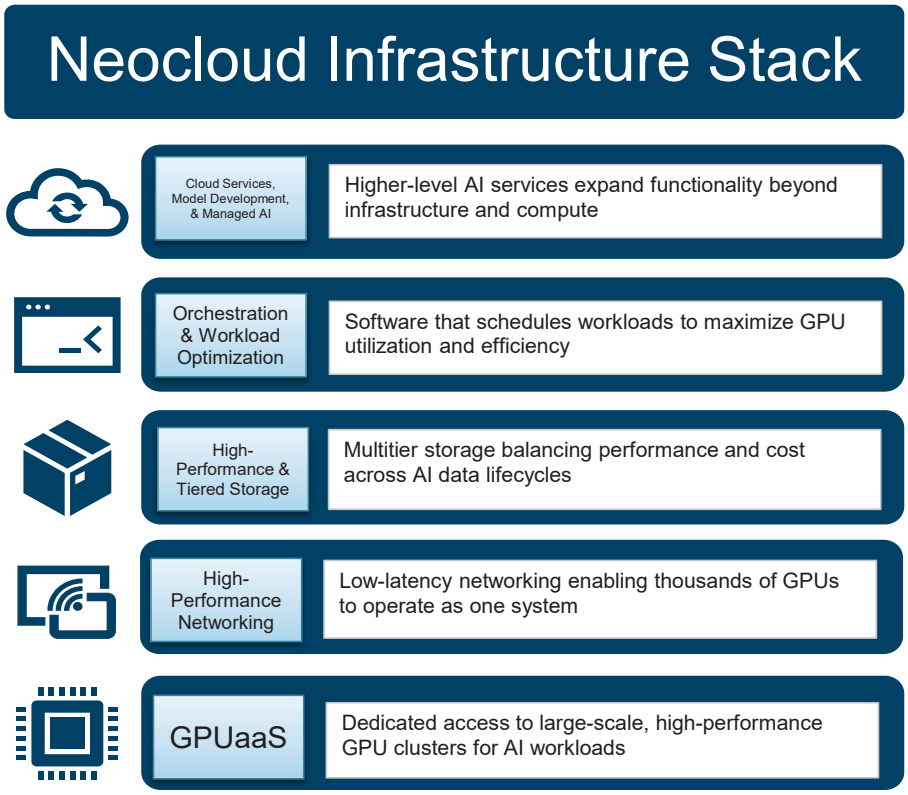
- **High-Performance and Tiered Storage.** Keeping expensive GPUs fed with data is equally important. Neoclouds therefore combine very high-performance flash and distributed file storage for active datasets, checkpoints, and model weights with lower-cost object-storage tiers for less latency-sensitive data. CoreWeave, for example, offers distributed file and object storage based on technology from VAST Data, while Nebius provides home-grown block, file, and object storage optimized for AI workloads. CoreWeave's recent $335 million Backblaze agreement illustrates how a neocloud's storage architecture can become increasingly tiered. Backblaze will supply multi-exabyte HDD-based capacity within CoreWeave's AI Object Storage offering, allowing lower-cost storage to complement its higher-performance flash storage tier.

- **Orchestration and Workload Optimization.** Sitting above the physical infrastructure is the software layer responsible for scheduling jobs, allocating GPUs, monitoring cluster health, balancing workloads, and maximizing utilization. Kubernetes and Slurm are common foundations, but neoclouds increasingly layer proprietary scheduling, telemetry, and automation on top. CoreWeave, for example, operates Kubernetes directly on bare metal and provides SUNK

to run Slurm workloads on top of Kubernetes, while Nebius supports both managed Kubernetes and Slurm-based clusters. This orchestration layer is strategically important because even modest improvements in utilization can materially improve infrastructure economics when individual GPU clusters cost hundreds of millions of dollars.

- **Cloud Services, Model Development, and Managed AI.** Above the infrastructure and orchestration layers, neoclouds are increasingly adding higher-level services that help customers build, customize, and run AI models and workloads. These include identity and access management, security, observability, managed databases, model serving, and inference, as well as tools for fine-tuning, reinforcement learning, evaluation, experiment tracking, and model monitoring. CoreWeave's acquisitions of Weights & Biases and OpenPipe, for example, extend its platform beyond core infrastructure into model development, evaluation, post-training, and reinforcement-learning workflows, while Nebius offers managed inference and model-serving capabilities through its Token factory offering. These higher-layer services can increase customer stickiness, raise revenue per MW, and allow neoclouds to capture more of the AI development lifecycle rather than competing solely on access to GPU capacity.

Exhibit 2
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Neocloud Infrastructure Stack



Source: William Blair Equity Research

Neocloud Customer Segments

We segment the neocloud customer base into five primary profiles, each driven by unique operational requirements:

1. **Hyperscalers.** Hyperscalers have become some of the neocloud industry's largest and most important customers, increasingly using third-party AI infrastructure as an extension of their own capacity supply chains. Microsoft alone has signed major capacity agreements with CoreWeave, Nebius, IREN, Nscale, Lambda, and Crusoe. Nebius's five-year Microsoft agreement carries a base contract value of approximately $17.4 billion, potentially increasing to $19.4 billion, while IREN's five-year agreement is worth approximately $9.7 billion. Nscale has contracted to provide Microsoft with infrastructure supporting approximately 200,000 Nvidia GB300 GPUs across the U.S. and Europe—an agreement reportedly worth as much as $14 billion—while Lambda has disclosed a separate multibillion-dollar Microsoft commitment. Crusoe is also developing a new 900 MW AI campus in Abilene, Texas, specifically to support Microsoft infrastructure.

   Meta has emerged as another major source of neocloud demand. CoreWeave expanded its relationship with Meta in April 2026 to approximately $21 billion of committed capacity through 2032. Nebius, meanwhile, has rapidly deepened its relationship with Meta: following an initial five-year, $2.9 billion agreement signed in late 2025, the companies entered a second five-year agreement in March 2026 with up to $27 billion of potential contract value, including $12 billion of dedicated Vera Rubin capacity and an additional commitment of up to $15 billion under which Meta effectively serves as a backstop purchaser for capacity Nebius does not sell to other customers.

   As these deals demonstrate, hyperscaler customers tend to sign the largest and longest-duration contracts, making them particularly important to the neocloud financing model. Their investment-grade credit and long-term commitments give lenders confidence to finance the GPUs and data centers required to fulfill the contracts. The trade-off is concentration: CoreWeave's top three customers represented 36%, 26% and 10% of second quarter 2026 revenue, respectively—72% combined. CoreWeave does not identify those three customers in the filing, although it identifies Microsoft and OpenAI as significant customers and expects Meta to become one.

2. **Frontier Labs.** Frontier AI labs are the most natural neocloud customer base because they need enormous, tightly interconnected GPU clusters for training and inference workloads. OpenAI is one of CoreWeave's largest customers, while Mistral has publicly described CoreWeave as an infrastructure partner for training its models. OpenAI is also the anchor participant in Nscale's Stargate Norway and Stargate UK projects.

   While neoclouds were historically used primarily by frontier labs for model training, those relationships are increasingly expanding into production inference. This shift is important for both durability and customer stickiness: training demand can be enormous but episodic, whereas inference generates more recurring, always-on compute consumption as models move into production.

3. **AI-Native Start-ups.** Below the frontier labs are a much larger group of AI-native start-ups building foundation models, coding tools, image/video models, vertical AI applications, and agents. These customers generally need less capacity individually but can grow rapidly—and based on our industry research, neoclouds can substantially cut training cycles and reduce

GPU hardware costs compared to traditional hyperscaler platforms. Nebius's disclosed customer list includes AI-native start-ups such as Cognition, Black Forest Labs, Higgsfield, and Recraft, while its second-quarter wins included Reflection AI.

This customer segment is strategically important because it offers neoclouds diversification from overdependence on a small number of mega contracts and customers. The economics here are also better for the neoclouds: AI-native start-ups typically purchase smaller blocks of capacity and receive less aggressive volume pricing. As a result, pricing per GPU-hour—and therefore revenue per MW—can be materially higher, particularly when compute is bundled with higher-value services such as managed inference, orchestration, storage, networking, and model tooling. The trade-off is greater churn potential, shorter contract duration, and higher customer-acquisition and support costs.

Nebius's second-quarter disclosures illustrate the emerging spectrum of customer contract duration: short-term capacity deals can run three to six months, core midterm contracts generally run one to three years, and investment-grade strategic customers sign longer commitments typically around five years.

4. **Large Enterprise.** Enterprise customers represent a new growth market for neoclouds, notwithstanding the aforementioned caveats on product breadth and distribution immaturity. Our research suggests that compared to traditional hyperscalers, neoclouds offer immediate provisioning, predictable bare-metal performance, and significant hourly cost advantages. CoreWeave, for example, is beginning to demonstrate tangible penetration of traditional enterprises, with customers such as Caterpillar, Bentley Systems, and Hudson River Trading announced during the second-quarter earnings call. As neoclouds move up the stack and build out a richer set of cloud services, the enterprise segment should become increasingly important.

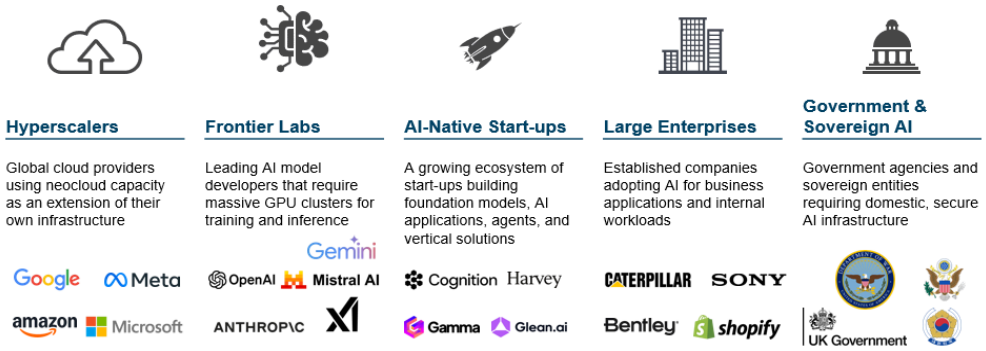
5. **Government Entities and Sovereign AI.** Government customers are an emerging segment for the neoclouds. Lambda already maintains a dedicated government business and says U.S. government agencies use its infrastructure through Cooperative Research and Development Agreements (CRADAs) and other strategic engagements. CoreWeave launched CoreWeave Federal in late 2025 specifically to serve U.S. agencies and the defense industrial base and is working toward the compliance infrastructure required for federal workloads.

   Outside the U.S., "sovereign AI" represents an underappreciated, multibillion-dollar growth opportunity for neoclouds. Governments are increasingly reclassifying AI as strategic national infrastructure, mandating that sensitive models, data, and compute remain strictly within domestic jurisdictional borders. This regulatory push favors regional neoclouds that can deploy rapidly within tier-III or -IV local data centers without the complex international data-sharing issues of global hyperscalers.

   The scale of this sovereign opportunity is illustrated by national commitments across Europe and Asia. As an example, Centuria Capital Group reports that South Korea has advanced a large-scale sovereign AI agenda with plans to deploy more than 260,000 GPUs, while Japan has committed about $65 billion in government support for AI and semiconductors through 2030. Simultaneously, the European Union has launched seven AI factories across 17 member states to build sovereign high-performance computing clusters.

   Nscale is developing sovereign U.K. infrastructure with BT and has positioned its European data-center footprint around in-country processing and data-residency requirements for public-sector customers. Crusoe's Edge Zones similarly target governments and regulated industries that need dedicated AI infrastructure located within their own jurisdictions.

Exhibit 3
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Neocloud Customer Segments



Source: William Blair Equity Research

Exhibit 4
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Major Disclosed Neocloud Contracts

Date	Neocloud	Ownership	Customer	Contract Amount*	Contract Length
3/10/2025	CoreWeave	Public	OpenAI	$11,900	5 years
5/1/2025	CoreWeave	Public	OpenAI	$4,000	5 years
9/8/2025	Nebius	Public	Microsoft	$17,400	5 years
9/23/2025	CoreWeave	Public	OpenAI	$6,500	5 years and 8 months
9/30/2025	CoreWeave	Public	Meta	$14,200	6 years and 4 months
10/15/2025	Nscale	Private	Microsoft	Up to $14,000	NA
11/3/2025	IREN	Public	Microsoft	$9,700	5 years
11/3/2025	Lambda	Private	Microsoft	Multibillion	Multiyear agreement
11/11/2025	Nebius	Public	Meta	$3,000	5 years
11/12/2025	Fluidstack	Private	Anthropic	$50,000	Multiyear agreement
3/16/2026	Nebius	Public	Meta	$27,000	5 years
4/9/2026	CoreWeave	Public	Meta	$21,000	6 years
4/14/2026	Nscale	Private	Microsoft	NA	NA
7/20/2026	IREN	Public	Undisclosed	$2,800	4 years
8/26/2026	Nscale	Private	Anthropic	$45,000	6 years
8/31/2026	Lambda	Private	Anthropic	$35,000	6 years
9/3/2026	Nscale	Private	Figure Robotics	$3,500-$6,000+	NA
9/3/2026	Crusoe	Private	Jane Street Partners	$13,000	5 years
9/24/2026	Akamai	Public	Anthropic	$11,600-$20,000	7 years

*Note: All numbers in millions unless otherwise specified

Sources: Company documents, public sources, and William Blair Equity Research

Exhibit 5
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Anthropic's Compute Commitments ($B) Since 2025



Note: Figures reflect disclosed headline commitments. Some are multi-year maximums or nominal contract values
Source: Company Reports and William Blair Equity Research

Neocloud Approaches to Owning the Physical Infrastructure

Beneath the GPU layer, neoclouds span a spectrum from relatively asset-light leasing to fully integrated data-center ownership. The key trade-off is capital efficiency and speed versus control and economics. Leasing can produce a faster, less capital-intensive expansion model, while data center ownership can improve long-term unit economics and create valuable infrastructure assets—but at the cost of significantly greater upfront investment and execution risk.

In general, we see three main ownership models:

1. **Lease/Colocation.** In the most asset-light model, the neocloud leases powered data-center capacity from a third-party operator (like Equinix or Digital Realty) and installs its own GPUs, servers, networking, and storage inside the facility. This approach minimizes upfront real-estate and infrastructure capex, speeds deployment, and gives the operator flexibility to enter new markets quickly. The trade-off is less control over power, cooling, and facility economics, while a portion of the infrastructure margin goes to the landlord. CoreWeave has historically relied on this colocation-centric approach, though by 2028 the company expects to have a meaningful portion of its powered capacity delivered through company-owned and operated data centers.

2. **Build-to-Suit/Joint Venture.** A middle-ground approach uses a third-party developer or infrastructure investor to fund and construct a facility specifically designed around the neocloud's requirements. The neocloud can retain significant control over technical design, power

density, and operations without funding the entire data center build itself. Joint ventures can further reduce the equity burden by bringing in outside infrastructure capital. Nebius and Crusoe make meaningful use of these structures, particularly for large bespoke deployments.

3. **Owned and Vertically Integrated.** At the asset-heavy end of the spectrum, the neocloud may own and operate more of the physical infrastructure stack, potentially including land, power interconnects, substations, cooling systems, and the data-center facility itself (in addition to the compute equipment). This model requires substantially more capital and carries greater development risk, but it provides maximum control over power costs, deployment timing, and long-term infrastructure economics while allowing the operator to retain more of the value created below the GPU layer. IREN is the clearest example among the major neoclouds of this model, largely owning the land, buildings, and electrical infrastructure supporting its AI cloud. Nscale and Crusoe are also increasingly pursuing a vertically integrated data center strategy.

Exhibit 6
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Major Neocloud Physical Ownership Models

Company	Data Center Model	What It Means
CoreWeave	Predominantly leased/build-to-suit, with selective JVs	CoreWeave primarily leases long-term capacity while owning GPUs and compute. At YE25, $38.5B of uncommenced lease payments generally spanned 5–17 years. Selective down-stack investments include the 250 MW Kenilworth, NJ, greenfield JV.
Nebius	Hybrid: owned + colo + build-to-suit	Nebius owns infra in Finland; colocates in Kansas City, France, and Iceland; and uses build-to-suit. Example: For its NJ data center, the developer owns land and power, while Nebius designs and leases the facility.
IREN	Predominantly owned/vertically integrated	IREN generally owns and operates its data centers, electrical infrastructure, compute, and much of the land. Its legacy crypto portfolio provides grid-connected sites now being converted and expanded for AI.
Crusoe	Vertically integrated developer/operator, heavily using infrastructure JVs	Crusoe designs, builds, and operates AI data centers while sharing funding and ownership. Its Abilene campus uses a Blue Owl/Primary Digital JV; Crusoe develops and operates it, with capacity leased long term to hyperscalers, limiting balance-sheet cost.
Nscale	Hybrid vertically integrated model	Nscale seeks infrastructure control through varied structures: It has purchased infrastructure; operated its own data centers; and used partner facilities as well. Its Norway projects have also used JVs, though Nscale consolidated Aker-Nscale within the company in 2026.

Sources: Company information, SEC, and William Blair Equity Research

Neocloud Contract Types

Neocloud contracts range from short-duration, variable consumption to multiyear capacity commitments. The basic trade-off is that shorter-duration contracts carry higher pricing but leave the neocloud with more utilization risk, while longer-term commitments provide greater revenue visibility and can be used to finance the underlying infrastructure.

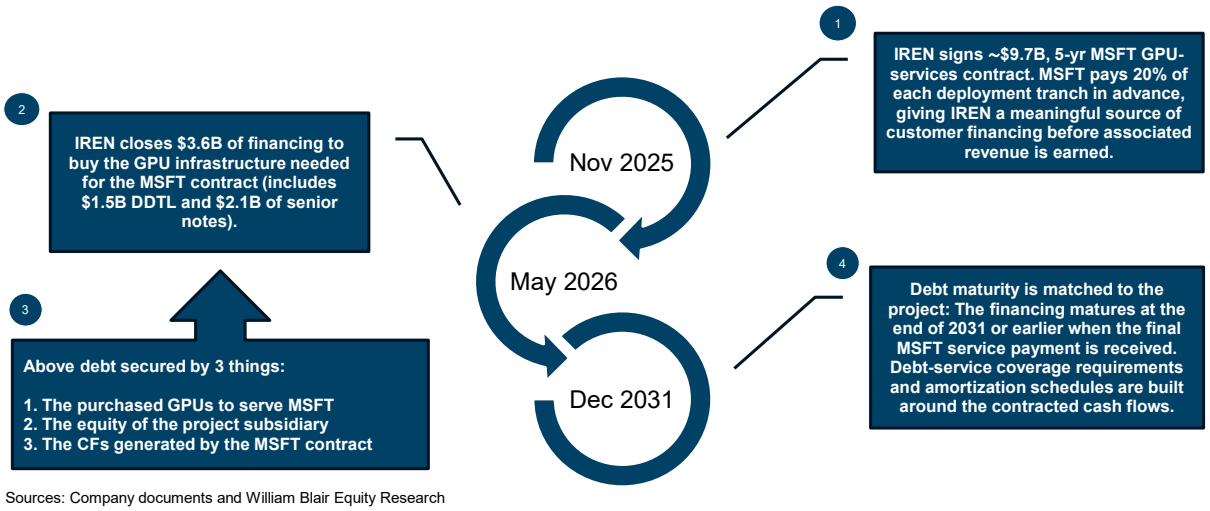
We see three primary contract types:

1. **Core AI Cloud Contracts.** These are typically one- to three-year capacity commitments serving large AI-native companies and enterprises. Contracts may reserve dedicated GPU capacity or require a minimum level of spend, giving customers greater capacity assurance while providing the neocloud with enough revenue visibility to support infrastructure investment. Customer prepayments can also be significant—in Nebius's case, upfront payments on recent deals have covered 50%-60% of associated capex.

2. **Short-Duration/On-Demand Capacity.** These contracts address immediate or time-bounded demand and generally range from true pay-as-you-go usage to commitments of several months. Because the neocloud retains more utilization risk and customers value immediate access to scarce capacity, pricing can be materially higher than under longer-term agreements; Nebius, for example, has cited short-duration capacity generating roughly $40 million-$50 million of ACV per MW, approximately twice the level of many of its larger core cloud contracts. The trade-off is substantially less revenue visibility and limited ability to finance new infrastructure against these commitments.

3. **Long-Term Strategic Capacity Agreements.** The largest hyperscaler and frontier-lab contracts typically extend three to five years or longer and are often structured as dedicated-capacity and/or take-or-pay agreements. Customers commit to pay for a specified amount of infrastructure regardless of actual utilization, materially reducing demand risk and allowing the neocloud to finance GPUs and data-center capacity against contracted cash flows. CoreWeave's committed contracts, for example, had a weighted-average duration of roughly five years at the end of 2025 and generally include customer prepayments, while Nebius's $17.4 billion Microsoft agreement provides dedicated capacity under a five-year term. These large, investment-grade contracts play an especially important role in underwriting the industry's capital-intensive buildout.

Exhibit 7
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
IREN + Microsoft Contract Example



Sources: Company documents and William Blair Equity Research

Asset-Light Models

In an asset-light model, the neocloud operates its cloud platform on infrastructure that it does not own. Ownership of the underlying facility and hardware can sit with the customer, as in CoreWeave's Omni model, or with third-party infrastructure and financial partners, as in Nebius's

announced structure. In either case, the neocloud contributes its systems architecture, cloud software, operating expertise, and platform management while avoiding much of the capital required to own the physical infrastructure itself. This can accelerate capacity expansion, improve capital efficiency, and reduce exposure to long-lived real estate and power investments, although the neocloud gives up some control over the underlying assets and, depending on the structure, may share economics with the infrastructure owner.

In March 2026, CoreWeave launched Omni, a cloud-as-a-service, asset-light offering under which CoreWeave deploys and operates its full cloud stack inside a customer's data center, while the customer retains ownership of the facility and hardware. In July 2026, Nebius announced a separate asset-light model in which infrastructure partners finance and own the facilities and hardware, while Nebius provides the systems architecture and software stack, deploys and maintains the platform, and brings the resulting capacity to market through its own sales organization. Nebius management said the model could begin adding incremental capacity in 2027, although it has already generated dozens of inquiries from prospective partners.

The Neocloud Business Model

The neocloud business model combines core competencies in data center systems engineering and specialized cloud software with expertise in infrastructure development and project finance. The model typically begins with securing multiyear customer commitments, which are then used to support the financing and deployment of GPUs and data center capacity. Returns are driven by utilization, pricing, asset life, and financing costs, while customer prepayments and long-term take-or-pay contracts can materially reduce both capital requirements and utilization risk. In effect, the best neoclouds are technology-enabled infrastructure platforms that use contracted demand to finance and monetize steadily depreciating computing assets (although older GPU prices have been appreciating, not depreciating).

Gross Margin

Reported gross margins are not particularly useful across the neoclouds, because companies classify GPU depreciation differently. As an example, CoreWeave's reported gross margin excludes depreciation on the GPUs and servers, which sits in technology and infrastructure expense, below cost of revenue. Second quarter 2026 D&A for CoreWeave was $1.39 billion on only $2.58 billion of revenue, so adjusted gross margin would be 12.1%.

Similarly, Nebius's cost of revenue was only 23% of revenue in the second quarter—a 77% reported gross margin—because it also separates out reported D&A (which in the second quarter was equal to 45% of revenue). Adjusting this out would yield a gross margin of 32.6%. This paints a clearer picture of actual gross margin (cost of physically delivering the service, which includes power/rent/DC operating expense and depreciation of GPUs/networking/DC infrastructure). Over the longer term, we believe that adjusted gross margin in the 40% range is realistic for the typical neocloud business, depending on utilization, pricing, and asset life.

Capex/Revenue Ratio

CoreWeave spent $14.1 billion on property and equipment in the first half of 2026 against only $4.65 billion of revenue (303% ratio). Nebius had capital expenditures of approximately $5.7 billion in the second quarter against $582 million of revenue (980% ratio). At this early stage of the buildout, these inflated ratios are not surprising.

For evaluating the underlying business, the more important metric is maintenance capex/mature revenue because all these companies are building capacity that will produce revenue in 2027 and beyond. If a fully utilized fleet costs 1.0-1.5 times its annual revenue base to construct and equipment lasts roughly five years, maintenance capex eventually should settle somewhere around 20%-30% of revenue for these companies.

Exhibit 8
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Table of Aggregate GW Commitments

Company	Contracted/Secured Capacity	Public Capacity Goal or Pipeline
CoreWeave	4.2 GW of contracted power, 1.5 GW of active power as of 2Q26	>8 GW of active power by YE 2030
Nebius	>3.5 GW as of May 2026	5 GW of contracted power by YE 2026
IREN	5 GW of secured power	Currently undisclosed
Nscale	1.37 GW of active and contracted capacity	Line of sight to 10 GW+ of potential DC capacity
Crusoe	4.9 GW of contracted AI infrastructure as of June 2026	Development pipeline over 40GW (no timeline given)
Lambda	Currently undisclosed	>2 GW deployed by YE 2030
Total	***18.5+ GW***	***55+ GW***

Sources: Bloomberg, company documents, and William Blair Equity Research

EBITDA Margin

At high utilization, the incremental economics are strong because much of the cost base is fixed. Once a cluster is built, the principal cash expenses are power, rent/colocation, networking, operating personnel, and support. GPU purchases do not affect EBITDA, because they are capitalized.

EBIT Margin

This is the more precise and practical profitability metric, in our view, as it includes equipment depreciation. We believe a mature, steady-state neocloud could generate an EBIT margin of roughly 25%, assuming approximately 50% EBITDA margin and 25% depreciation as a percentage of revenue. Management teams at both CoreWeave and Nebius have indicated that economics in this range are achievable over time.

Exhibit 9
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Gross Margin, EBITDA Margin, and EBIT Margin of Public Neoclouds

Company	Ticker	2026 Adjusted Gross Margin*	2026 EBITDA Margin	2026 EBIT Margin	2027 Adjusted Gross Margin*	2027 EBITDA Margin	2027 EBIT Margin
CoreWeave	CRWV	16.1%	58.3%	8.2%	23.1%	62.8%	15.4%
Hut 8	HUT	0.7%	39.5%	-235.9%	1.1%	58.8%	-30.3%
IREN Ltd	IREN	20.1%	59.9%	-36.6%	31.0%	73.6%	12.7%
Nebius	NBIS	15.2%	40.8%	-20.8%	15.1%	52.9%	-1.2%
TeraWulf	WULF	8.2%	21.0%	-145.2%	44.4%	60.5%	11.9%
White Fiber	WYFI	34.5%	33.2%	-20.6%	44.1%	51.5%	19.2%
Median		***15.6%***	***40.1%***	***-28.7%***	***27.0%***	***59.7%***	***12.3%***
Mean		***15.8%***	***42.1%***	***-75.1%***	***26.5%***	***60.0%***	***4.6%***

*Adjusted gross margin includes depreciation and amortization expense

Sources: FactSet, William Blair Equity Research

Operating Cash Flow

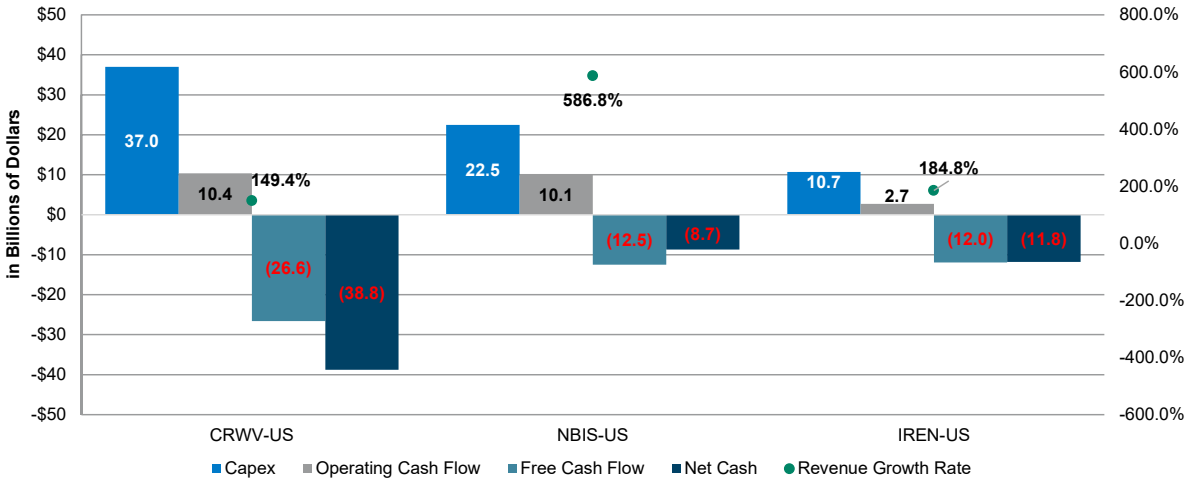
This is another somewhat misleading (and inflated) metric, in our view, due to the impact of customer prepayments. CoreWeave generated $3.7 billion of operating cash flow in the first half of 2026 despite reporting a $1.37 billion net loss, aided by a $1.37 billion increase in deferred revenue as customers paid in advance for future capacity, as well as $2.54 billion of noncash depreciation and amortization. CoreWeave's committed contracts typically include upfront payments, which historically have averaged 15%-25% of total contract value.

The effect is even more pronounced at Nebius. The company generated $2.25 billion of operating cash flow in the second quarter on just $582 million of revenue, and $4.50 billion of operating cash flow in the first half. The primary driver was customer prepayments: deferred revenue increased by approximately $3.2 billion in the first quarter and another $1.2 billion in the second quarter, bringing the first-half increase to $4.4 billion. Nebius now expects to receive more than $9 billion of customer prepayments during 2026, which management views as a source of funding that reduces the amount of debt and equity required to finance its capacity buildout.

Free Cash Flow

This is the "rubber meets the road" metric for an asset-heavy business. However, it is not a useful metric in the early stage of a massive infrastructure buildout—the neoclouds are spending aggressively today to generate revenue several years into the future. In first half 2026, CoreWeave reported free cash flow margin of roughly -225%, while Nebius reported free cash flow margin of roughly -623%. Over the long term, we believe roughly 20% unleveraged free cash flow margins are realistic for these businesses, assuming a sustainable 25% EBIT margin (maintenance capex roughly equals annual depreciation), a roughly 20% tax rate, and neutral working capital.

Exhibit 10
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
CY26 FCF, Capex, and Revenue Growth Consensus for Major Neoclouds



Note: CRWV & NBIS capex numbers are based on annual guidance; consensus estimates as of 09/28/2026. All other CRWV & NBIS numbers are William Blair estimates.
Sources: Company reports, FactSet estimates, and William Blair Equity Research

Payback Period

CoreWeave disclosed at the time of its IPO that committed contracts at the end of 2024 implied an average cash payback of roughly 2.5 years on GPUs and associated infrastructure, including the benefit of customer prepayments. More recently, Nebius indicated that the major deals it signed in second quarter 2026 carried a revenue payback period of just 1 year and 10 months, with an even shorter cash payback period when including prepayments. For assets with an expected useful

life of roughly five to six years, these are highly attractive economics: the initial investment can be recovered well before the hardware reaches the back half of its life, leaving meaningful upside even if pricing and utilization decline as GPUs age.

Revenue/MW Versus Capex/MW

Revenue per megawatt is a useful starting point for evaluating neocloud economics, but it is not, by itself, a measure of profitability or capital efficiency. Nebius has said its original 2026 baseline was roughly $12 million of annual contract value per MW, while second quarter 2026 deals averaged more than $20 million/MW, including four large contracts in the $20 million-$25 million/MW range. For scarce, short-duration capacity, both Nebius and CoreWeave have cited pricing as high as $40 million-$50 million/MW.

DigitalOcean provides a useful benchmark here: its general-purpose cloud business generated more than $22 million of ARR/MW historically. However, for incremental AI capacity, management has been underwriting roughly $13 million of ARR/MW, with potential upside as customers layer on inference, core cloud, storage, and other higher-value services.

On the cost side, our industry research suggests that $25 million-$35 million of capex per MW is a reasonable range for the compute-heavy portion of a modern neocloud deployment, although cutting-edge configurations and more vertically integrated projects can run materially higher. As a reference point, DigitalOcean management has said its prior $20 million-$25 million/MW capex assumption is likely to rise for new capacity coming online in 2027 due to component inflation and more advanced hardware.

CoreWeave's Pennsylvania data center project illustrates the wide range of possibilities for more integrated builds: the company has discussed investing up to $6 billion in a campus initially sized at 100 MW and potentially expanding to 300 MW, implying roughly $20 million-$60 million/MW depending on ultimate scope and what is included in the investment.

If a neocloud can generate $10 million-$20 million of annual revenue on $25 million-$35 million of installed capital, that implies roughly 0.3x-0.8x annual asset turnover, with a midpoint near 0.5x. Those are attractive economics for a capital-intensive infrastructure business, particularly if gross margins after depreciation can settle in the 35%-45% range. The strongest platforms could push asset turnover toward or above 1.0x through higher utilization, stronger pricing, and greater software, inference, and managed-service attach.

ROIC and Equity Returns

Customer prepayments can further enhance returns on the neocloud's own capital. Nebius, for example, has said prepayments are covering roughly 50%-60% of capex associated with many recent contracts. A $100 infrastructure investment may therefore require only about $40-$50 of funding from the neocloud at the outset, with the balance effectively funded by the customer. If much of the remaining requirement is financed with debt, the equity contribution can be relatively modest, allowing equity returns to exceed the underlying unleveraged return on the infrastructure.

Exhibit 11
Rise of the Neoclouds: From GPU Specialists to Long-Term Hyperscaler Challengers
Illustrative Economics Per MW

Illustrative metric	Downside case	Base case	Upside case
Installed capital per MW	$35 million	$30 million	$25 million
Annual revenue per MW	$10 million	$15 million	$20 million
Asset turnover	0.3x	0.5x	0.8x
Cash gross margin	65%	70%	75%
Depreciation as % of revenue	30%	25%	20%
EBIT margin	20%	25%	30%
After-tax unlevered ROIC	5%	10%	19%
Customer prepayment as % of capex	20%	30%	40%
Indicative cash payback	>5 years	~2 years	<2 years

Source: William Blair Equity Research

Utilization Is the Key Driver of Neocloud Returns

Revenue per MW receives considerable investor attention because it provides a simple way to compare contract values with infrastructure costs. However, neocloud economics also depend heavily on utilization. Once a cluster is installed, many costs—including GPUs, networking, data-center space, depreciation, and financing—are fixed or semi-fixed. Even modest changes in utilization can therefore have an outsized impact on EBIT, free cash flow, and returns on invested capital.

The challenge is that utilization is not always reported consistently. Capacity can be described as contracted, deployed, available, reserved, or utilized, and those terms are not interchangeable. For investors, the most useful measure is revenue-producing utilization: the percentage of installed, service-ready capacity that is actually generating contracted or consumption-based revenue. That is different from physical GPU utilization, which simply measures how intensively the hardware is running at a given point.

High Fixed Costs Create Significant Operating Leverage
Consider a deployment capable of generating $20 million of annual revenue per MW at full utilization. At 90% utilization, revenue would be approximately $18 million; at 70%, $14 million; and at 50%, $10 million. Yet operating expenses do not fall proportionately. Data-center rent, personnel, depreciation, financing costs, and much of the contracted power expense remain largely fixed, making margins and returns highly sensitive to utilization.

This is particularly important for operators carrying substantial debt, leases, or minimum power commitments. A deployment can remain EBITDA positive while generating insufficient cash to service financing obligations or fund the next hardware refresh cycle.

Contract Structure Determines Who Bears Utilization Risk
Long-term take-or-pay contracts shift much of the near-term utilization risk to the customer because reserved capacity generates revenue regardless of actual consumption. These agreements provide revenue visibility and support project financing during the period when providers are seeking to recover their upfront infrastructure investment.

Short-duration and on-demand capacity leave more risk with the neocloud. Pricing may be higher, but providers must continually remarket capacity as contracts expire, creating greater exposure to customer churn, spot pricing, and transition downtime. One- to three-year cloud commitments sit between these extremes, with the degree of protection determined by minimum-spend requirements, reservation terms, and the provider's ability to remarket capacity.

Investors should therefore consider not only total contracted revenue, but also how much deployed capacity is protected by firm minimum payments. High utilization backed by short-duration demand may be less durable than somewhat lower utilization supported by long-term take-or-pay contracts.

Training and Inference Create Different Utilization Profiles

Training workloads are typically large, concentrated, and time-bounded, creating periods of extremely high utilization followed by potential gaps between major jobs. Inference is more recurring but can fluctuate by time of day, geography, model, and traffic levels, often requiring spare capacity to meet peak demand and service-level commitments.

Agentic workloads add further complexity because they combine model inference with CPU orchestration, data retrieval, code execution, networking, and storage. This makes whole-system utilization increasingly important: expensive GPUs can remain idle if other parts of the infrastructure stack become bottlenecks.

The strongest platforms should therefore be able to pool diverse workloads across a common infrastructure base. A mix of training, recurring inference, short-duration capacity, and workloads suited to older hardware should support higher fleet utilization and more stable returns than reliance on a single workload or customer type.

Older GPUs and Software Can Extend Asset Utilization

Older accelerators can remain useful for inference, fine-tuning, and less performance-sensitive workloads after they are no longer optimal for frontier training. Successfully matching workloads to the most economical hardware can extend asset lives, improve lifetime utilization, and reduce the effective cost of hardware refreshes.

Software plays an important role in this process. Better scheduling and orchestration can reduce idle capacity, shift workloads across heterogeneous accelerators, recover capacity from failed jobs, and place lower-priority workloads into otherwise unused operating windows. Because much of the infrastructure cost has already been incurred, even modest utilization improvements can generate meaningful incremental profit.

The Renewal Period Is the Critical Utilization Test

Take-or-pay contracts can protect initial utilization and support attractive payback periods, but they do not eliminate long-term remarketing risk. The critical test comes when an anchor contract expires and the provider must renew the customer, reprice the capacity, or remarket older hardware to another workload.

The initial contract determines whether the provider can finance and recover its investment; subsequent contracts help determine the lifetime return. Platforms capable of moving capacity efficiently among customers, workloads, and hardware generations should therefore be better positioned than operators running highly customized infrastructure with limited alternative uses.

Over time, sustaining fleet utilization may become a more durable competitive advantage than access to any specific GPU generation. Hardware availability should broaden, but maintaining high utilization across a heterogeneous fleet requires workload breadth, sophisticated scheduling software, engineering expertise, and customer diversification.

Utilization and Pricing Must Be Evaluated Together
High utilization does not create attractive returns if pricing is too low, while high pricing cannot compensate for substantial idle capacity. A provider operating at 90% utilization and $20 million of annual revenue potential per MW generates approximately $18 million of realized revenue. At 70% utilization, that falls to $14 million; with both 70% utilization and a 20% pricing decline, revenue falls to approximately $11.2 million per MW.

The key debate is therefore not simply how quickly GPU pricing normalizes as supply expands, but whether neoclouds can offset that pressure through higher utilization, better workload mix, greater software and services attach, lower infrastructure costs, and longer asset lives. Stronger operators may successfully trade some pricing for higher fleet utilization; weaker providers risk seeing both pricing and utilization deteriorate, putting meaningful pressure on margins, ROIC, and debt-service capacity.

Neocloud Competitive Landscape

According to ABI Research and McKinsey, there are over 100 distinct neocloud operators and specialized AI cloud providers globally. The flood of activity is a direct response to two structural forces: a global scarcity of high-end compute, and the revenue diversification strategies of the largest advanced-chip producers (especially Nvidia). From a competitive standpoint, the market appears to be coalescing around different tiers of providers, based on scale, with the top tier including CoreWeave, Nebius, IREN, Nscale, Crusoe, Lambda, and Fluidstack.

CoreWeave (Nasdaq: CRWV)

History
CoreWeave is arguably the company that defined the neocloud category. The founders began mining Ethereum with GPUs in 2016/2017, accumulated large amounts of GPU hardware through the 2018-2019 crypto downturn, and subsequently realized that the infrastructure could address GPU-intensive workloads such as rendering and AI. The company gradually abandoned its crypto origins, built a specialized Nvidia cloud, and went public in March 2025.

Product Offering
CoreWeave has moved well beyond basic GPU rental. It provides GPU and CPU compute, bare metal, managed Kubernetes and Slurm, high-performance networking, object and distributed file storage, observability, training and inference services, and increasingly AI-development tooling. Acquisitions such as Weights & Biases and OpenPipe have extended it further into model development, experimentation, and reinforcement learning, while newer products such as Interconnect, SUNK Anywhere, and LOTA are designed to make CoreWeave part of a customer's broader multi-cloud environment.

Differentiation
CoreWeave's biggest advantages are scale, early access to Nvidia's newest architectures, operating expertise around very large GPU clusters, and increasingly low-cost access to capital. The company has developed a strong record of bringing new GPU platforms into production quickly, supported by networking, storage, orchestration, and control-plane software designed specifically for AI workloads. CoreWeave is further along than most competitors in converting long-term, take-or-pay customer commitments into project-style infrastructure financing, creating a reinforcing cycle in which high-quality contracts lower its cost of capital and support faster capacity deployment. Its growing software stack—including Weights & Biases and other acquired capabilities—also gives it an opportunity to differentiate beyond raw GPU access over time.

Key Customers

CoreWeave has assembled the deepest roster of large AI customers in the neocloud sector, including Microsoft, OpenAI, Meta, Anthropic, Cohere, and Mistral, alongside a growing enterprise base that includes Caterpillar, Bentley Systems, Databricks, Jane Street, and Hudson River Trading. OpenAI has committed up to roughly $22.4 billion across several agreements, while Meta followed an initial $14.2 billion agreement with an additional $21 billion commitment in April 2026. CoreWeave also has a significant Microsoft relationship, although the aggregate contract value has not been disclosed. Reported revenue remains highly concentrated: the company's three largest customers represented 36%, 26%, and 10% of second quarter 2026 revenue, respectively, or 72% combined. However, concentration is materially lower when viewed against contracted backlog, which is more relevant to the future.

Financial Profile

CoreWeave is by far the largest pure-play neocloud in terms of scale—its second quarter 2026 revenue of $2.6 billion was likely greater than that of all other neoclouds combined. Second-quarter revenue grew 112% year-over-year, with adjusted EBITDA of $1.5 billion, or a 59% margin. However, the company generated just $128 million of adjusted operating income (roughly equivalent to non-GAAP EBIT), or a 5% margin, which includes substantial depreciation expense, and reported an adjusted net loss of $567 million, driven largely by $640 million of net interest expense. Free cash flow remains deeply negative as the company funds rapid capacity expansion: operating cash flow was $679 million in the second quarter against roughly $6.4 billion of cash purchases of property and equipment, implying approximately negative $5.7 billion of free cash flow. On a year-to-date basis, free cash flow was negative $10.5 billion.

Nebius Group (Nasdaq: NBIS)

History

Nebius emerged in 2024 from the former Yandex N.V. after Yandex divested all of its Russia-based businesses for approximately $5.4 billion and severed its Russian ties. The remaining company was renamed Nebius Group, resumed Nasdaq trading as NBIS, and brought with it a substantial engineering organization with years of experience building search, distributed computing, and cloud infrastructure.

Product Offering

Of the major neoclouds, Nebius most closely resembles a genuine public cloud built around AI. Beyond GPU clusters, it provides CPU/GPU VMs, InfiniBand, Slurm, managed Kubernetes, block and file storage, S3-compatible object storage, container registry, observability, IAM, and managed PostgreSQL. Above this infrastructure sits its token factory for managed inference, MLflow, and other AI-development services.

Differentiation

Nebius's key advantages are in software and cloud engineering depth. Whereas several competitors start with power and GPUs and work upward, Nebius inherited expertise in building distributed cloud software and is now scaling the physical layer. That makes it perhaps the neocloud furthest along in creating a broader hyperscaler alternative rather than just an AI training/inference platform.

Key Customers

Nebius's customer set spans hyperscalers, AI-native companies, and select enterprises. Its two largest disclosed relationships are Microsoft, which signed a $17.4 billion agreement through 2031 that can expand to approximately $19.4 billion, and Meta, whose latest five-year agreement has a potential value of up to $27 billion. The Meta agreement includes $12 billion of dedicated capacity and up to $15 billion of additional capacity that Meta has committed to if Nebius is unable to sell that capacity to third-party Nebius AI Cloud customers. In second quarter 2026, Nebius closed four

major AI Cloud deals—Reflection, Cohere, an unnamed U.S. AI lab, and a large U.S. quantitative-trading firm—with average TCV exceeding $1 billion per contract and ACV of $20 million-$25 million/MW. Nebius does not specify revenue concentration by customer, but hyperscaler deals (typically bare-metal deals only) will be roughly 50% of the company's ARR target by the end of 2026. Even as the broader customer base diversifies, the scale of the Microsoft and Meta commitments (which do not fully ramp up until 2027) should lead to meaningful future customer concentration.

Financial Profile

Second quarter 2026 group revenue reached $582 million, of which $575 million came from AI Cloud, and AI Cloud ARR exited June at $3.0 billion. AI Cloud adjusted EBITDA margin reached roughly 50%, although this excludes heavy D&A associated with capex investments (second-quarter capex alone was about $5.7 billion). Nebius says it has more than $40 billion of customer commitments, expects over $9 billion of customer prepayments in 2026, and raised its year-end 2026 contracted-power target to 5 GW. The company also demonstrated the viability of the project-finance model with a $775 million GPU/cash-flow-backed financing priced at SOFR + 250 basis points.

IREN (Nasdaq: IREN)

History

IREN is a good example of a crypto infrastructure firm converting itself into an AI infrastructure cloud. The company built large renewable-power-based Bitcoin mining campuses and is now aggressively reallocating those power, land, and data center assets toward AI. That transition accelerated meaningfully in the June 2026 quarter, when IREN AI Cloud revenue more than doubled sequentially to $70.5 million and surpassed Bitcoin mining revenue of $66.7 million for the first time, versus $34 million and $111 million, respectively, in the March quarter.

Product Offering

IREN historically offered a relatively infrastructure-heavy AI cloud centered on dedicated Nvidia GPU clusters, high-speed networking, and storage. However, its recently completed $625 million acquisition of Mirantis materially broadens the software layer, adding Kubernetes-based orchestration, cluster management, enterprise support, and AI infrastructure management capabilities. IREN owns and operates the underlying data centers, giving it tighter control of the infrastructure than many competitors dependent on colocation.

Differentiation

IREN's moat starts below the cloud layer: access to enormous parcels of grid-connected land and power, the ability to develop the facilities itself, and its experience operating power-intensive compute campuses—all of which should translate into an attractive cost structure and potentially faster deployment of data centers. At the same time, IREN's software and go-to-market capabilities are less mature than those of CoreWeave or Nebius, meaning it has work to do if it wants to capture a greater share of customer spending above the infrastructure layer.

Key Customers

The IREN AI cloud business is anchored by two large strategic customers. Microsoft signed a $9.7 billion, five-year agreement expected to generate roughly $1.94 billion of annualized revenue at full deployment, while Nvidia subsequently signed a $3.4 billion, five-year managed-cloud agreement worth roughly $680 million annually. IREN has since expanded its customer roster to include Perplexity, Figure AI, Together AI, Fluidstack, Fireworks AI, Fal AI, Hume AI, and others, announcing another $2.8 billion of multiyear contracts in July. Nevertheless, customer concentration remains high: Microsoft and Nvidia together represent roughly $2.6 billion of annualized contracted revenue, equivalent to 75%-80% of IREN's currently contracted AI cloud ARR and roughly two-thirds

of its more than $4 billion year-end 2026 ARR target. The $2.8 billion of recently signed contracts should materially broaden the customer base, but customer concentration with Microsoft and Nvidia will remain high in the near term.

Financial Profile

On its second-quarter earnings call, IREN reported approximately $4 billion of contracted AI cloud ARR for 2026 capacity, which management says is largely sold out, and roughly $1 billion of ARR as of late August. The company is targeting approximately 0.3 GW of cumulative AI cloud capacity by the end of 2026 and 0.8 GW by the end of 2027. Recent contracts include customer prepayments equal to roughly 45%-55% of GPU capex, while new GPU financings are funding about 90% of the associated hardware cost. As of June 30, IREN had $5.9 billion of cash and cash equivalents plus $1.7 billion of restricted cash, and management cited approximately $14 billion of combined cash, committed GPU financing, and customer prepayments available to support the buildout.

Nscale (Private, S-1 Filed)

History

Nscale's roots go back to 2022-2023, when founder Josh Payne and his team began assembling powered land in low-cost energy markets while still part of Arkon Energy. Nscale was spun out from Arkon in May 2024 with an infrastructure-first strategy focused on building against contracted customer demand and controlling more of the physical infrastructure than the typical colo-based neocloud. Since then, the company has raised more than $3.3 billion through its series financings and expanded aggressively through both organic development and acquisitions. In March 2026, Nscale acquired AIPCorp, including the Monarch Compute Campus in West Virginia, and established Nscale Energy & Power to bring power origination and development capabilities in-house. The company filed for a U.S. IPO in September 2026 and intends to list on the NYSE under the ticker NSCL.

Product Offering

Nscale describes its offering as two integrated products spanning "power to token." Nscale Infrastructure combines power generation, liquid-cooled data centers, GPUs, networking, and fleet operations under long-term take-or-pay contracts, while Nscale Cloud provides Kubernetes, managed Slurm, bare-metal and virtual instances, identity and security, model and data registries, fine-tuning, and serverless inference. Its pending acquisition of Anyscale—the commercial platform built by the creators of Ray (a distributed, open-source AI compute engine)—will materially deepen the software layer by adding distributed workload orchestration across data processing, training, reinforcement learning, and inference. Ray has reached roughly 740 million cumulative downloads, and about 200 Anyscale employees are expected to join Nscale when the acquisition closes.

Differentiation

Nscale is betting on vertical integration, low-cost power, open-source software, and sovereign AI rather than simply renting GPUs. The company increasingly controls the full stack from power origination and data-center development through compute, networking, fleet management, and cloud services. The Monarch Compute Campus in West Virginia has a development runway of more than 8 GW of gross power and more than 6.5 GW of IT load; its 200 MW Sines campus in Portugal is expected to support more than 66,000 Vera Rubin GPUs for Microsoft beginning in late 2027; and its Norwegian footprint combines renewable power with large-scale Microsoft deployments.

The result increasingly resembles an attempt to build an AI-native hyperscaler from the ground up. Nscale argues that controlling infrastructure beneath the workload allows it to optimize scheduling, workload placement, GPU utilization, and ultimately cost per token in ways that pure infrastructure or software providers cannot. The Anyscale acquisition adds an open-source-led developer funnel, potentially giving Nscale a route from developer adoption to paid software,

cloud consumption, and ultimately infrastructure usage. Its U.K. domicile and European infrastructure footprint provide an additional sovereign-AI angle as governments and enterprises increasingly emphasize data residency and control.

Key Customers

Nscale's expansion is anchored by several unusually large customer commitments. Between September 2025 and April 2026, the company entered agreements with Microsoft, providing for payments of up to about $43.8 billion through 2033, excluding optional extensions. In August 2026, Nscale signed agreements with Anthropic, providing for up to about $44.6 billion of payments tied to Vera Rubin infrastructure at its Monarch campus. The company also signed a multiyear agreement with Figure that could ultimately deploy as many as 100,000 Vera Rubin GPUs, while OpenAI remains a major participant in Stargate Norway. As of August 31, Nscale reported about $103.4 billion of active and contracted total contract value under long-term take-or-pay agreements supporting roughly 461,000 active or contracted GPUs and a weighted-average life of 5.7 years.

With this scale comes substantial concentration risk. Nscale's largest customer accounted for 52% of revenue during the first half of 2026, down from 73% in 2025 and essentially all revenue in 2024. While the addition of Anthropic, Figure, and the Anyscale customer base should broaden the mix over time, the business in the near term remains highly dependent on a small number of very large deployments.

Financial Profile

According to the S-1, Nscale's revenue increased from $19.1 million in 2024 to $33.0 million in 2025 and reached $140.6 million in the first six months of 2026, up from $10.4 million in the prior-year period. Nscale had about $1.48 billion of cash on hand as of June 30, 2026, and expects to continue funding its rapid buildout through a combination of debt, equity, customer prepayments, strategic partnerships, and project-level financing. In September 2026, the company raised $3.36 billion of pre-IPO funding, which includes a $2.36 billion tranche led by Third Point with participation from Apollo Global Management (APO), Citadel, Hudson Bay Capital, Abu Dhabi Investment Council, and 8090 Industries, alongside a $1 billion commitment from Nvidia.

The financing base has expanded rapidly alongside customer commitments. In addition to more than $3.3 billion raised through series financings, Nscale has secured about $1.4 billion under its GPU financing facility, a $900 million revolving facility, up to $790 million for its Kvandal South data center, $332 million for Iceland, $1.85 billion for Ward County GPUs, $1.2 billion for North Carolina GPUs, and approximately $2.54 billion of initial-term vendor financing obligations with Dell. The increasingly project-specific structure reinforces the emerging neocloud financing model: secure long-duration, creditworthy customer demand and then finance the associated GPUs and infrastructure against those contracted cash flows.

Crusoe (Private)

History

Crusoe was founded in 2018 to monetize stranded natural gas that oil producers would otherwise flare (i.e., burn off at the wellhead because it could not economically be transported or sold). Crusoe was able to convert that gas into electricity on-site and initially used the power for Bitcoin mining, giving the company expertise in colocating large-scale compute with energy generation. In 2025, Crusoe agreed to sell its Bitcoin-mining business to NYDIG and focus entirely on vertically integrated AI infrastructure.

Product Offering

Crusoe's vertically integrated approach to the AI infrastructure stack includes energy sourcing and data-center development through GPU infrastructure and cloud software. On the physical infrastructure side, the company designs, engineers, and constructs hyperscale AI campuses optimized for high-density GPU deployments, including direct liquid-to-chip cooling, high-bandwidth networking, and dedicated power infrastructure. Its 1.2 GW Abilene, Texas, campus provides a useful proof point. Construction began in June 2024 and the first two buildings were energized within a year, with the eight-building campus designed to support hundreds of thousands of GPUs on a single network fabric.

Crusoe is also pushing data-center construction toward a more modular, manufacturing-based model. Through Crusoe Industries, the company manufactures power-distribution centers, switchgear, electrical enclosures, industrial controls, and other critical components in-house rather than relying on traditional construction supply chains. Its new 352,000-square-foot Spark Factory in Colorado is designed to mass-produce prefabricated Crusoe Spark AI data centers that integrate power, cooling, fire suppression, monitoring, and high-density GPU racks into standardized modules. Individual units can be combined into deployments ranging from hundreds of kilowatts to tens or hundreds of megawatts, with Crusoe targeting deployment in as little as three months.

Crusoe Cloud sits above this infrastructure with Nvidia and AMD GPU compute, virtual machines, storage, networking, orchestration, and observability through its Command Center. More recently, Crusoe has moved higher into the model-development and inference stack through its Intelligence Foundry. The platform includes serverless inference, dedicated self-service inference deployments, and serverless fine-tuning of open-weight models using LoRA-based workflows. Customers can fine-tune models on their own data, manage resulting checkpoints, and deploy those models directly onto Crusoe-managed inference infrastructure, while Crusoe handles engine selection, autoscaling, rate limiting, and other serving optimizations. This gives Crusoe a broader role across the AI lifecycle—from provisioning GPU clusters for training to customizing models and running them in production.

Differentiation

Crusoe's core advantage is its high degree of vertical integration across the AI infrastructure stack, spanning energy development, data-center engineering, manufacturing, physical infrastructure, and cloud services. Rather than treating a data center as a conventional construction project assembled from dozens of independent contractors and suppliers, Crusoe increasingly treats the facility as a manufactured product: major components are engineered and prefabricated off-site, then assembled into repeatable modules that can be customized for different power, cooling, and GPU configurations. By controlling more of the stack internally, Crusoe can optimize power, cooling, electrical systems, and compute around each generation of hardware while reducing reliance on constrained third-party supply chains. The company says its vertically integrated manufacturing process can deliver certain data-center systems in less than half the time of traditional alternatives.

This capability gives Crusoe two distinct deployment models. At one end, it can develop gigawatt-scale campuses such as Abilene for hyperscalers and frontier AI workloads; at the other end, Spark modules can place smaller GPU clusters closer to customers for inference, sovereign AI, or on-premises workloads. The combination should shorten deployment timelines, improve control over infrastructure economics, and allow Crusoe to capture more of the value created below the cloud layer. In that sense, Crusoe's differentiation lies as much in its ability to vertically integrate and industrialize the construction of AI infrastructure as in the cloud services it ultimately sells on top of it.

Key Customers

Crusoe's customer relationships are unique because the company operates both an AI cloud and a large-scale build-to-suit data-center business. On the infrastructure side, its flagship 1.2 GW Abilene campus is purpose-built for Oracle Cloud Infrastructure and forms the initial U.S. Stargate site, while a second adjacent 900 MW campus is being developed specifically for Microsoft. Crusoe also has contracts to develop two additional hyperscale campuses in Texas and another in Missouri, although those customers have not been publicly identified. Based on Bloomberg reporting, Meta is also a Crusoe customer. In total, Crusoe reported 4.9 GW of contracted AI infrastructure capacity as of June 2026 across its data-center developments and capacity supporting Crusoe Cloud. The two disclosed Abilene projects alone represent 2.1 GW, or roughly 43% of that total.

Crusoe Cloud serves a different customer base consisting primarily of AI-native companies and model developers, covering use-cases such as coding, voice AI, weather models, world models, and video generation. Customers include Cognition, Cartesia, Linum, Odyssey, SSI, Baseten, Together AI, and Figure. Windsurf, which was acquired by Cognition, used Crusoe Cloud to scale its AI coding platform to more than 800,000 active developers, while Linum trained its frontier video model on Crusoe's H200 infrastructure.

More recently, Crusoe Cloud has begun expanding beyond traditional AI-native customers into very large enterprise compute workloads, reportedly signing a roughly $13 billion, five-year contract to provide quantitative trading firm Jane Street with GPU clusters and AI infrastructure. In September 2026, the company signed a $65 million annual agreement with Thinking Machines Lab to power inference for the lab's open models on Crusoe Cloud.

This diversified customer mix gives Crusoe exposure to both large, long-duration hyperscale infrastructure contracts and higher-value AI cloud workloads, but the company does not disclose revenue, backlog, or customer concentration by segment, making it difficult to determine how much of the economics currently come from hyperscale development versus Crusoe Cloud.

Financial Profile

In early September 2026, Crusoe raised more than $3 billion in a series F financing at a post-money valuation of roughly $30 billion. The round was co-led by Atreides Management and Valor Equity Partners, with Mubadala Capital participating. By June 2026 the company had 4.9 GW of contracted AI infrastructure across data-center developments and Crusoe Cloud, with a stated development pipeline above 40 GW. Importantly, some of that capacity represents data-center development for hyperscale customers rather than capacity Crusoe itself will sell through Crusoe Cloud. Crusoe has also used third-party infrastructure capital to support this buildout, including a roughly $15 billion joint venture with Blue Owl and Primary Digital Infrastructure to fund Crusoe's 1.2 GW Abilene campus for Oracle/OpenAI. This reflects its hybrid business model, which straddles the lines between neocloud, data center developer, and infrastructure operator. As evidence of its expanding hybrid model, Crusoe disclosed over $100 million of Managed Inference ARR even though the service launched less than a year ago.

Lambda (Private)

History

Founded in 2012 by machine-learning engineers, Lambda began by building AI workstations and servers before expanding into GPU cloud infrastructure and large-scale AI clusters. This gives Lambda a somewhat different heritage than newer neoclouds: it grew out of serving AI researchers and developers rather than crypto mining or data-center development.

Product Offering
Lambda focuses on purpose-built AI infrastructure spanning on-demand GPU instances, 1-Click Clusters, and large dedicated single-tenant superclusters. Its latest deployments combine Nvidia Blackwell GPUs with high-speed InfiniBand or RoCE networking, tiered storage, and managed Kubernetes, while larger private-cloud configurations can scale from several thousand to more than 165,000 GPUs.

Differentiation
Lambda's differentiation is less about owning the full physical infrastructure stack and more about deep machine-learning expertise, simplicity, and large-scale cluster engineering. The company has spent more than a decade designing AI systems and increasingly positions itself around dedicated, single-tenant "AI factories" optimized specifically for training and inference rather than general-purpose cloud workloads. Lambda also retains a broad self-service developer funnel through its cloud business, giving it customer breadth at the smaller end of the market alongside increasingly large hyperscale contracts. In contrast to more vertically integrated players such as Crusoe, IREN, or Nscale, Lambda often partners with third-party data-center operators—including Aligned, EdgeConneX, Prime Data Centers, and ECL—to secure AI-ready capacity while focusing its differentiation in compute, networking, orchestration, and customer support.

Key Customers
Lambda's largest disclosed commercial relationship has historically been a multibillion-dollar, multiyear agreement with Microsoft announced in November 2025 covering tens of thousands of Nvidia GPUs. More recently, Anthropic reportedly signed a roughly $35 billion cloud-computing agreement with Lambda for about 350 MW of capacity at a new Texas data center, substantially expanding Lambda's exposure to large frontier-model customers. Lambda also serves a broad mix of AI labs, enterprises, start-ups, and researchers and has publicly highlighted customers such as Hudson River Trading and Voltron Data. In May 2026, Hudson River Trading selected Lambda to expand quantitative research and development as its internal infrastructure reached capacity. Lambda says it has more than 150,000 cloud users, providing considerably greater customer breadth at the smaller end of the market.

Financial Profile
Lambda raised more than $1.5 billion of series E equity in November 2025, following earlier equity and GPU-backed financings, and has increasingly shifted toward contract-backed debt as customer commitments have grown. In May 2026, it expanded an existing $275 million secured facility to $1 billion, and in August closed a separate $926 million Baa2-rated senior secured term loan B at SOFR + 300 basis points. The facility funds GPU infrastructure for a committed investment-grade customer, is secured by the underlying GPUs and related cash flows, and fully amortizes through 2030 in line with the customer contract and hardware life. Shortly thereafter, Lambda reportedly raised another $1 billion of private, short-dated debt to purchase Nvidia GPUs for a Microsoft deployment.

These transactions illustrate an increasingly repeatable financing model: secure a creditworthy customer contract first, then finance the associated GPU deployment against the underlying assets and contracted cash flows. Lambda's recently reported $35 billion agreement with Anthropic provides another substantial base of committed demand against which future infrastructure could be financed.

Fluidstack (Private)

History
Founded in 2017 at Oxford University by Gary Wu, César Maklary, Jamie Cox, and others, Fluidstack initially operated as a distributed compute marketplace, aggregating underutilized GPU capacity and connecting it with developers and AI companies. As demand shifted toward increasingly large dedicated AI clusters, the company moved up-stack into private cloud and eventually into full-scale infrastructure development. Fluidstack now describes itself as operating across the entire infrastructure stack—securing power, designing and building data centers, deploying compute, and operating both the hardware and software layers. The company relocated its global headquarters from London to New York in December 2025 as its business shifted toward large U.S. deployments.

Product Offering
Fluidstack provides dedicated AI infrastructure for large-scale training and inference, spanning bare-metal compute, managed Kubernetes, high-performance networking and storage, and customized single-tenant environments. While the company historically provided access primarily to Nvidia GPU clusters, its model has evolved toward designing and operating large-scale infrastructure around customer- or partner-financed accelerators as well. Fluidstack is now supporting both GPUs and custom silicon, including Google TPUs and Broadcom XPUs, while its software layer handles provisioning, orchestration, observability, health monitoring, and cluster management across large fleets. This gives the company greater hardware flexibility than neoclouds built exclusively around Nvidia GPU fleets.

Differentiation
Fluidstack's primary differentiation is speed of deployment and flexibility across the infrastructure stack. Unlike neoclouds that began with owned GPU fleets or crypto infrastructure, Fluidstack grew out of an aggregation model that required it to source compute and data center capacity across a broad partner ecosystem. It has since pushed substantially toward vertical integration and today claims the ability to acquire power, design and construct facilities, install compute, and bring gigawatt-scale capacity online materially faster than typical data center development timelines. Recently, Forbes reported that the company leased more than 1 million square feet near Glendale, Arizona, and is developing modular data-center infrastructure that can be manufactured and shipped to deployment sites, part of an effort to shorten construction timelines. The company also appears to be emerging as an important third-party deployment platform for Google TPUs, with Anthropic's TPU infrastructure representing the first publicly known large-scale operation of Google TPUs in customer-owned data centers outside Google itself.

Key Customers
Fluidstack's most important customer relationship is with Anthropic, which selected the company in November 2025 to deliver custom-built AI data centers in New York and Texas as part of Anthropic's planned $50 billion U.S. compute infrastructure investment, with initial sites coming online during 2026. In June 2026, Apollo, Blackstone, and Broadcom subsequently established a $35 billion financing platform supporting more than 1 GW of Anthropic compute expected to deploy at Fluidstack-based sites, underscoring the scale to which the relationship has grown. Fluidstack has also publicly identified Mistral AI, Character.AI, Poolside, Black Forest Labs, and Meta as customers, and by early 2025 said it had more than 100,000 GPUs under management.

Financial Profile
In January 2026, Fluidstack raised $830 million in a series A round at a $7.5 billion valuation, led by hedge fund Situational Awareness. Situational Awareness subsequently suffered a severe liquidity crisis in July 2026 and liquidated most of its leveraged public-equity portfolio, but reportedly retained its private holdings, including Fluidstack. Prior to this, Fluidstack had also been

discussing an additional financing of about $1 billion at a roughly $18 billion valuation, potentially co-led by Situational Awareness and Jane Street, although that transaction has not been confirmed as completed.

Like other neoclouds, Fluidstack is increasingly pairing customer commitments with asset- and project-backed financing rather than funding infrastructure entirely with equity. In April 2025, Macquarie agreed to finance GPUs deployed through Fluidstack using structures secured by the hardware itself. In addition, Fluidstack entered long-term leases for more than 360 MW at TeraWulf's Lake Mariner campus, with Google ultimately providing roughly $3.2 billion of credit support to help finance the projects. Fluidstack and TeraWulf subsequently formed a separate 168 MW Texas joint venture backed by another $1.3 billion Google guarantee.

Fluidstack increasingly resembles an infrastructure developer and financial orchestrator as much as a conventional GPU cloud—using long-duration customer commitments and investment-grade counterparties to mobilize third-party capital for scaled AI deployments. The Apollo and Blackstone–led $35 billion financing for more than 1 GW of custom AI compute scheduled to deploy at Fluidstack sites is supported by contracted Anthropic payments and technology-partner credit enhancement.

Emerging Neoclouds
Other neoclouds that we have come across in our research include (in alphabetical order):

- ***Boost Run (Nasdaq: BRUN).*** Boost Run is a publicly traded Nvidia-focused AI cloud offering GPU compute, managed Kubernetes, CPU, and storage, with roughly $1.9 billion of contracted revenue as of second quarter 2026 and capacity expected to expand to more than 100 MW.

- ***ChronoScale (Nasdaq: CHRN).*** ChronoScale was created from Applied Digital's (Nasdaq: APLD) former GPU cloud business and operates dedicated accelerated-compute infrastructure for AI workloads across Nvidia, AMD, and TPU-based capacity.

- ***Core Scientific (Nasdaq: CORZ).*** Formerly focused primarily on Bitcoin mining, Core Scientific is increasingly an AI data-center infrastructure provider, with roughly 590 MW contracted to CoreWeave under agreements representing more than $10 billion of potential revenue.

- ***Core42.*** A subsidiary of Abu Dhabi–based G42, Core42 operates a sovereign AI cloud supporting government, sovereign, and enterprise AI workloads globally.

- ***Firmus.*** Australia-based Firmus is emerging as a large vertically integrated AI infrastructure provider, with more than 900 MW of contracted capacity across Asia-Pacific and a recently announced multiyear OpenAI agreement for dedicated compute in Malaysia.

- ***Genesys Cloud.*** Germany-based Genesys Cloud is one of the earliest examples of a European GPU-first cloud, offering Nvidia-based training, inference, and HPC capacity from renewable-energy-powered data centers.

- ***GMI Cloud.*** GMI Cloud is a GPU-focused AI infrastructure provider operating across the U.S. and Asia, with a growing emphasis on dedicated Nvidia clusters, sovereign AI deployments, and large-scale infrastructure projects.

- ***Hyperstack.*** Hyperstack, operated by NexGen Cloud, is a European GPU cloud offering on-demand and dedicated Nvidia infrastructure for training, inference, and HPC workloads, supported by asset-backed financing for new GPU deployments.

- ***Lightning AI.*** PyTorch specialist Lightning AI combined with GPU infrastructure provider Voltage Park in January 2026, creating a full-stack AI cloud that pairs Lightning's developer and orchestration software with Voltage Park's large owned Nvidia GPU fleet; the combined company reported more than $500 million of ARR at the time of the merger and focuses specifically on model training.

- ***Massed Compute.*** Massed Compute is a smaller U.S. GPU cloud offering on-demand Nvidia GPUs, bare-metal servers, and dedicated clusters for AI and HPC workloads.

- ***QumulusAI (Nasdaq: QMLS).*** Recently direct-listed on Nasdaq, QumulusAI is an inference-first neocloud that identifies available or stranded pockets of power and data-center capacity, then rapidly deploys GPUs across distributed colocation sites to serve AI-native customers that need fast access to inference-optimized compute; management is targeting 18 MW of capacity and about $300 million of ARR by the end of 2026.

- ***Sharon AI (Nasdaq: SHAZ).*** Australia-based Sharon AI is rapidly scaling an Nvidia-focused neocloud, with about $8.8 billion of contracted value, including a six-year Nvidia agreement for up to 40,000 GB300 GPUs.

- ***STN.*** STN is a Chicago-based neocloud responsible for GPU One, a private AI cloud offering dedicated Nvidia-powered GPU clusters and managed environments for compliance-focused deployments of enterprise AI training and inference workloads.

- ***TensorWave.*** TensorWave differentiates itself as an AMD-exclusive AI cloud, operating large clusters of Instinct GPUs for training and inference and positioning itself as an alternative to Nvidia-centric neoclouds.

- ***Verda.*** Finland-based Verda (formerly DataCrunch) operates a vertically integrated AI cloud spanning data-center infrastructure, GPU compute, and platform software, with operations across more than 50 countries and plans to exceed 250 MW of capacity in 2027.

- ***Vertical Data (OTC: VDTA).*** Las Vegas–based Vertical Data is building an integrated AI infrastructure platform spanning GPU cloud capacity, financing, managed services, and data-center development, with a $29.5 million B200 cloud agreement and a newly announced letter of intent for a five-year, $192.1 million take-or-pay deployment of 1,024 Nvidia B300 GPUs for a global AI developer in Western Europe.

- ***WhiteFiber (Nasdaq: WYFI).*** Majority-owned by Bit Digital (Nasdaq: BTBT), WhiteFiber combines GPU cloud services with owned AI/HPC data-center infrastructure and represents Bit Digital's former AI compute business following its August 2025 IPO.

Neocloud-Adjacent Vendors (in Alphabetical Order)

- ***Akamai Cloud/Linode.*** Akamai Cloud is the cloud computing business built around Linode, the developer-focused IaaS provider that Akamai (Nasdaq: AKAM) acquired in 2022. The platform combines traditional cloud compute, storage, and other infrastructure services with GPU-based AI infrastructure, while leveraging Akamai's much larger global edge, content-delivery, and security footprint to differentiate around distributed applications and workloads that benefit from proximity to end-users. In September 2026, Akamai secured a landmark seven-year, $11.6 billion cloud infrastructure commitment from Anthropic—with potential expansion to over $20 billion—highlighting the growing scale of AI compute demand and validating Akamai's emergence as a major AI infrastructure provider beyond its legacy CDN and security businesses.

- ***Baseten.*** Baseten is a specialized AI infrastructure platform focused on deploying and serving models in production. The company abstracts the underlying GPU infrastructure across clouds and regions and provides model APIs, dedicated inference, training, autoscaling, observability, and its proprietary inference stack, allowing customers to run open-source or custom models without directly managing the underlying serving infrastructure.

- ***Cloudflare/Replicate.*** Replicate is an AI inference and model-deployment platform that allows developers to run more than 50,000 public and custom models through a simple API without managing the underlying GPU infrastructure. Its platform handles model packaging, deployment, autoscaling, and hardware selection, making it more of a software and inference layer than a traditional neocloud. Replicate was acquired by Cloudflare (NYSE: NET) in December 2025 and is being integrated into Workers AI, giving it access to Cloudflare's global network and broader developer platform while retaining the Replicate brand and APIs.

- ***DigitalOcean (NYSE: DOCN).*** DigitalOcean is a developer-focused cloud platform that has historically targeted start-ups, digital-native businesses, and SMBs with a simpler, lower-cost alternative to the hyperscalers. Originally centered on general-purpose compute, storage, databases, and Kubernetes, the company has expanded aggressively into AI through GPU infrastructure and an integrated "AI-native cloud" spanning accelerated compute, model inference, agents, and data services. Please see our research report from June 1, 2026 for an in-depth view of DigitalOcean's evolving strategy and positioning.

- ***Fireworks AI.*** Fireworks AI is an AI-native infrastructure platform focused primarily on high-performance model inference and customization. Rather than providing a broad general-purpose cloud, Fireworks operates an optimized inference stack spanning GPU infrastructure, custom kernels, memory management, quantization, and serving software, allowing customers to deploy and fine-tune open and custom models with an emphasis on throughput, latency, and cost efficiency at production scale.

- ***OpenRouter.*** OpenRouter is a model aggregation and routing platform that provides developers with a single API for accessing hundreds of models from providers such as OpenAI, Anthropic, Google, and leading open-model developers. Unlike an inference cloud that primarily serves models on its own GPU infrastructure, OpenRouter sits largely at the aggregation layer, routing requests across upstream inference providers based on model availability, price, and performance and giving developers a standardized interface across a fragmented model ecosystem. The company recently disclosed that it has 8 million global users, and on August 16, 2026, Bloomberg reported that Stripe finalized a deal to acquire OpenRouter for more than $7 billion.

- ***OVHcloud.*** OVHcloud is one of Europe's leading cloud infrastructure providers, offering a broad portfolio spanning public cloud, hosted private cloud, dedicated/bare-metal servers, and web hosting. The company is relatively vertically integrated—including ownership and operation of much of its data-center and server infrastructure—and differentiates through competitive economics, data sovereignty, interoperability, and European residency/compliance requirements, positioning it as a regional alternative to the U.S. hyperscalers.

- ***Runpod.*** Runpod is a developer-focused GPU cloud offering serverless and dedicated GPU infrastructure for AI training and inference, with particular strength among smaller developers, start-ups, and open-source AI users.

- ***Together AI.*** Together AI is a full-stack AI-native cloud focused on open and custom models, with offerings spanning accelerated GPU compute, serverless and dedicated inference, fine-tuning/model shaping, and large-scale pre-training. The company differentiates through

vertically optimized software—including proprietary kernels, quantization, and inference techniques—designed to improve GPU utilization, model performance, and serving economics, making it one of the closer AI-native analogs to a specialized cloud platform rather than simply an inference API provider.

- ***Vultr.*** Vultr is an independent cloud infrastructure provider positioned as a simpler and often lower-cost alternative to the hyperscalers, offering virtual machines, bare metal, storage, Kubernetes, managed databases, and a growing portfolio of GPU infrastructure. More recently, Vultr has increasingly focused on enterprise AI, combining Nvidia and AMD GPU capacity with its globally distributed cloud footprint to support both traditional cloud applications and AI training and inference workloads.

Key Risks

Slowdown in AI Demand
The neocloud market opportunity depends heavily on sustained growth in both enterprise and consumer AI usage. Slower enterprise adoption, weaker-than-expected monetization of AI applications, or softer consumer engagement could reduce inference and training demand, while algorithmic breakthroughs or lower-cost open-weight models could further reduce compute intensity. More restrictive AI safety regulation could also slow the pace of deployment by increasing compliance requirements, limiting certain use-cases, or raising the cost and time required to train and release advanced models.

Importantly, a significant portion of neocloud revenue to date—and likely much of the capacity already under long-term contract—has been tied directly or indirectly to frontier-model development and the enormous compute requirements of model training. If increasingly capable open-weight models begin to commoditize portions of the model layer, the economic returns from repeatedly training ever-larger proprietary frontier models could become less compelling. In that scenario, frontier labs could moderate training investment, particularly if incremental model-performance gains fail to generate commensurate improvements in monetization or customer willingness to pay.

The impact would ripple quickly through neocloud economics. Lower training demand, revenue per MW, or utilization would lengthen infrastructure payback periods, reduce returns on invested capital, pressure the residual value of older GPUs, and make it harder to service debt and long-term data center commitments. Because much of the industry is adding capacity against expectations of continued rapid demand growth, a meaningful slowdown in frontier-model investment could also create excess supply, further pressuring GPU pricing and amplifying the downside. Growing inference demand could offset some of this risk, but only if production AI workloads scale rapidly enough to absorb capacity initially built around large training clusters.

Supply-Side Constraints
Neocloud growth is ultimately constrained by access to GPUs, power, data-center capacity, networking equipment, memory, transformers, and other critical infrastructure. Shortages or deployment delays can push out revenue recognition, increase capex, and leave contracted demand waiting for capacity to come online. Community opposition is also becoming a more meaningful constraint, as concerns about power consumption, grid reliability, water usage, and local development drive tighter permitting, zoning restrictions, and project delays. At the same time, shortages of skilled labor—including electricians, engineers, construction workers, and data-center technicians—can further slow deployment and raise project costs. While supply tightness can support pricing, it also creates execution risk for operators unable to secure power, hardware, permits, and labor on schedule.

Narrowing Window of Opportunity?
Much of the hyperscalers' legacy data-center infrastructure was designed for general-purpose cloud workloads, with lower rack densities, air cooling, broad multitenancy, and networking optimized for a wide range of independent applications. The shift to AI initially left hyperscalers short of infrastructure capable of supporting dense, large-scale GPU clusters, creating an opening for neoclouds—particularly for model training. That advantage, however, is beginning to narrow. Microsoft, Google, and AWS are investing heavily in purpose-built AI infrastructure, including liquid cooling, high-density accelerator fabrics, and facilities designed around future generations of AI hardware.

As this capacity comes online, hyperscalers should become less dependent on third-party providers for incremental capacity, potentially curtailing neocloud demand and increasing competition for workloads. The long-term risk is that "purpose-built for AI" becomes table stakes rather than a durable source of differentiation, forcing neoclouds to compete increasingly on cost, software, deployment speed, and customer experience. The counterargument is that the AI infrastructure TAM may be large enough to support both hyperscalers and neoclouds without materially constraining growth for either.

Customer Concentration
Customer concentration remains a significant risk for leading neoclouds, whose large infrastructure buildouts are often anchored by a small number of hyperscalers and frontier AI labs. While these companies are working to diversify their revenue bases, near-term financial performance still depends heavily on a handful of multibillion-dollar contracts. The loss, downsizing, or nonrenewal of a major customer could leave operators with substantial underutilized capacity, pressuring the P&L statement.

Questionable Product-Level Differentiation
A key question for the neocloud model is how to distinguish and differentiate the underlying product across the various providers. Most platforms ultimately sell access to similar Nvidia accelerators, networking architectures, and core infrastructure services, making performance, availability, and price the primary decision factors for many customers. This could limit pricing power and increase churn if/when capacity becomes more widely available. Scale may prove to be a durable advantage, however, by improving hardware procurement, financing costs, fleet utilization, geographic reach, and the ability to support very large customer deployments. While software, support, orchestration, and managed services can further drive differentiation, it remains unclear whether these features will be sufficient to create strong customer lock-in or prevent the neocloud market from becoming increasingly price-competitive over time, especially as hyperscalers deploy more of their own capacity.

Capex/Leverage
The neocloud model requires substantial upfront investment in GPUs, networking, storage, and data-center capacity, much of which is financed with debt. Long-term customer commitments and prepayments can substantially reduce funding risk, but leverage remains meaningful for several operators and creates significant fixed obligations. Still, financing conditions are improving as the asset class matures and lenders gain confidence in contracted cash flows. CoreWeave's recent $8.5 billion nonrecourse facility, for example, included a fixed-rate tranche priced at roughly 5.9%, well below the company's roughly 10% borrowing costs around the time of its IPO. Even so, absolute leverage remains high and creates significant fixed obligations. If utilization, pricing, or customer demand falls short of expectations, neoclouds could be left servicing large debt and lease commitments against underutilized infrastructure, magnifying downside to equity holders.

Cash Burn

Rapid capacity expansion is producing significant negative free cash flow across much of the industry, as capex materially exceeds the cash generated from the existing installed base. This is not necessarily indicative of poor unit economics—much of the spending supports contracted future revenue—but it leaves neoclouds dependent on continued access to debt, equity, and customer prepayments to fund growth. A tighter capital environment or slowdown in contracted demand could force operators to curtail expansion or raise capital on less attractive terms.

Depreciating Assets

Neoclouds typically depreciate GPU assets over useful lives of roughly five to six years, creating a risk that rapid hardware advances could shorten their economic lives and reduce residual values before the end of the accounting schedule. So far, however, pricing has been much more resilient than expected: rental rates for prior-generation GPUs have remained firm, and older hardware continues to find productive use in inference and other less performance-sensitive workloads. The risk is that this durability proves temporary. If supply eventually catches up with demand and older GPUs experience sharper pricing declines or weaker utilization, returns on invested capital could come under pressure and operators may face accelerated depreciation or asset write-downs.

Circular Financing

One of the more controversial aspects of the AI buildout is the rise of circular financing structures, where chip vendors and hyperscalers invest in customers that, in turn, use that capital to purchase their infrastructure. Nvidia has been particularly active here with neocloud customers, including a $2.35 billion equity investment in CoreWeave, a $2 billion-plus investment in Nebius, a €500 million investment in Nscale, an up to $2.1 billion potential equity investment in IREN, and undisclosed investments in Crusoe and Lambda. Similar dynamics extend to frontier labs: Nvidia and Microsoft have committed $10 billion and $5 billion, respectively, to Anthropic, while OpenAI's February 2026 financing included $30 billion from Nvidia and $50 billion from Amazon, alongside a large incremental AWS commitment. These arrangements have drawn scrutiny over whether supplier financing is artificially accelerating reported demand, although the rapid growth in AI usage and frontier-lab revenue indicates the presence of substantial underlying end-demand.

The prices of the common stock of other public companies mentioned in this report follow:

Accenture (Market Perform)	$174.47
Advanced Micro Devices (Market Perform)	$607.87
Akamai (Outperform)	$108.99
Alphabet (Outperform)	$339.16
Amazon (Outperform)	$246.15
Applied Digital	$24.53
Arm Holdings (Outperform)	$283.33
Backblaze (Market Perform)	$13.54
Bentley Systems (Outperform)	$31.29
Bit Digital	$1.68
Blackstone	$114.56
Blue Owl Capital	$9.15
Boost Run	$16.24
Broadcom (Outperform)	$349.57
Caterpillar	$819.95
ChronoScale	$19.82
Cloudflare (Outperform)	$353.99
Core Scientific	$16.52
CoreWeave (Outperform)	$85.07
Dell Technologies	$543.43
Digital Realty Trust	$177.18
DigitalOcean (Outperform)	$136.29
Equinix	$1,011.07
IREN	$41.72
Meta (Outperform)	$715.62
Microsoft (Outperform)	$509.22
Moody's (Outperform)	$460.08
Nebius (Outperform)	$231.88
Nvidia (Outperform)	$228.86
Oracle (Outperform)	$132.60
QumulusAI	$7.16
Sharon AI	$52.91
Space Exploration Technologies (Outperform)	$145.47
TeraWulf (Outperform)	$15.12
Vertical Data	$5.94
White Fiber	$18.83

IMPORTANT DISCLOSURES

William Blair or an affiliate was a manager or co-manager of a public offering of equity securities for Backblaze, Inc. within the prior 12 months.

William Blair or an affiliate beneficially own or control (either directly or through its managed accounts) 1% or more of the equity securities of Everpure, Inc. as of the end of the month ending not more than 40 days from the date herein.

William Blair or an affiliate is a market maker in the security of Arista Networks, Inc., AvePoint, Inc., Backblaze, Inc., Box, Inc., Cisco Systems, Inc., Commvault Systems, Inc., CoreWeave, Inc., DigitalOcean Holdings, Inc., Dropbox, Inc., Everpure, Inc., F5, Inc., GitLab, Inc., JFrog, Ltd., Microsoft Corporation, MongoDB, Inc., N-able, Inc., Nebius Group N.V., NetApp, Inc., Nutanix, Inc., Oracle Corporation, Rubrik, Inc. and Varonis Systems, Inc.

William Blair or an affiliate expects to receive or intends to seek compensation for investment banking services from Arista Networks, Inc., AvePoint, Inc., Backblaze, Inc., Box, Inc., Cisco Systems, Inc., Commvault Systems, Inc., CoreWeave, Inc., DigitalOcean Holdings, Inc., Dropbox, Inc., Everpure, Inc., F5, Inc., GitLab, Inc., JFrog, Ltd., Microsoft Corporation, MongoDB, Inc., N-able, Inc., Nebius Group N.V., NetApp, Inc., Nutanix, Inc., Oracle Corporation, Rubrik, Inc. and Varonis Systems, Inc. or an affiliate within the next three months.

William Blair or an affiliate received compensation for investment banking services or non-investment-banking services from Backblaze, Inc. within the last 12 months. Backblaze, Inc. is or was, within the last 12 months, an investment banking client of William Blair & Company and/or one or more of its affiliates.

Officers and employees of William Blair or its affiliates (other than research analysts) may have a financial interest in the securities of Arista Networks, Inc., AvePoint, Inc., Backblaze, Inc., Box, Inc., Cisco Systems, Inc., Commvault Systems, Inc., CoreWeave, Inc., DigitalOcean Holdings, Inc., Dropbox, Inc., Everpure, Inc., F5, Inc., GitLab, Inc., JFrog, Ltd., Microsoft Corporation, MongoDB, Inc., N-able, Inc., Nebius Group N.V., NetApp, Inc., Nutanix, Inc., Oracle Corporation, Rubrik, Inc. and Varonis Systems, Inc.

This report is available in electronic form to registered users via R*Docs™ at https://williamblairlibrary.bluematrix.com or www.williamblair.com.

Please contact us at +1 312 236 1600 or consult https://www.williamblair.com/equity-research/coverage for all disclosures.

Jason Ader attests that 1) all of the views expressed in this research report accurately reflect his/her personal views about any and all of the securities and companies covered by this report, and 2) no part of his/her compensation was, is, or will be related, directly or indirectly, to the specific recommendations or views expressed by him/her in this report. We seek to update our research as appropriate. Other than certain periodical industry reports, the majority of reports are published at irregular intervals as deemed appropriate by the research analyst.

DOW JONES: 51349.90
S&P 500: 7670.84
NASDAQ: 26797.50

Additional information is available upon request.

Current Rating Distribution (as of September 30, 2026):

Coverage Universe	Percent	Inv. Banking Relationships *	Percent
Outperform (Buy)	71	Outperform (Buy)	16
Market Perform (Hold)	28	Market Perform (Hold)	4
Underperform (Sell)	1	Underperform (Sell)	0

*Percentage of companies in each rating category that are investment banking clients, defined as companies for which William Blair has received compensation for investment banking services within the past 12 months.

The compensation of the research analyst is based on a variety of factors, including performance of his or her stock recommendations; contributions to all of the firm's departments, including asset management, corporate finance, institutional sales, and retail brokerage; firm profitability; and competitive factors.

OTHER IMPORTANT DISCLOSURES

Stock ratings and valuation methodologies: William Blair & Company, L.L.C. uses a three-point system to rate stocks. Individual ratings reflect the expected performance of the stock relative to the broader market (generally the S&P 500, unless otherwise indicated) over the next 12 months. The assessment of expected performance is a function of near-, intermediate-, and long-term company fundamentals, industry outlook, confidence in earnings estimates, valuation (and our valuation methodology), and other factors. Outperform (O) - stock expected to outperform the broader market over the next 12 months; Market Perform (M) - stock expected to perform approximately in line with the broader market over the next 12 months; Underperform (U) - stock expected to underperform the broader market over the next 12 months; not rated (NR) - the stock is not currently rated. The valuation methodologies include (but are not limited to) price-to-earnings multiple (P/E), relative P/E (compared with the relevant market), P/E-to-growth-rate (PEG) ratio, market capitalization/revenue multiple, enterprise value/EBITDA ratio, discounted cash flow, and others. Stock ratings and valuation methodologies should not be used or relied upon as investment advice. Past performance is not necessarily a guide to future performance.

The ratings and valuation methodologies reflect the opinion of the individual analyst and are subject to change at any time.

Our salespeople, traders, and other professionals may provide oral or written market commentary, short-term trade ideas, or trading strategies-to our clients, prospective clients, and our trading desks-that are contrary to opinions expressed in this research report. Certain outstanding research reports may contain discussions or investment opinions relating to securities, financial instruments and/or issuers that are no longer current. Investing in securities involves risks. This report does not contain all the material information necessary for an investment decision. Always refer to the most recent report on a company or issuer. Our asset management and trading desks may make investment decisions that are inconsistent with recommendations or views expressed in this report. We will from time to time have long or short positions in, act as principal in, and buy or sell the securities referred to in this report. Our research is disseminated primarily electronically, and in some instances in printed form. Research is simultaneously available to all clients. This research report is for our clients only. No part of this material may be copied or duplicated in any form by any means or redistributed without the prior written consent of William Blair & Company, L.L.C.

This is not in any sense an offer or solicitation for the purchase or sale of a security or financial instrument.

The factual statements herein have been taken from sources we believe to be reliable, but such statements are made without any representation as to accuracy or completeness or otherwise, except with respect to any disclosures relative to William Blair or its research analysts. Opinions expressed are our own unless otherwise stated and are subject to change without notice. Prices shown are approximate.

This report or any portion hereof may not be copied, reprinted, sold, or redistributed or disclosed by the recipient to any third party, by content scraping or extraction, automated processing, or any other form or means, without the prior written consent of William Blair. Any unauthorized use is prohibited.

If the recipient received this research report pursuant to terms of service for, or a contract with William Blair for, the provision of research services for a separate fee, and in connection with the delivery of such research services we may be deemed to be acting as an investment adviser, then such investment adviser status relates, if at all, only to the recipient with whom we have contracted directly and does not extend beyond the delivery of this report (unless otherwise agreed specifically in writing). If such recipient uses these research services in connection with the sale or purchase of a security referred to herein, William Blair may act as principal for our own account or as riskless principal or agent for another party. William Blair is and continues to act solely as a broker-dealer in connection with the execution of any transactions, including transactions in any securities referred to herein.

For important disclosures, please visit our website at williamblair.com.

This material is distributed in the United Kingdom and the European Economic Area (EEA) by William Blair International, Ltd., authorised and regulated by the Financial Conduct Authority (FCA). William Blair International, Limited is a limited liability company registered in England and Wales with company number 03619027. This material is only directed and issued to persons regarded as Professional investors or equivalent in their home jurisdiction, or persons falling within articles 19 (5), 38, 47, and 49 of the Financial Services and Markets Act of 2000 (Financial Promotion) Order 2005 (all such persons being referred to as "relevant persons"). This document must not be acted on or relied on by persons who are not "relevant persons."

This report is being furnished in Brazil on a confidential basis and is addressed to the addressee personally, and for its sole benefit. This does not constitute an offer or solicitation for the purchase or sale of a security by any means that would constitute a public offering in Brazil under the regulations of the Brazilian Securities and Exchange Commission (*Comissão de Valores Mobiliários*) or an unauthorized distribution under Brazilian laws and regulations. The securities are authorized for trading on non-Brazilian securities markets, and this report and all the information herein is intended solely for professional investors (as defined by the applicable Brazilian regulation) who may only acquire these securities through a non-Brazilian account, with settlement outside Brazil in a non-Brazilian currency.

“William Blair” and “R*Docs” are registered trademarks of William Blair & Company, L.L.C. Copyright 2026, William Blair & Company, L.L.C. All rights reserved.

William Blair & Company, L.L.C. licenses and applies the SASB Materiality Map® and SICSTM in our work.

William Blair

Equity Research Directory

John Kreger, Partner Director of Research +1 312 364 8612

Kyle Harris, CFA, Partner Operations Manager +1 312 364 8230

CONSUMER

Sharon Zackfia, CFA, Partner +1 312 364 5386
Group Head–Consumer
Lifestyle and Leisure Brands, Restaurants, Automotive/E-commerce

Jon Andersen, CFA, Partner +1 312 364 8697
Consumer Products

Phillip Blee, CPA +1 312 801 7874
Home and Outdoor, Automotive Parts and Services, Discount and Convenience

Dylan Carden +1 312 801 7857
E-commerce, Specialty Retail

ECONOMICS

Richard de Chazal, CFA +44 20 7868 4489

ENERGY AND POWER TECHNOLOGIES

Jed Dorsheimer +1 617 235 7555
Group Head–Energy and Power Technologies
Generation, Efficiency, Storage

Neal Dingmann +1 312 801 7835
Oil and Gas, Rare Earth

Tim Mulrooney, Partner +1 312 364 8123
Energy and Environmental Services

Mark Shooter +1 617 235 7593
Generation, Storage

FINANCIAL SERVICES AND TECHNOLOGY

Adam Klauber, CFA, Partner +1 312 364 8232
Group Head–Financial Services and Technology
Financial Analytic Service Providers, Insurance Brokers, Property & Casualty Insurance

Andrew W. Jeffrey, CFA +1 415 796 6896
Fintech

Cristopher Kennedy, CFA +1 312 364 8596
Fintech, Specialty Finance

Jeff Schmitt +1 312 364 8106
Wealthtech, Wealth Management, Capital Markets Technology

GLOBAL SERVICES

Tim Mulrooney, Partner +1 312 364 8123
Group Head–Global Services
Commercial and Residential Services

Andrew Nicholas, CPA +1 312 364 8689
HR Technology, Information Services, Professional Services

Trevor Romeo, CFA +1 312 801 7854
Staffing, Waste and Recycling

HEALTHCARE

Biotechnology

Matt Phipps, Ph.D., Partner +1 312 364 8602
Group Head–Biotechnology

Lachlan Hanbury-Brown +1 312 364 8125

Andy T. Hsieh, Ph.D., Partner +1 312 364 5051

Myles R. Minter, Ph.D., Partner +1 617 235 7534

Healthcare Technology and Services

Ryan S. Daniels, CFA, Partner +1 312 364 8418
Group Head–Healthcare Technology and Services
Healthcare Technology, Healthcare Services

Jared Haase, CFA +1 312 364 5135
Healthcare Delivery, Pharmaceutical Outsourcing and Services

Steven Lichtman +1 646 561 9373
Medical Technology

Brandon Vazquez, CFA +1 212 237 2776
Dental, Animal Health, Medical Technology

Life Sciences

Matt Larew, Partner +1 312 801 7795
Life Science Tools, Bioprocessing, Healthcare Delivery

INDUSTRIALS

Brian Drab, CFA, Partner +1 312 364 8280
Co-Group Head–Industrials
Advanced Manufacturing, Industrial Technology

Ryan Merkel, CFA , Partner +1 312 364 8603
Co-Group Head–Industrials
Building Products, Specialty Distribution

Louie DiPalma, CFA, Partner +1 312 364 5437
Aerospace and Defense, Smart Infrastructure

Ross Sparenblek, CFA +1 312 364 8361
Diversified Industrials, Robotics, and Automation

TECHNOLOGY, MEDIA, AND COMMUNICATIONS

Jason Ader, CFA, Partner +1 617 235 7519
Co-Group Head–Technology, Media, and Communications
Infrastructure Software

Arjun Bhatia, Partner +1 312 364 5696
Co-Group Head–Technology, Media, and Communications
Software

Dylan Becker, CFA +1 312 364 8938
Software

Louie DiPalma, CFA, Partner +1 312 364 5437
Government Technology

Jonathan Ho, Partner +1 312 364 8276
Cybersecurity, Security Technology

Patrick McIlwee, CFA +1 312 364 8655
Software

Sebastien Naji +1 212 245 6508
Infrastructure Software, Semiconductor and Infrastructure Systems

Maggie Nolan, CPA, Partner +1 312 364 5090
IT Services

Ralph Schackart III, CFA, Partner +1 312 364 8753
Internet and Digital Media

Stephen Sheldon, CFA, CPA, Partner +1 312 364 5167
Vertical Technology – Real Estate, Education, Restaurant/Hospitality

EDITORIAL AND SUPERVISORY ANALYSTS

Steve Goldsmith, Head Editor and SA +1 312 364 8540
Katie Anderson, Editor and SA +44 20 7868 4451
Audrey Majors, Editor and SA +1 312 364 8992
Beth Pekol Porto, Editor and SA +1 312 364 8924
Lisa Zurcher, Editor and SA +44 20 7868 4549