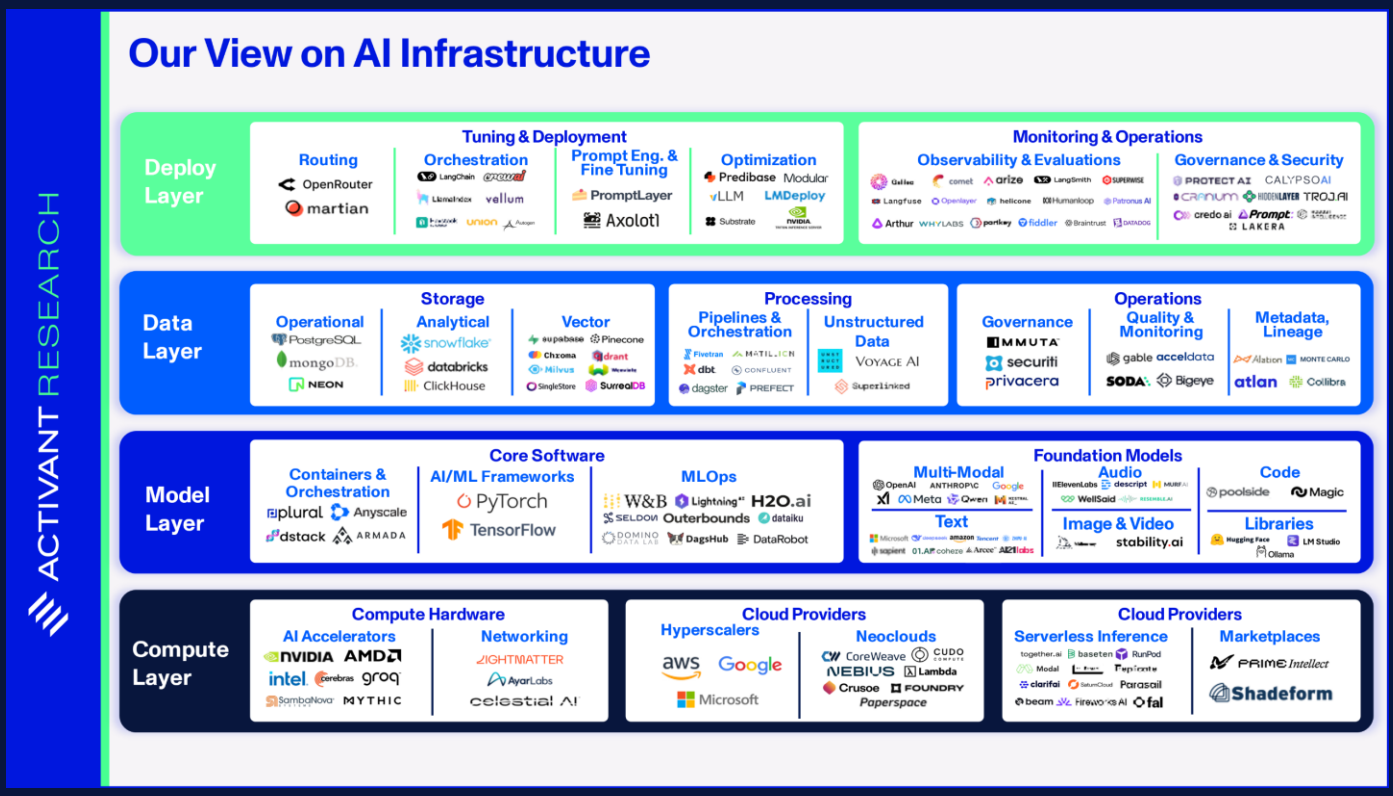
ACTIVANT RESEARCH

AI Infrastructure: Compute (3/4)

Neoclouds and the New Economics of AI Compute



Q1 2026

Torben Wiesbach, Jono Vickery

The Rise of Neoclouds: AI's Latest Infrastructure Innovation?

In our last article, we saw how hyperscale cloud providers built enduring competitive advantages. Yet, their broad service model and historically CPU-centric infrastructure left them poorly positioned for the AI boom of 2023 and created openings for new challengers. As demand for AI compute surged and outstripped supply, even the hyperscalers found themselves "compute starved." GPU instances on AWS, Azure, and Google became scarce and expensive — renting just four H100 instances for a month could cost nearly $300,000.[1] Startups and researchers faced waitlists, rate limits, and eye-watering bills. It was during this peak of explosive demand and constrained capacity that a new class of cloud provider emerged to fill the vacuum: neoclouds.

Neoclouds are the GPU-centric upstarts of the cloud world: purpose-built data centers delivering high-performance AI infrastructure through a GPUaaS business model. What began as a workaround for hyperscaler limitations has **rapidly evolved into a permanent fixture of the cloud landscape**. Today, neocloud providers are securing multi-billion-dollar contracts, deploying cutting-edge AI datacenters, and capturing a sizable share of AI compute supply.[2]

However, a central **question remains unresolved:** Are neoclouds building sustainable businesses on top of massive venture capital and debt or are they merely beneficiaries of a transient compute shortage that will fade once the scarcity eases?

In this third part of our series, we explore how these GPU-first "AI clouds" came to prominence, the playbooks and models that enabled their ability to scale rapidly, and the structural forces that will determine whether they endure.

From Cloud Oligopoly to AI Gold Rush

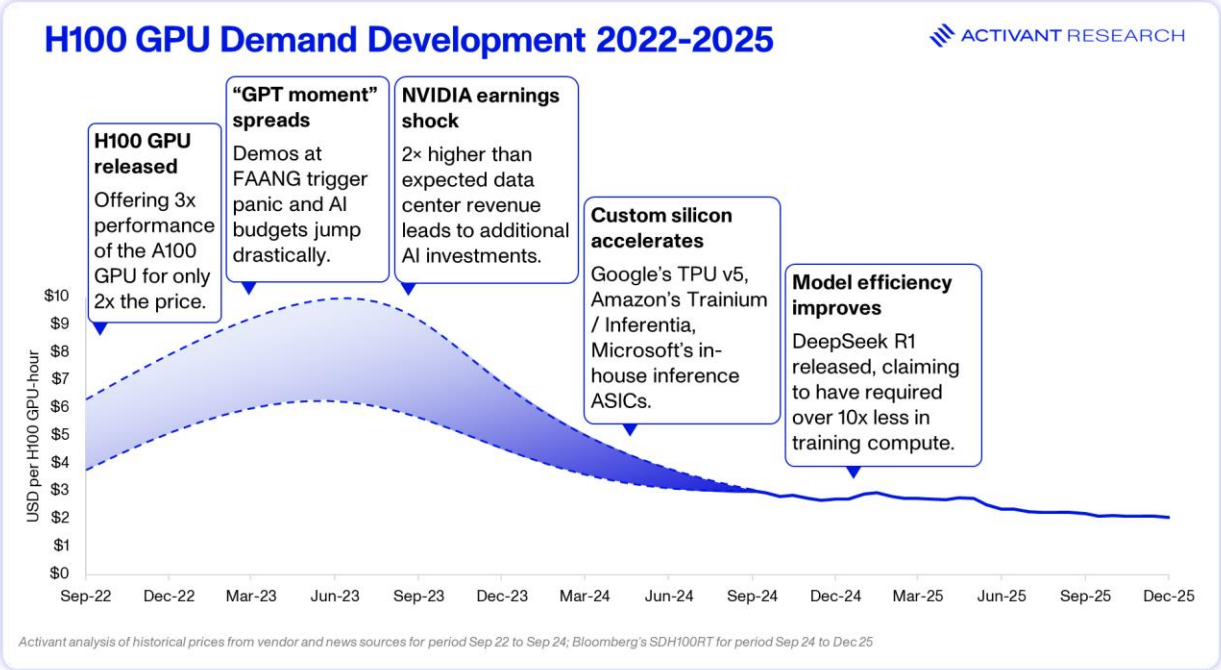
Cloud computing has been dominated by an oligopoly of hyperscalers for over a decade. These giants built vast, general-purpose infrastructure and responded to early AI trends by layering GPU instances, custom AI chips, and managed ML services onto their cloud platforms. Under normal circumstances, their scale advantages and integrated offerings would likely have been sufficient to defend their market position. But the generative AI explosion of 2023 was anything but normal. It triggered a gold rush for GPU compute and exposed limitations in the hyperscalers' business models.

In the span of months, applications such as ChatGPT sparked a race to train ever-larger models and deploy AI features, driving an insatiable demand for GPU power. Hyperscalers simply could not provision new GPU capacity fast enough. By late 2023, a significant share of Azure's latest H100 GPUs were being funneled to a single customer, OpenAI.[3] Hyperscalers split their attention and budgets between serving general cloud needs, developing alternative hardware, and

supporting a few mega-clients' AI efforts, leaving a long tail of startups and enterprises facing lengthy wait times and usage quotas. Even AI's elite felt the crunch:

> ***We're not trying to get [people] to use [ChatGPT] more**. Actually we'd love it if they use it less because **we don't have enough GPUs**.*
>
> ***Sam Altman, US Senate AI Hearing** (May 2023)* [4]

This compute scarcity was also evident in pricing. Hourly rates for H100 compute climbed to $8 or more, creating a powerful incentive for new players to bring additional supply to a "compute starved" market.

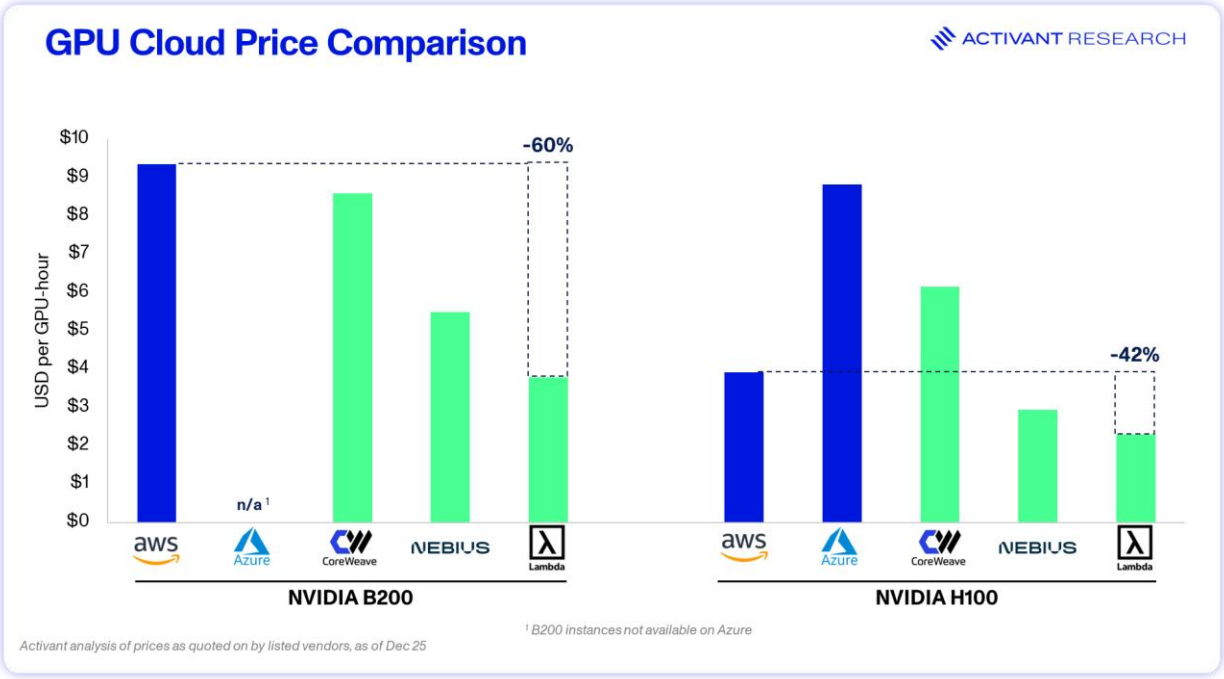


That opening was seized by the neoclouds, stepping in on short notice. Built entirely around high-performance GPUs, these operators were able to deliver scarce capacity and charge higher prices than the hyperscalers for doing so. Nvidia recognized the opportunity early and allocated supply to these emerging providers, helping to seed an alternative ecosystem of GPU buyers eager for their latest chips. By 2024, these newcomers had established themselves in the AI compute landscape. From 2022 to 2024, Nebius's revenue grew from $13.5M to $117.5M, while CoreWeave's revenue rose from $15.8M to $1.9B.[5] [6] All backed by substantial investor capitals and strong alliances with Nvidia.

The shift was perhaps best illustrated by Microsoft's move in mid-2023. Facing its own GPU shortfall on Azure, Microsoft signed a multibillion-dollar deal with CoreWeave to supply additional capacity.[7] In effect, a hyperscaler turned to neoclouds to meet demand that its own datacenters couldn't satisfy. What began as a tactical workaround solidified into a durable business model. By 2025, CoreWeave had grown to thirty-three data centers with more than 250.000 GPUs, allowing customers to launch thousands of H100 instances within minutes.[8]

It's All About Cost

So how have these upstart providers been able to thrive where hyperscalers struggled after the initial 2023 capacity shortage? Neoclouds understand they are very much selling a commodity: access to the same Nvidia chips everyone wants. Lacking brand power and entrenched relationships, they compete primarily on price-performance. The key to running a cost-effective AI data center lies in fundamentally rethinking infrastructure for the AI era. Neoclouds operate under a different playbook than the traditional cloud giants, prioritizing focus and efficiency over breadth. Unlike the big three, which sell every virtual machine (VM) size in every region, neoclouds typically offer only a handful of GPU server types. This reduces operational complexity, lowers overhead, and allows them to run on thinner margins while passing savings to customers.



To deliver maximum compute at minimum cost, neocloud operators have optimized facilities for the age of 100kW servers and heat-dense silicon. The result is rack densities and cooling solutions that go beyond traditional enterprise datacenters. Most use liquid-cooled racks supporting over 100 kW each, allowing more GPUs to be packed into less space. State-of-the-art liquid cooling

enables greater rack density, while improving energy efficiency and lowering operating costs per GPU. In practice, this allows neoclouds to put more computing power into each square foot. Some providers have purpose-built AI clusters, claiming approximately 20% better GPU throughput than alternative solutions, directly translating into better performance per dollar.[9]

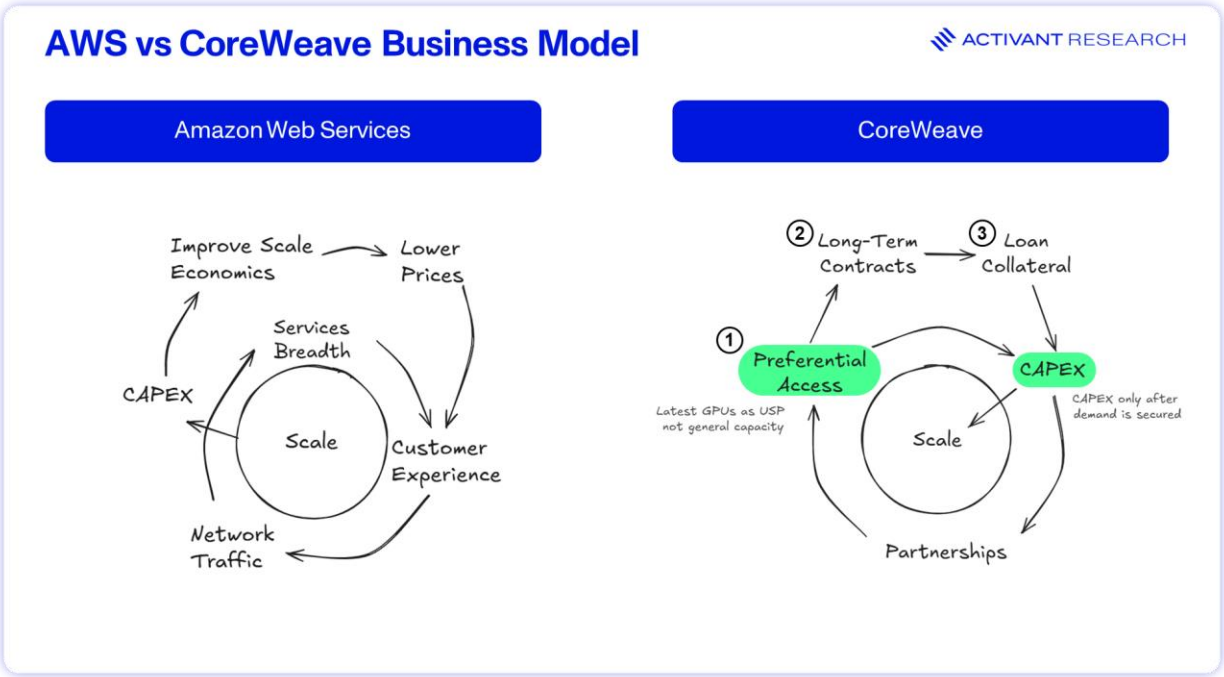
However, as neoclouds attempt to build more durable business models, they are increasingly orienting their strategies away from pure price competition and toward differentiated offerings with the potential to support higher, more sustainable margins over time:

- Some have tried to **turn their narrow focus into a feature by offering bespoke cloud solutions** for specific industries, regions, or use cases. Players like Salad operate a fully decentralized GPU cloud that harnesses idle consumer GPUs from volunteers, positioning it as an "Airbnb for GPUs".[10] TensorWave, has taken a hardware-centric approach building its cloud exclusively on non-Nvidia hardware. Its AMD Instinct GPUs serve customers looking for alternative price-performance or larger memory per GPU. TensorWave claims that with AMD's 192 GB-memory MI300X accelerators, some of its clients fine-tuned enormous 400+ billion parameter models fit on a single 8-GPU node — something not possible on 80 GB Nvidia cards without quantizing down to 4-bit precision.[11]
- Others have tried to **develop AI-first managed services layered on top of GPU compute**. Paperspace's (now part of DigitalOcean) Gradient platform is a leading example, positioning itself as an end-to-end platform to develop, train, and deploy AI models.[12] Similarly, CoreWeave's acquisition of Weights & Biases highlights a deliberate push further up the software stack, embedding experiment tracking, model management, and MLOps tooling directly alongside infrastructure.[13] Nebius has taken a comparable approach by offering managed Kubernetes, Postgres, and MLflow tooling, targeting teams that want a more integrated, production-ready AI platform rather than raw GPU capacity.[14] Collectively, these providers recognize that AI teams need an end-to-end development pipeline, not just raw GPU access.

While such models come with trade-offs, they underscore the creativity of this new market. The common thread is that neoclouds are playing by their own rules rather than trying to replicate the hyperscalers. They can't "out-Amazon" Amazon in general cloud services and, in our view, intensifying competition will push neoclouds to specialize and move beyond pure GPU leasing, competing less on price and more on differentiated software, platforms, and tightly integrated solutions. By doubling down on doing one thing exceptionally well, neoclouds aim to attract workloads that value unique capabilities, build stickier customer relationships, and mitigate the margin pressure inherent in undifferentiated infrastructure.

The Neocloud Playbook: High Stakes and Heavy Metal

To understand how the neoclouds have scaled so quickly, it helps to look at the underlying playbook that several of the biggest players follow, even though not every neocloud fits the mold precisely. Their model is built on a self-reinforcing cycle: gain access to the latest GPUs before anyone else, use that advantage to lock in customers through long-term contracts, and leverage those commitments to secure the financing needed for the expansion. This cycle has allowed neoclouds to build AI-optimized data centers at extraordinary speed, deepen customer dependence, and continually pull forward future growth.



Step 1: Gain Access to the Latest GPUs

Neoclouds gained an edge by being first in line for the latest GPUs. Many of them forged tight partnerships with Nvidia, translating into priority access to new hardware. CoreWeave, for example, reportedly secured a “most favored” supply position that allowed it to ramp H100 inventories faster than any competitor in 2023. By mid-year, it was fulfilling substantial H100 orders while many Azure and AWS customers were still waiting for availability.[15]

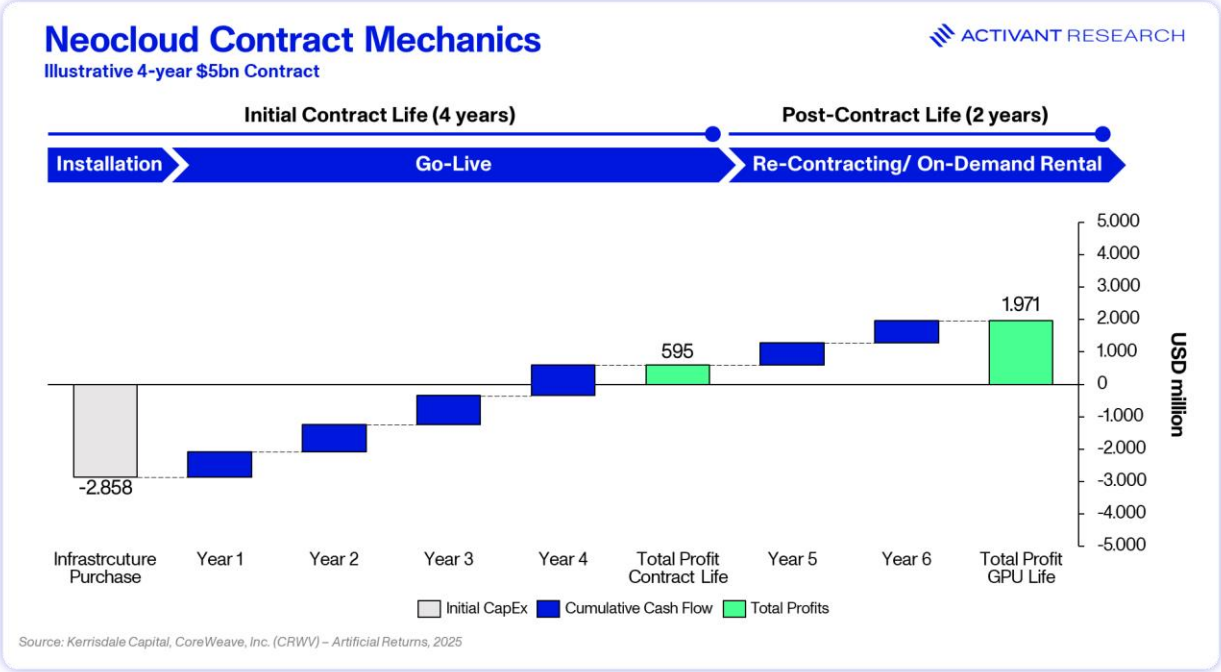
Nvidia’s motivation went well beyond short-term volume. A fragmented buyer base strengthens Nvidia’s pricing power. When demand is spread across many customers, no single buyer can exert meaningful pricing or contractual leverage. By contrast, a concentrated customer base shifts bargaining power toward hyperscalers—customers that are not only large and sophisticated, but also actively developing their own custom silicon. Over the long term, that concentration represents a structural disruption risk to Nvidia’s core business.

Seeding the neocloud ecosystem was therefore a strategic response to an existential threat. By creating independent buyers, Nvidia preserved pricing leverage, reduced dependence on hyperscalers, and ensured that access to cutting-edge silicon did not become monopolized.

Step 2: Lock in Customers Through Long-term Contracts

Many neoclouds still attract users through on-demand GPU rentals: spin up thousands of GPUs in minutes, pay by the hour, and shut everything down once the job is done. This model, however, exposes a fundamental economic challenge. Training demand comes in bursts, followed by periods of lower activity. For neoclouds operating billions of dollars of GPU infrastructure with large fixed-cost bases and significant debt service obligations, utilization becomes the deciding factor between profitability and loss. To mitigate this risk, neoclouds increasingly push customers toward capacity reservations. Contracts typically span months or years and guarantee payment for a fixed block of GPUs regardless of usage. For customers, the benefit is predictable access to scarce hardware. For neoclouds, it is predictable revenue.

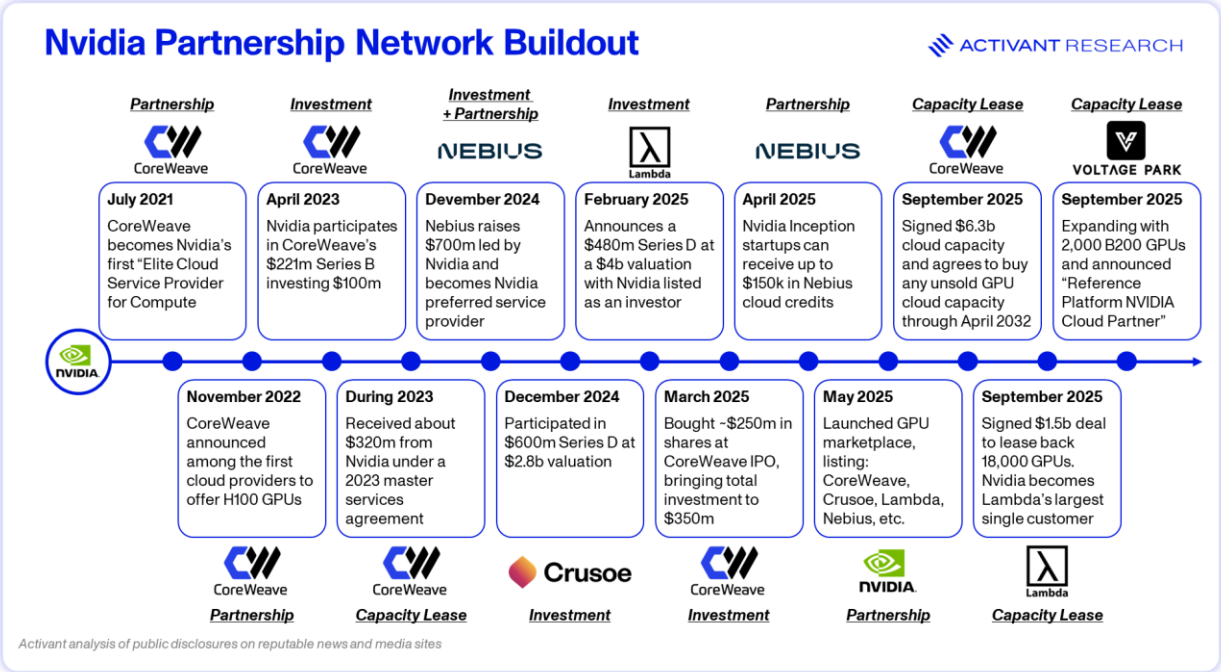
These agreements can be substantial. Microsoft, for example, has signed significant multi-year capacity agreements with Lambda, Nebius and CoreWeave.[16] [17] [18] By the end of 2025 Microsoft's neocloud spending surpassed $60bn.[19] These long term commitments form the financial backbone of the neocloud business model. Guaranteed multiyear revenue becomes collateral that neoclouds use to raise large debt facilities. The illustration below of a "typical" Neocloud contract demonstrates how this works in practice. Infrastructure is paid off during the contract period while on-demand rentals or contract extensions thereafter generate attractive returns.



Step 3: Secure the Financing Needed for the Expansion

The final step is converting contracted demand into financing that funds the next wave of expansion. Standing up an AI-optimized cloud is extraordinarily expensive. A single Nvidia H100 server can cost $25,000–$30,000 or more once networking and system components are factored in. Scaling to tens of thousands of GPUs, while also constructing datacenters, power and cooling systems, and networking fabrics, quickly pushes capital requirements into the billions. Unlike software startups that scale on rented cloud capacity, neoclouds must buy or lease hardware and infrastructure from day one. Yet despite the capital intensity and rising interest rates, the leading neoclouds have managed to raise staggering amounts of financing by tapping three primary sources.

1. A large portion of neocloud funding has come from **venture capital and GPU-backed private credit**. Investors have poured billions into providers like Lambda, Crusoe, CoreWeave and Voltage Park. But equity alone is not enough. What truly unlocked scale was the emergence of hardware-secured private credit. Neoclouds began treating racks of Nvidia GPUs like real estate assets, using them as collateral to raise debt. This financing approach emerged in early 2023, with large credit facilities structured directly around high-end GPU inventories.[20] The market initially validated the model, but Credit rating agencies have classified much of this debt as junk grade, underscoring the high risk of this unconventional form of collateral. Lenders, however, have remained willing and are betting that GPU demand will stay high enough to service debt obligations.
2. **Customer prepayments** represent a second financing lever. Multi-year capacity agreements often include significant upfront payments to secure guaranteed access to scarce GPUs. CoreWeave is the most visible example. As of September 2025, it carried $5.2bn in deferred revenue on its balance sheet.[21] These large prepayments effectively act as low-cost financing.
3. The final and most unconventional financing source is **Nvidia itself**. Beyond simple supply priority, Nvidia has gradually built a broader partnership network across multiple neoclouds, including equity investments, revenue-backed capacity agreements, cloud credits, and preferred-provider designations. In CoreWeave's case, Nvidia reportedly provided a revenue backstop by agreeing to buy unused GPU capacity if customer demand fell short.[22] This assurance materially reduces downside risk for lenders, allowing neoclouds to raise more debt at larger scales. While this arrangement is unusually "circular" (as pointed out in our article AI bubble signals intensifying), it serves Nvidia's strategic objective of sustaining demand for its chips and preventing under-utilization from dampening future orders.



Too Good to Be True?

The neocloud business model looks compelling, but several structural risks complicate the bullish narrative. Not all long-term contracts are as secure as they appear. Recent examples have shown that missed delivery timelines and service shortfalls can give customers flexibility to renegotiate or even exit portions of their commitments.[23] [24] In a highly leveraged model, even one major customer scaling back can trigger cascading challenges. And while utilization is guaranteed during a three- or four-year term, once GPUs roll off contract they must be re-leased into a market where prices have fallen sharply. H100 rentals dropped from about $8 per hour in 2023 to around $2 in 2025, while cheaper competitors like AMD and Tesla accelerators are gaining traction. That raises real doubts about profitability for older hardware or for clusters with shorter initial commitments.

Financing is another pressure point. Today's neocloud boom relies heavily on cheaply priced private credit, but this depends on continued lender confidence. If sentiment turns, borrowing costs will rise or capital may dry up entirely, making the model far harder to sustain. This concern is aggravated by uncertainty around the long-term customer base. The best customers — those who are willing to commit early, reserve large blocks of cutting-edge GPUs, and sign multi-year contracts — are also long-term competitors. Hyperscalers depend on neoclouds to bridge short-term GPU shortages and serve near-term training surges, but they are simultaneously accelerating efforts to build their own AI chips and high-performance fabrics. Microsoft, Google, Amazon, and OpenAI through its Stargate initiative, are all scaling their own chips and supercomputers with the clear intent to internalize a growing share of future inference and training

demand. As the supply crunch eases, these customers may not need neoclouds after all and the big providers will be left chasing small enterprise contracts rather than multi-billion-dollar hyperscaler contracts.

Underpinning everything is the assumption that compute demand will continue to grow exponentially. If scaling laws flatten or efficiency gains reduce the need for ever-larger clusters, demand may fall short of current projections. All of this suggests that while the neocloud flywheel works extremely well today, it depends on conditions that may prove more fragile than they currently appear.

Looking Ahead: Consolidation or Coexistence?

The central strategic question for neoclouds is whether their aggressive capacity build-out will ultimately pay off. The answer hinges on the structural realities shaping this market.

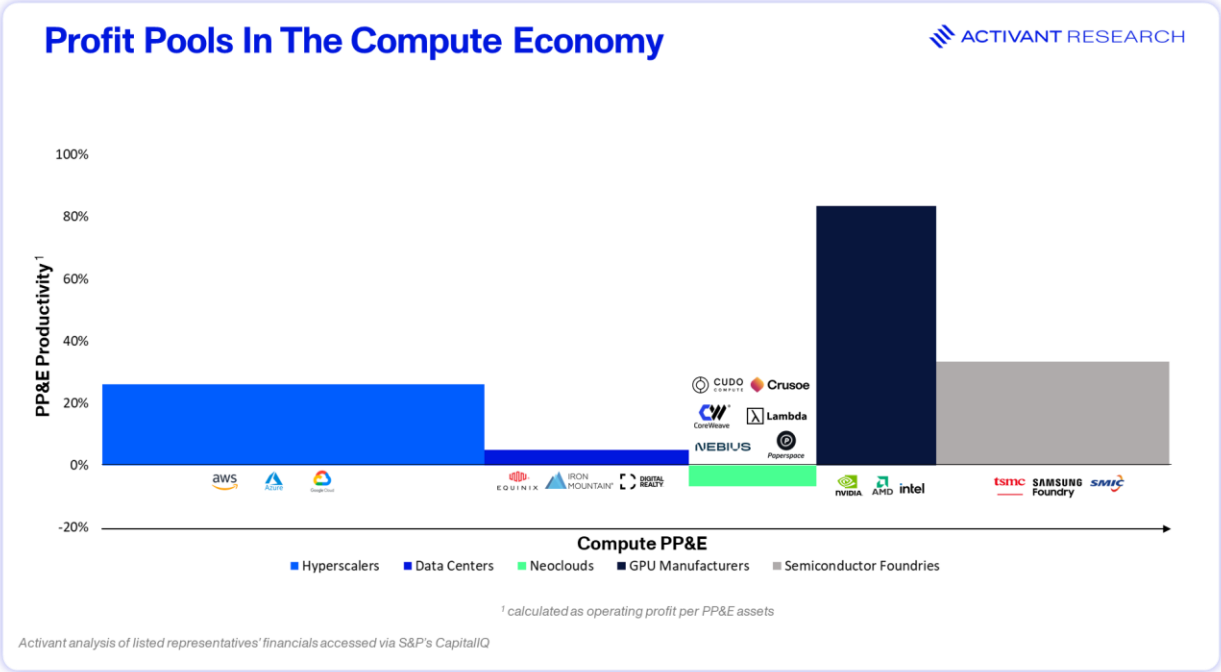
Securing long-term demand and locking in predictable workloads is what separates the winners from the players who will be forced into a race to the bottom. As supply expands through new fabs, more GPUs coming online, and alternative accelerators, idle capacity quickly becomes toxic. We are already seeing early signs of this dynamic in the sharp decline in GPU rental prices over the past two years. Any neocloud unable to keep its fleet fully booked risks being squeezed by falling prices while still servicing debt on expensive hardware.

Bare metal alone is a commodity. Sustainable margins require real differentiation. While many neoclouds present themselves as full-stack platforms, customer sentiment often tells a different story. Without genuinely valuable orchestration, tooling, or higher-level services (similar to how hyperscalers layer software on top of generic compute), GPU providers will inevitably compete on price. The long-term sustainability of any neocloud depends on building a platform that customers adopt and that meaningfully increases switching costs.

The business model remains tightly coupled to hardware access and financing, and most profits sit downstream. The economics of neoclouds depend heavily on how cheaply they can procure leading-edge GPUs and structure their financing. Nvidia, seeking to diversify demand beyond the hyperscalers, is actively fueling an alternative ecosystem. Yet most of the value in the stack still accrues to Nvidia and the foundries. It will be important for neoclouds to build sufficient scale and bargaining power to set terms and not be left competing over high-priced inventory that hyperscalers opted not to purchase.

Customer concentration is high, and the largest customers are also long-term competitors. While hyperscalers remain the biggest buyers of neocloud capacity today, they are simultaneously investing heavily in their own AI chips and fabrics with the clear intention of internalizing a large share of inference and potentially also training workloads over time. In many ways, hyperscalers

have also benefited from the rise of neoclouds: they were able to offload short-term demand spikes and let neoclouds take the upfront risk of securing scarce Nvidia GPUs, effectively using them as a bridge during a two to three-year window while building out their own silicon. While this development mainly concerns inference workloads, neoclouds will need to invest heavily in the latest tech to maintain an edge on the training front.



In our view, the most likely outcome is a consolidation of the neoclouds role within the AI infrastructure stack. The AI build-out is real and durable, and demand for accelerated compute will continue to grow over the coming years. As a result, the capacity neoclouds have built will not prove redundant. Instead, it anchors them as an enduring layer of the ecosystem, particularly for large-scale training and specialized workloads that require flexibility, speed of deployment, or hardware optionality.

However, the economics of bare-metal GPU rental present a challenging risk profile. The profit pools in AI infrastructure concentrate elsewhere in the stack and will influence how the market evolves from here. Capital intensity, pricing pressure, and customer bargaining power compress returns at the pure infrastructure layer. From their current starting point as hardware lessors, not all neoclouds will evolve into high-margin, high-ROIC businesses. The inevitable outcome is consolidation. As competition intensifies and hyperscalers reassert their scale advantages, the market will narrow to a smaller number of neoclouds with sufficient scale, balance-sheet resilience, and platform integration. Some players will become redundant, others will merge, and a handful will emerge as “hyper AI clouds”.

The shake-out will not determine whether neoclouds exist, but what they become. Only a handful will successfully move up the stack and capture more attractive profit pools, while others remain trapped in capital-heavy, low-return infrastructure roles. The winners of the next phase will be those that translate their early infrastructure advantage into differentiated software, platforms, and managed services. Effectively, this is the pole position where pricing power, switching costs, and sustainable returns reside.

[1] Uptime Institute, Neoclouds: a cost-effective AI infrastructure alternative, 2025

[2] CoreWeave, CoreWeave Expands Agreement with OpenAI by up to $6.5B, 2025

[3] Data Gravity, 2023 Year in Review: The Great GPU Shortage and the GPU Rich/Poor, 2024

[4] CBS News, OpenAI CEO Sam Altman testifies at Senate artificial intelligence hearing, 2023

[5] Yahoo!Finance, Nebius Group N.V., 2025

[6] Yahoo!Finance, CoreWeave, Inc. (CRWV), 2025

[7] CNBC, Microsoft signs deal for A.I. computing power with Nvidia-backed CoreWeave, 2023

[8] CoreWeave, Company Website, Accessed December 2025

[9] CoreWeave, Industry-Leading AI Infrastructure

[10] Salad, Company Website, Accessed December 2025

[11] TensorWave, Company Website, Accessed December 2025

[12] Paperspace, Company Website, Accessed December 2025

[13] CoreWeave, CoreWeave Completes Acquisition of Weights & Biases, 2025

[14] Nebius, Managed Service for Kubernetes, 2025

[15] Sacra, CoreWeave Equity Research, 2025

[16] Financial Times, Data centre operator CoreWeave lays groundwork for IPO, 2025

[17] CNBC, Lambda, Microsoft agree to multibillion-dollar AI infrastructure deal with Nvidia chips, 2025

[18] Nebius, Nebius announces multi-billion dollar agreement with Microsoft for AI infrastructure, 2025

[19] Bloomberg, Microsoft Neocloud Deals Cross $60 Billion in AI Spending Frenzy, 2025

[20] Reuters, CoreWeave raises $2.3 billion in debt collateralized by Nvidia chips, 2023

[21] CoreWeave, Q3 2025 Results, 2025

[22] Reuters, CoreWeave, Nvidia sign $6.3 billion cloud computing capacity order, 2025

[23] Financial Times, Microsoft drops some CoreWeave services ahead of $35bn IPO, 2025

[24] Semafor, Microsoft chose not to exercise $12 billion Coreweave option, 2025