FUTURIOM
FUTURE OF CLOUD TECH

Cloud Market Trend Report

AI, GPU Clouds, and Neoclouds in the Age of Inference

October 2025

Sponsored by:

f5®

Key Findings:

- **AI datacenter spending is enormous and shows no signs of slowing.** It's likely that more than $1 trillion has been committed to AI datacenter buildouts.
- **Neoclouds—clouds purpose-built for AI—are grabbing headlines in the datacenter industry.** Most prominent in this group is CoreWeave, which went public in 2025 and now has a market cap exceeding $60 billion.
- **NVIDIA is providing fuel by funding neoclouds, which are also NVIDIA customers.** It's legal, and it's in NVIDIA's best interest, but the circular nature of the NVIDIA AI economy is raising uncomfortable bubble questions.
- **Alternative clouds have a promising niche in enterprise AI.** These clouds, which tout lower prices than the likes of AWS, provide both GPU clusters and enterprise cloud computing services. Vultr in particular is finding traction with this combination.
- **The need for power is driving AI cloud expansion.** The heated competition to build more AI capacity is primarily a race to lock down power guarantees.
- **This new market is driving the evolution of a new infrastructure stack with needs in networking, storage, and security.** Each layer of the infrastructure has new demands to support the scale and complexity of hosting AI clouds.
- **Debt is starting to outweigh equity in neocloud funding.** The scale is tipped by companies like CoreWeave. The debt usually goes toward buying GPUs, but for those neoclouds that are building datacenters themselves, the debt also finances construction projects.
- **Hyperscalers are implementing diversification strategies to offset risks.** The market abhors a monopoly, which in this case includes not only NVIDIA GPUs but also the rest of the NVIDIA software and hardware stack. Chip and networking companies are courting neoclouds with alternatives—and hyperscalers are diversifying suppliers.

- **Companies mentioned in this report:** Accton, AMD, Arista, AWS, Cerebras, Cisco, CoreWeave, Crusoe, DDN, Dell, DigitalOcean, DriveNets, Eclypsium, Fluidstack, Fortanix, Google, Groq, Hedgehog, HPE, Juniper, Lambda Labs, Nebius, Nscale, NVIDIA, OpenAI, Runpod, Shakti Cloud, Supermicro, TensorWave, Together AI, VAST Data, Voltage Park, Vultr, WEKA, and WhiteFiber.

Enterprise AI delivery and security

Solve AI deployment and security challenges to unlock emerging opportunities.

- Deliver data efficiently for AI training and inference
- Scale AI factories with optimized GPU utilization
- Secure AI deployments with end-to-end protection against app and model threats
- Connect data and applications to scale distributed AI deployments

Learn more about how we can help you with these challenges, visit f5.com/ai

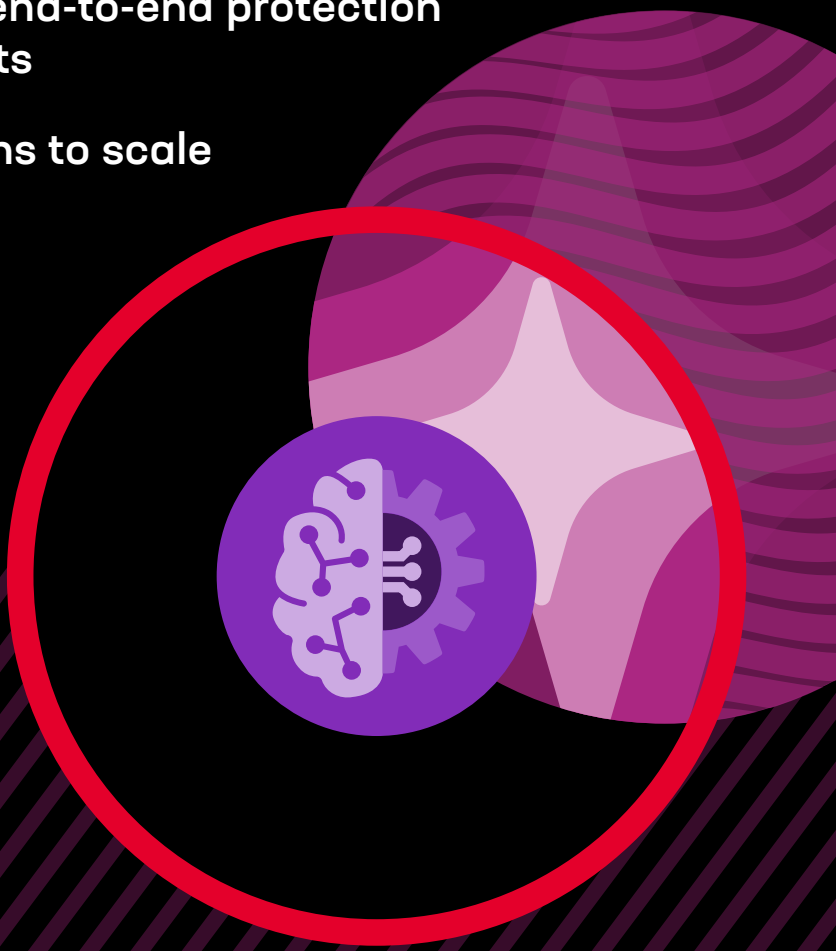


Table of Contents

1. Introduction: A Seismic Shift for Datacenters 4

About Our Methodology 6

2. What's in a Neocloud? The Rise of the GPU Cloud 7

The Neoclouds that Really Matter 8
Frontier AI Models Pay the Bills, but Small Jobs Are Welcome Too 9
It Really Is All About Power 10
Massive Funding, Massive Debt 10
The Circle of Life (Is OpenAI Too Big to Fail?) 11
What About Differentiation? 12

3. Hyperscaler and GPU Cloud Strategies 13

Hyperscalers and Other GPU Clouds 13
Alt Clouds Are In on GPUs, Too 14
Hyperscaler Cloud Differentiation 14

4. GPU Cloud Infrastructure: Vendor Landscape 15

GPUs: NVIDIA Rules, but Customers Want Alternatives 15
Networking: Applying Lessons from Hyperscalers 16
What About Security? 18
Storage Needs to Be More than Just "Storage" 19
OEMs Can Think Big (Rack-Scale) 20
Liquid Cooling Finally Has Its Day 20

5. Conclusion: Potential Risks and Rewards of the GPU Cloud Market 22

The Depreciation Question 22
Have the Hyperscalers Outsourced Risk? 23
Future Evolution of Neoclouds, Hyperscalers, and Alt Clouds 23

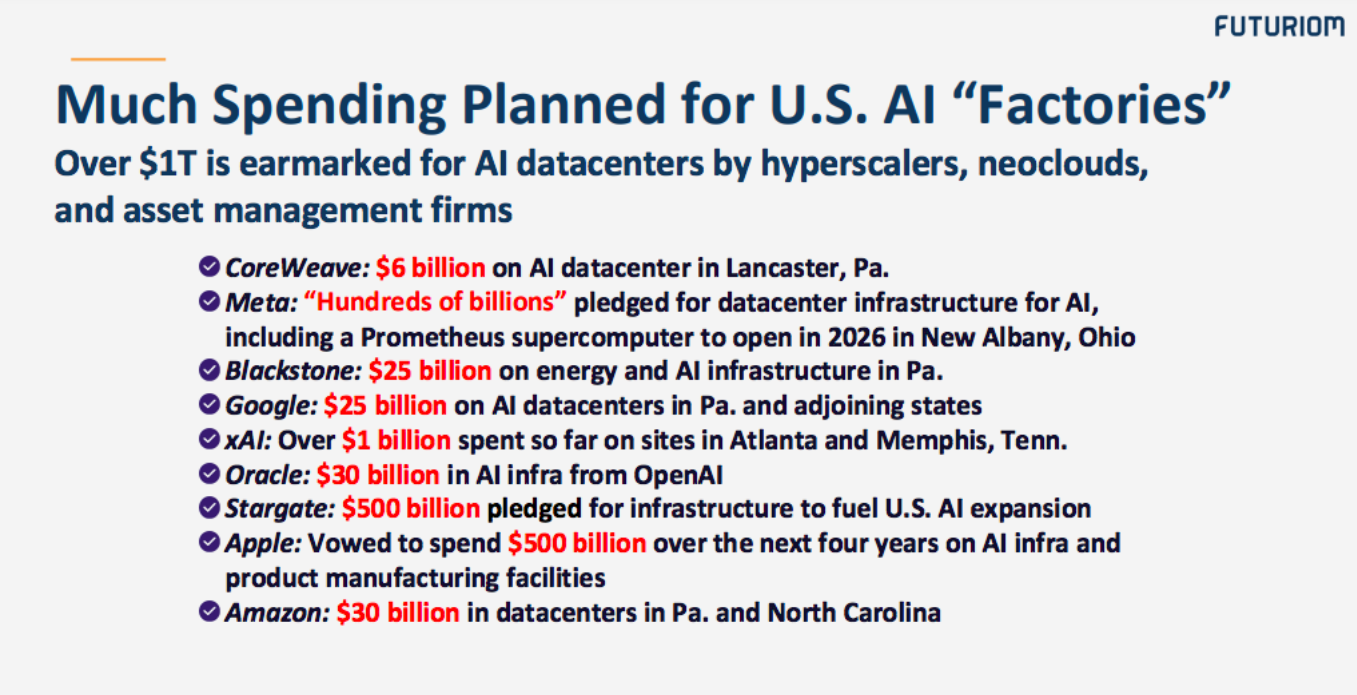
Appendix: Selected AI and GPU Cloud Ecosystem Players to Watch 24

1. Introduction: A Seismic Shift for Datacenters

The AI boom has generated big dreams in the tech and finance worlds, and that's most blatantly on display when it comes to clouds and datacenters. What we have on our hands is a worldwide arms race to build out AI clouds of all stripes, an endeavor which requires lots of resources—land, water, power—and of course money.

This is the domain of GPU clouds as well as neoclouds, which are GPU-laden clouds dedicated to AI training and inference. They're not the only clouds deploying GPUs, but they've grabbed the spotlight due to their audacious scale. The splashiest neocloud projects are campuses costing billions and designed to consume gigawatts of power.

And yet, it's still not enough, according to AI company executives. Neocloud activity intensified in 2025, with bigger datacenter plans, enormous funding rounds, and cavernous debt financing. As we noted in a recent Cloud Tracker Pro report, in total more than $1 trillion has been pledged for AI datacenters over the next few years. See below for a summary of some recent AI factory projects.



Some of this money will be internal investment (Apple, xAI), but most of it, including likely all $500 billion of the Stargate figure, is fueling the AI cloud boom, and the pace is accelerating. Nine-figure deals seem to get announced nearly every day, meaning by the time this report goes to market, some of it will already be out of date.

The surge has also propelled neocloud-related Initial Public Offerings (IPOs) by CoreWeave and WhiteFiber, providing public capital to pick up the pace of buildouts. CoreWeave projects it will spend $20 billion to $23 billion in capex this year, up roughly three times what it spent last year.

Other publicly traded neoclouds show similar increases. Nebius projects $2 billion in capex for 2025, about 2.5 times what it spent last year. Oracle expects fiscal 2026 capex (for the year ending May 31) of around $35 billion, up from $21.2 billion the previous year; most of that money will go toward equipping rapid cloud expansions.

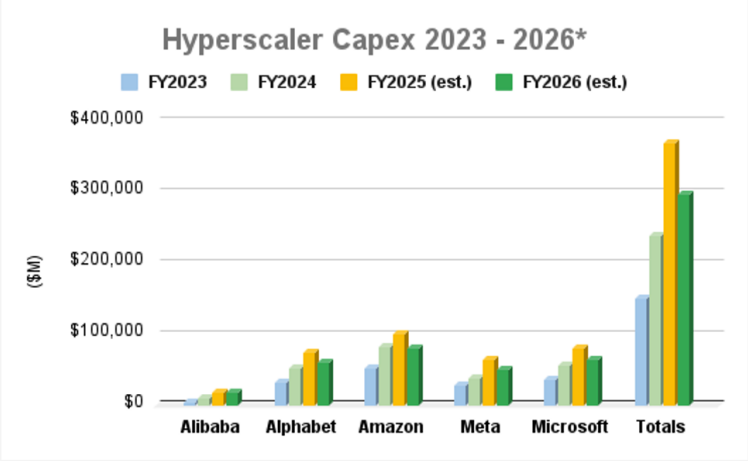
Then there are the hyperscale public clouds, which have joined the race to build AI factories. As the chart below shows, hyperscalers (ex-neoclouds) expect to spend 55% more capex than in 2024. Despite the projections, we are expecting these levels to slow.

FUTURIOM

Hyperscaler Spending Grows

Hyperscalers plan about 55% more capex this year than in 2024. But 2026 could see less investment.

*Based on Microsoft's retreat from several large-scale datacenter projects this year, Futuriom projects an approximate 20% reduction in hyperscaler spending in 2026.



You might ask: Why would the pace of datacenters slow? We are simply expecting that after the initial surge, the largest cloud providers might revert to their mean of capex spending (as a percentage of revenue). The spending has reduced profitability. With the exception of Oracle, which has recently been increasing its projections, it appears that the major hyperscalers are also outsourcing some of the risk of building more GPU clouds to their partners. Incremental AI spending has been focused on new projects such as the neoclouds and OpenAI's Stargate.

Finally, the neocloud business itself continues to change. The companies offering GPU clusters now want to oversee datacenter construction themselves. OpenAI, a major tenant of the

neoclouds, says it will begin building its own datacenters, too. "Neocloud" tends to refer to the cloud service, but it's starting to get associated with being a datacenter landlord as well.

The rules are still being written for this sector of the cloud market. Will 2025 prove to be an anomaly, or is it the start of a new normal?

This report presents a snapshot of how AI clouds, GPU clouds, and neoclouds fit together, examining the sector's history and the direction of recent deals and financing. We also look at the GPU efforts of the public clouds, including hyperscalers. Finally, we consider the infrastructure ecosystem targeting AI datacenters, where storage and networking take on heightened importance.

About Our Methodology

Futuriom is focused on what's happening with technology in the real world. We spend many hours talking to technology users and vendors and researching deployments. When we can, we consult user data rather than marketing documents. Where possible, we cite sources.

The basis of our research is our weekly, primary conversations with end users as well as observations of what technical leaders are doing in the community. Additional sources are public survey data, both our own as well as data from other reputable sources.

As always, we also spend a lot of time with startups, investigating the companies we feel currently are special to watch. As in all our Cloud Market Trend reports, we have included in this report all of the major players in the market, regardless of whether or not they sponsored the report. Our goal is to present an independent view of the market. Our free reports are supported by sponsorships, in which sponsors receive additional benefits such as distribution rights, an ad in the report, and demographic data about report downloads.

2. What's in a Neocloud? The Rise of the GPU Cloud

There's no denying that new types of clouds and datacenters have emerged for AI. Neoclouds sprang into the AI conversation to fulfill additional GPU capacity from both other cloud providers as well as enterprise customers. Some neoclouds have evolved from other businesses, such as cryptomining operations or edge compute services. The business models continue to evolve. Let's review how this all happened.

Most neoclouds started by offering GPU capacity to cryptocurrency miners. As that market cooled, some of them, notably CoreWeave, began repurposing those GPUs toward AI processing. Then OpenAI launched ChatGPT in November 2022, launching the current AI craze and providing a ready market for those GPUs. A subsequent run on NVIDIA chips led to shortages, prompting even more cryptomining operations to pivot to AI.

Some of the larger neoclouds did not evolve this way. Lambda originally targeted the self-driving vehicle market. And Nebius began life as the infrastructure behind Russian search engine Yandex. After Russia's invasion of Ukraine, NASDAQ announced Yandex's delisting. The company responded by incorporating Nebius in The Netherlands and divesting Yandex. Nebius continues to be publicly traded on NASDAQ.

Here's our view on the current market segmentation:

Neoclouds: These are the new breed of cloud, specializing in hosting AI and high-performance computing. They differ from other public clouds in their single-mindedness, emphasizing availability of GPU clusters and performance suitable for AI training and inference.

Hyperscalers: GPU cloud services are available from conventional sources, particularly the hyperscale public clouds such as AWS. Their advantage is the ability to more easily integrate AI work with the enterprise's existing cloud environment. That said, they have also begun building massive campuses dedicated to AI, just like the neoclouds.

Hybrid AI clouds: There's a group of alternative clouds (alt clouds)—economical alternatives to hyperscale public clouds—that offer a mix of GPU clusters and conventional public cloud services. They like to differentiate from neoclouds in terms of services and from hyperscalers in terms of price. They are unlikely to build their own datacenters dedicated to AI.

The Neoclouds that Really Matter

Many estimates put the number of neoclouds at around 100, but only a subset seem likely to have an impact. Below is a chart of the emerging players we expect to be leaders in the industry.

Provider	Size	Chip Suppliers	Year Founded	Total Funding (Includes Debt)
Core42	>35,000 GPUs in the UAE, plus leased capacity in the U.S. and Europe; about 75 MW, >200 MW planned short term	AMD, Cerebras, NVIDIA, Qualcomm	2023	>$1.5 billion
CoreWeave	>250,000 GPUs in 32 datacenters; >2.2 GW	NVIDIA	2017	CoreWeave is public (Nasdaq: CRWV) but raised >$10 billion prior to IPO
Crusoe	Claims 946,000 GPUs; about 200 MW; up to 10 GW planned	NVIDIA, AMD	2018	>$2.2 billion
DigitalOcean	Number of GPUs N/A; 16 datacenters worldwide; wattage capacity N/A	AMD, NVIDIA	DigitalOcean has sold GPU services since acquiring Paperspace in 2023	DigitalOcean is public (NYSE: DOCN)
Fluidstack	>100,000 GPUs; 1.5 GW	NVIDIA	2017	$4.5 million, with access to $10 billion debt financing
Lambda Labs	Approximately 18,000 to 25,000 GPUs and >30 MW in 14 datacenters worldwide	NVIDIA	2012	$1.65 billion
Nebius	>30,000 GPUs; about 400 MW with 1 GW planned capacity	NVIDIA	Founded as Yandex in 1989; refounded as Nebius in 2024	Public (Nasdaq: NBIS); raised about $8.6 billion between 2024 and Sept. 2025
Nscale	>40,000 GPUs in 50 datacenters; >400 MW; 700 MW planned	NVIDIA	2024	$1.7 billion
Runpod	Claims "thousands of GPUs" in 31 regions worldwide; >40 MW	NVIDIA, AMD	2022	$22 million
Shakti Cloud	>16,000 GPUs, 5 datacenters, >100 MW, 1 GW planned	NVIDIA	2024	N/A
TensorWave	About 8,192 GPUs in three data centers; 1 GW planned	AMD	2023	$143 million
TogetherAI	>36,000 GPUs; 100,000 (2 GW) planned	NVIDIA	2022	$533.5 million
Voltage Park	24,000 GPUs, >10 MW in 6 U.S. datacenters	NVIDIA	2023	$500 million
Vultr	32 leased datacenter locations across 6 continents; GPU and wattage capacity N/A	AMD, NVIDIA	2014	$662 million
WhiteFiber	>4,500 GPUs in 5 datacenters; up to 24 MW, with 1.3 GW planned	NVIDIA	2024	Public (Nasdaq: WYFI) raised undisclosed funding from majority shareholder Bit Digital before IPO in 2025

Frontier AI Models Pay the Bills, but Small Jobs Are Welcome Too

For the largest neoclouds, massive scale lends itself to more complex use cases:

AI models. Neoclouds are best known for hosting so-called frontier AI models, with OpenAI as the canonical example. This includes both training and inference, although training continues to motivate the largest neocloud buildouts.

Additional hyperscaler capacity. Neoclouds give hyperscalers a faster route to GPU capacity while they continue deploying their own AI clusters. Microsoft has signed long-term contracts with both CoreWeave and Nebius to guarantee it can satisfy its own GPU needs.

Neoclouds do cater to enterprises, too. The use cases tend to cover training, tuning, and inference of companies' specialized models. Below, we list a few examples. Note that some of them are relatively small, evidence that neoclouds aren't exclusively catering to OpenAI-sized jobs.

Drug discovery. Pharmaceutical company SieveStack used Nebius tools to assemble a 150-million-point dataset that was then used to train a drug discovery model. Data preparation pipelines are one way for the neoclouds to court enterprise customers. Also, Athos Therapeutics uses AI clusters to research novel drug targets, running model training, fine-tuning, and inference on Vultr's cloud.

Back-end Optimization. Lablup is software company that uses Vultr as a partner. It has developed Backend.AI, an AI infrastructure management platform. It is used to simplify deployment and scaling of AI workloads by resource management, automated scheduling, and MLOps pipelines.

Video generation. Higgsfield AI is an all-AI studio; it uses AI for every step from script generation to rendering and editing the final video. It runs on AMD processors in TensorWave's neocloud environment.

Research. Some neoclouds do offer small-scale GPU capacity on demand. MIT researchers building Muon, a deep learning optimization tool, use Lambda occasionally for training runs that use only eight GPUs—jobs too big for laptop GPUs or MIT's own network.

It Really Is All About Power

Before we get into the massive amount of funding going to this sector, let's talk about power.

Before ChatGPT emerged in 2022, power densities were around 30 kW per rack for high-performance computing workloads. Contrast that with NVIDIA's GB200 NVL72, a rack-scale system that consumes 120 kW. NVIDIA's Kyber rack architecture, due in late 2027 as the Rubin Ultra chips arrive, will pack 576-GPU die and is expected to consume 600 kW per rack. NVIDIA has speculated that this might push power, cooling, and switching to operate in separate standalone racks.

Securing power contracts has become the game. This is one factor driving aggressive datacenter expansion—companies are racing to lock down power. They must work with local utilities to get guarantees, in addition to the nontrivial work of securing land and permits.

Despite the enormous amounts of power in play, many neoclouds still describe themselves as champions of sustainability. This doesn't necessarily mean renewables. Crusoe, for example, powered its early deployments via surplus natural gas, a resource that refineries normally waste by burning it off. It's an available resource that the energy companies never expected to sell, so Crusoe reasoned it could become a cheap power source for cryptocurrency mining. This also gave Crusoe a climate-friendly angle, as its sites diminished natural gas flaring.

That's still a carbon-based option, however. Renewable energy sources are required in regions such as the EU and Australia. Norway's cool climate and hydropower resources have made it a particularly popular destination for neoclouds. Nscale, which focuses on renewable energy, has been enlisted to build Stargate Norway for OpenAI. CoreWeave and Crusoe also operate datacenters there.

Nuclear power is another option that cloud providers and datacenter operators are investigating. AWS, Equinix, and Microsoft have all procured power from nuclear plants. More exotic solutions are in play as well. Lambda recently announced installing all-hydrogen-powered NVIDIA GB300 NVL72 systems in Mountain View, California.

Massive Funding, Massive Debt

Some counts place the number of neoclouds at more than 100, but only a handful have been impactful. Their rapid pace has called for lots of funding and lots of debt financing.

Here's a look at how some of the most well-funded neoclouds got their capital:

Neocloud	Inception	Equity Funding	Debt Funding*
CoreWeave	2017	$2.3B	$21B
Crusoe	2018	$1.11B	$1.15B
Fluidstack	2017	$4.5M	$10B
Lambda	2012	$937M	$775M
Nscale	2024	$1.3B	$0
Together AI	2022	$533.5M	$0

Source: Company reports
* Includes announced line-of-credit financing

CoreWeave's debt financing overtook its equity financing a couple of years ago, and other neoclouds are heading in that same direction. That debt can go toward buying GPUs, but it's increasingly going toward datacenter construction. Using GPU inventories and long-term contracts as collateral, this ideally sets up a flywheel where the activity tied to sites under development can secure further debt financing for future buildouts.

The Circle of Life (Is OpenAI Too Big to Fail?)

NVIDIA has played no small role in keeping the neocloud economy humming, having invested in CoreWeave, Lambda, Nebius, and Nscale. Some of that money goes back to NVIDIA in the form of chip purchases, of course.

This circular economy is uncomfortable but is in NVIDIA's interest, partly because those clouds enable the continued growth of large language model (LLM) developers—particularly OpenAI—which in turn drives more consumption of NVIDIA products.

Even more uncomfortable is the thought that some companies have become "too big to fail." CoreWeave is on track to report net losses exceeding $1 billion in 2025—but its future also appears secure, because NVIDIA has agreed to purchase all unused CoreWeave capacity up until Aprll 2032, a deal worth $6.3 billion initially, CoreWeave says.

There are many industry dependencies on OpenAI. The company is on the hook for enormous infrastructure deals such as a $300 billion contract with Oracle, which is part of the Stargate initiative to build $500 billion in AI infrastructure through 2028. NVIDIA is also connected here, recently agreeing on paper to invest $100 billion in OpenAI, contingent on certain infrastructure milestones. (The deal was still not signed as we went to press.) All of this indicates the sector is highly dependent on both OpenAI and NVIDIA to provide capital to keep the sector humming. It has created a co-dependent economy.

What About Differentiation?

To date, neoclouds have emphasized GPU cluster performance and have focused on bringing even more GPUs online quickly. Sheer availability and speed of development were the competitive advantages.

While that's still somewhat true, the market has clearly divided into strata. For the largest-scale work, such as LLM training, there's a group that has a proven track record for operating at that scale. This includes CoreWeave, Crusoe, Lambda, Nebius, and Nscale.

Someday, though, you would think the supply-demand imbalance will abate, making neoclouds more fungible. Neoclouds might then have to turn their attention to the broader enterprise market, beyond the early adopters they currently serve—but those customers would prefer the familiarity and breadth of services offered by public clouds like AWS. In the short term, this doesn't seem to be a factor, and most of the neoclouds do not appear to be looking ahead toward this future.

Then again, certain neoclouds don't have to worry too much about differentiation. The NVIDIA-CoreWeave deal, mentioned above, means CoreWeave can keep building datacenters without fear.

3. Hyperscaler GPU Cloud Strategies

You might ask: What about the cloud giants? The largest public clouds want a say in the future of AI, too. They are the producers of some of the largest frontier models—Microsoft Copilot, Amazon Bedrock, and Google Vertex AI are leading AI services from the hyperscaler clouds, according to our own Cloud Tracker Pro research—and they also offer many AI services for both consumers and enterprises.

The major cloud providers are also building their own AI-specific datacenters, arguably playing catchup to the neoclouds. But they're also using the neoclouds as partners to provide extra GPU capacity.

Hyperscalers and Other GPU Clouds

Here's a summary of the hyperscale public clouds' plans:

AWS has launched so-called AI Zones with HUMAIN in Saudi Arabia and with SK Group in South Korea. Project Rainier in Indiana is a campus planned for up to 30 AWS datacenters with a total power draw estimated at 2.2 GW. AWS also offers "UltraClusters" based on NVIDIA chips or its own Trainium or Inferentia chips.

Microsoft is building a campus in Wisconsin, expected to be operational in 2026, but the company is also hedging its bets by leasing capacity from CoreWeave and Nebius (long-term contracts worth $10 billion and $19.4 billion, respectively). It's also partnered with Nscale on a relatively small UK facility: 50 MW and roughly 23,000 GPUs.

Google has been aggressively building its own AI datacenters, with campuses in Ohio and Oklahoma and on the Iowa/Nebraska border, plus one recently announced campus in West Memphis, Arkansas. There's no indication yet of AI datacenter builds in other countries.

Oracle provides GPUs in Oracle Cloud Infrastructure, but the company's more provocative work lies in serving up datacenter capacity to OpenAI. Oracle hosts OpenAI at the 1.2-GW Abilene, Texas, site (built by Crusoe, leased by Oracle) and recently announced plans to provide OpenAI with 4.5 GW of additional capacity at sites in Texas, New Mexico, and an unspecified Midwest location. It's apparently going well; the two recently signed a five-year, $300 billion agreement to build even more.

Alt Clouds Are In on GPUs, Too

Alternative public clouds, such as Cloudflare and Vultr, are in the mix as well. Some of them are sensitive about being lumped in with neoclouds, as they existed pre-ChatGPT, offering cloud services well beyond AI. Their original charter was to be more economical and less complex alternatives to hyperscale public clouds, and this strategy applies to their GPU clusters as well.

Like neoclouds, the alternative clouds must grapple with aggressive GPU roadmaps. Vultr points out that long-term cloud operations require continual upkeep anyway—there's always something that's due to be upgraded, and there's budget allocated for that need. It's an operational juggling act that neoclouds haven't had time to master, Vultr argues.

Being inherently multi-tenant, Vultr also monetizes older gear partly through virtualization, chopping up resources among multiple customers. It's not clear that neoclouds are ready for that step. So far, they haven't needed it; some AI customers reportedly continue running workloads on whatever generation of GPU they initially used, reserving the newest GPUs for their newest models.

Hyperscaler Cloud Differentiation

It’s early in the AI era. Inferencing and agentic AI are evolving rapidly; meanwhile, many enterprises are still in an experimental phase with AI in general. Some early adopters might be working with neoclouds, but a broader swath of enterprises will want to blend AI activity with their existing cloud workloads. This is where the hyperscalers should focus. They have these customers today and can give them a familiar setting for integrating AI into the rest of the business workflow.

Alternative clouds are moving to seize this ground too. Vultr, for example, is an established non-hyperscale public cloud that was also quick to embrace GPU clusters. Vultr uses performance as a lever for attracting business, but it also courts enterprises by promising lower prices and simpler operations than in the hyperscale clouds.

Finally, some enterprises live under thorny security and compliance rules. As Vultr points out, those teams want to see certifications from the likes of the American Institute of Certified Public Accountants (AICPA) and/or proof of adopting guidelines and frameworks from the National Institute of Standards and Technology (NIST). Vultr and its peers spent years building that expertise. Neoclouds can certainly build those skills too, but in the short term, public clouds should leverage their advantages here.

4. GPU Cloud Infrastructure: Vendor Landscape

We've talked about massive datacenters; now let's take a closer look at what goes inside them. It starts with GPUs, but AI also pushes boundaries in storage and networking, and it's made liquid cooling hot, figuratively.

GPUs: NVIDIA Rules, but Customers Want Alternatives

NVIDIA is unquestionably the center of the AI supernova. Its chips dominate the GPU field, and its aggressive roadmap is part of what's pushing neoclouds to expand so ambitiously. New chip generations are due to arrive at a rate of one per year:

Year	GPU Architecture	FP4 performance
2025	Blackwell Ultra	15 PFLOPs
2026	Rubin	50 PFLOPs
2027	Rubin Ultra	100 PFLOPs
2028	Feynmann	TBA

Source: NVIDIA

Neoclouds, especially those serving OpenAI, seem eager to ride this train. The dominant narrative is that AI training requires denser and denser GPU clusters. AI inference could be just as high-powered. NVIDIA CEO Jensen Huang says that agent-driven inference consumes orders of magnitude more tokens than his company had previously projected, due to the emergence of reasoning models and agents.

Depreciation could be just as aggressive, making it fair to ask what happens to neocloud datacenters filled with "last year's" GPUs. We'll address this further in our concluding section.

AMD

In 2025, AMD kicked into gear as the primary alternative to NVIDIA. AMD launched the Instinct MI300 family, and its current flagship processor, the MI355X, got early support from Crusoe, Oracle, TensorWave, and Vultr.

AMD can claim some advantages over NVIDIA. It packs more on-chip memory—288 GB for the MI355X versus 180 GB for NVIDIA's GB200. This helps large models run on fewer GPUs. TensorWave tells us that AMD's architecture makes it easier to virtualize the chip, letting it run multiple smaller jobs at once. That's a good way for a cloud to serve smaller AI models.

AMD must compete not only with NVIDIA's high-performance chips but also with its ability to provide the entire AI stack, including hardware, software, models, and developer tools. Another factor in NVIDIA's favor is the installed base of applications built on and for its CUDA framework. On the other hand, customers and clouds alike want to cultivate alternatives and avoid a single-vendor ecosystem. AMD is in this game for keeps and will enjoy plenty of support.

Hyperscaler Chips

Even before the ChatGPT era, hyperscale cloud providers were designing their own AI chips, as noted above. For the cloud providers, this represents a controllable, tailored alternative to NVIDIA's dominance.

The first to emerge was Google's tensor processing unit (TPU) in 2015, and chips have also emerged from Alibaba (Hanguang), AWS (Inferentia and Trainium), Meta (MTIA), and Microsoft (Maia).

Startup Alternatives

Startups such as Cerebras and Groq offer more radical architectures. Cerebras designed a wafer-scale product that's reportedly in use at Core42 and WhiteFiber. Groq was founded by CEO Jonathan Ross, formerly of Google's TPU team, who wanted to design a new chip architecture specifically for AI inferencing.

OpenAI is trying its hand at this, too. Reports emerged in September 2025 that the company was enlisting Broadcom—which co-designed Google's TPUs—to produce AI chips for internal use. This followed Broadcom's announcement that an unnamed new customer was committing to $10 billion in orders. Neither company has confirmed the deal, but news reports tie the two together, saying the chips could emerge in 2026. What we can say for sure is that this move would be in character with OpenAI planning its own datacenters and becoming its own cloud.

Networking: Applying Lessons from Hyperscalers

In networking an AI datacenter, the path of least resistance is to simply use NVIDIA's technologies. The company's NVLink and NVSwitch provide Ethernet connectivity to GPUs throughout the rack and datacenter. NVIDIA is also the lone remaining source of InfiniBand, which was expressly designed for high-performance computing.

That walled-garden situation creates an opening for other vendors, provided they can show prowess enough to supplant NVIDIA. Moreover, the rise of AMD as a GPU provider motivates

cloud vendors of all types to seek more open alternatives. Efforts such as the Ultra Ethernet Consortium are helpful in this regard, as they are building standardized ways to fortify Ethernet for AI factory requirements.

Customers need more than the usual speeds-and-feeds of networking, however. The more subtle requirement is time to revenue, or "time-to-first-job," Arista notes. Neoclouds hunger to get the latest silicon installed as soon as possible, because of the fast depreciation. Here's how networking vendors are approaching the neoclouds market:

Arista is building on its success in hyperscaler datacenters, saying neoclouds want to take advantage of those proven blueprints. The company has had major success with multiple large AI operators as well as neoclouds on both frontend and backened networks. It says it's also working with neoclouds on custom designs for scaling GPU-to-GPU networking within a rack; these are next-generation designs targeting 2027.

Cisco's AI cloud strategy centers on pre-validated AI PODs, modular building blocks such as Cisco UCS servers (with NVIDIA or AMD GPUs), Nexus switches, and Hypershield security. Cisco is a preferred technology partner for the Stargate UAE consortium and has partnerships with HUMAIN and G42. It also just announced that it's expanding its partnership with NVIDIA to ship integrated solutions that combine Nexusu, Silicon One, and NVIDIA's Spectrum-X Ethernet networking platform.

DriveNets' Network Cloud-AI provides a single network stack for both GPU networking and storage networking. WhiteFiber, previously an all-InfiniBand house, is using the technology in its newest datacenter, located in Iceland.

F5's BIG-IP services were recently validated to work with the NVIDIA Cloud Partner reference architecture. Moreover, F5's BIG-IP Next for Kubernetes can also be integrated with NVIDIA's BlueField-3 DPUs, a solution that reached general availability in April. These integrations augment a GPU cluster with F5's usual capabilities, including DDOS defense, zero-trust security, and data protection that's in line with compliance and privacy standards. When it comes to inferencing, F5 and NVIDIA have validated that their joint solution results in a 30% reduction in cost per token, the result of an increase in token generation and a dramatic (60%) decrease in time to first token.

Hedgehog likewise provides a networking software stack and initially catered to the hyperscalers. In marketing to the AI market, the company emphasizes performance and rapid, turnkey deployment. One announced customer is FarmGPU, a new neocloud based in California.

Juniper lacks an integrated rack-scale product, although it could produce one now that the HPE acquisition is done. Juniper does offer pre-validated multivendor products, highlighted by its own Mist AI for automated network management, and it provides a lab for testing AI models' performance on those solutions.

What About Security?

The rise of AI-specific infrastructure has created an entirely new set of cybersecurity risks. Shared GPUs, highly sensitive training data, and distributed supply chains are expanding the attack surface of datacenter infrastructure.

The major hyperscalers as well as alt clouds are familiar with the security needs of running cloud services for enterprises, and they have had decades of experience with secure data, networking, compute, and storage. But GPU clouds are quickly implementing new architectures and models for infrastructure, which brings new security holes and challenges. Enterprises are loath to put their most sensitive data in a GPU cloud without knowing it is properly secured. Here are some of the key security aspects that customers will be looking for as they consider the use of AI-specific GPU clouds:

Data assurance: Companies are seeking solutions to protect data and ensure its integrity, reliability, and security as AI consumes more of the data infrastructure.

Data governance: As AI adoption scales, enterprises need to securely leverage their datasets and use AI models with robust governance in place. They also need to comply with worldwide regulations for data sovereignty.

Comprehensive data security: The explosion of GPU clouds means an explosion of data. How is it handled and stored? Cloud providers must protect data in all the places it goes—in motion (across networks), at rest (in storage), and in use (memory).

Hardware security: Hardware-level security is a rising concern, with recent exploits having been discovered in GPUs from the major vendors. Neoclouds and GPU providers need strategies in place to secure GPUs, firmware, and the entire hardware and software supply chain. Some techniques include cryptographic scanning and hardware root of trust.

Confidential computing: This approach concentrates on the security of data that resides in device memories (data in use). It uses techniques to creates a trusted execution environment (TEE) for privately processing sensitive information.

Supply-chain security: The GPU clouds must secure the entirety of their supply chain, which includes any software or hardware devices being deployed, including firmware, which is being increasingly targeted by nation-state actors in large-scale events such as the Typhoon hacks, which compromised large pieces of critical infrastructure.
There are many vendors working in the areas of hardware security, supply-chain security, and confidential computing. This can protect data and code from unauthorized use. Secure enclaves, such as Intel SGX (Software Guard Extensions) or AMD SEV (Secure Encrypted Virtualization), are practical examples of confidential computing, which Futuriom has covered in detail on our website. Fortanix has been a leader in this area, offering runtime encryption solutions using Intel SGX to encrypt sensitive data. Industry leaders and companies such as Accenture, ANT Group, Arm, Google, Huawei, Intel, Meta, Microsoft, and Red Hat, together with Fortanix and other members, are collaborating to expand and advance the use of confidential computing via the Confidential Computing Consortium.

Startup Eclypsium is an innovator to watch in the hardware and supply-chain security area. Customers use Eclypsium's platform to scan hardware, firmware, and software components across their IT infrastructure and then flag vulnerabilities, threats, and inventory issues. The platform establishes trust in every endpoint, server, and network appliance in enterprise infrastructure (IT, cloud, datacenters, networks) by identifying, verifying, and fortifying third-party software, firmware, and hardware in every device.

Earlier this year, Futuriom Principal Analyst R. Scott Raynovich spoke with a large Eclypsium customer that was building AI datacenters to meet today's market demand. His take was:

> *"It's our job to provide access to the most bleeding-edge AI infrastructure while making it consumable for customers. However, it's madness thinking through how we secure infrastructure from dozens of suppliers. We want to know for a fact if their code has vulnerabilities or has been altered. If something changes in the firmware of one of our servers, we need to know instantly."*

Storage Needs to Be More than Just "Storage"

VAST Data and WEKA are best known for storage architectures based on parallel file systems, but they're seizing the AI moment to explain how they go beyond "storage" to improve overall datacenter efficiency and economics.

VAST Data declared itself the "AI operating system" earlier this year. It refers to the fact that VAST has a datacenter-spanning architecture that allows any compute nodes to access any data. This eliminates the bottlenecks that arise during the many parallel jobs that comprise an AI workload,

thus reducing the time that GPUs spend waiting for data. VAST is working with all the major neoclouds, including Core42, CoreWeave, Crusoe, Lambda, Nebius, and Nscale.

WEKA pools memory resources so that GPUs can access any memory location within the datacenter. It allows for features like persistent cache, so that GPUs don't generate redundant tokens due to running out of cache memory. WEKA champions token efficiency and touts ways to reduce the number of tokens required for inferencing.

DDN is among the leaders in high-performance parallel file systems. It has announced Core42 and Nebius as customers. It also says its DDN Infinia service is the "Platform for End-to-End AI."

OEMs Can Think Big (Rack-Scale)

The frenetic pace of AI cloud construction lends itself to turnkey equipment. That's an opening for the OEMs such as Accton, Dell, HPE, and Supermicro, which can gather servers, storage, and networking into rack-scale offerings. Supermicro notes that in addition to speed, the value lies in being a common point of contact between the customer and the various providers involved.

Liquid Cooling Finally Has Its Day

Cooling is the most obvious challenge facing GPU datacenters. Liquid cooling is getting its moment as a result, although the jury is still out on whether AI will push cloud providers to use exotic varieties of the technology. The field includes too many vendors to list, but here are the major options in play:

Single-phase, direct-to-chip cooling is the simplest option and is widely available. The liquid—usually water or glycol—circulates to cold plates attached to GPUs and CPUs, where it absorbs the heat. The liquid then continues circulating until it reaches a coolant distribution unit (CDU), which transfers the heat away.

Two-phase cooling is similar, but in this case the liquid is boiled away by the heat of the chips. The resulting hot vapor can be blown outside the rack. This remains a niche option, partly due to its cost.

Immersion cooling submerges blades, or even entire servers, in liquid. It's more disruptive to datacenter design—you need a tank, for starters—but is available from OEMs and has been proven out in large-scale settings.

The chart below outlines some of the vendors and components involved in building the world's largest AI datacenters:

Provider	Chip Suppliers	Networking	Chief data platform, infra, and storage vendors
Core42	AMD, Cerebras, NVIDIA, Qualcomm	InfiniBand, Ethernet, RoCE	VAST Data, WEKA
CoreWeave	NVIDIA	InfiniBand, Ethernet/RoCE	VAST Data, Pure Storage, IBM Storage, Backblaze
Crusoe	NVIDIA, AMD	InfiniBand, Ethernet	VAST Data, Lightbits Labs, MinIO
DigitalOcean	AMD, NVIDIA	Ethernet, InfiniBand, RoCE	Proprietary
Fluidstack	NVIDIA	InfiniBand, RoCE	DDN, VAST Data
Lambda Labs	NVIDIA	InfiniBand, Ethernet	VAST Data, DDN
Nebius	NVIDIA	InfiniBand, Ethernet	DDN, WEKA, VAST Data, proprietary
Nscale	NVIDIA	InfiniBand, Ethernet, RoCE	Dell, Supermicro, Cloudian
Runpod	NVIDIA, AMD	InfiniBand, Ethernet/RoCE	Amazon S3, Google Cloud Storage, Microsoft Azure Blob, Dropbox, Backblaze
Shakti Cloud	NVIDIA	InfiniBand, Ethernet	WEKA
TensorWave	AMD	Ethernet	WEKA
TogetherAI	NVIDIA	InfiniBand, Ethernet	VAST Data, WEKA
Voltage Park	NVIDIA	InfiniBand	VAST Data
Vultr	AMD, NVIDIA	InfiniBand, Ethernet/RoCE	Proprietary platforms, NetApp, VAST, WEKA, and DDN.

5. Conclusion: Potential Risks and Rewards of the GPU Cloud Market

The acceleration of the GPU cloud market has generated excitement across the board including in financial markets, in the utility sector, and among AI enthusiasts. AI figureheads such as OpenAI's CEO Sam Altman and NVIDIA's CEO Jensen Huang have portrayed their work as a mission to accelerate the innovation in AI.

However, excitement about innovation also requires caution. The shift from equity to debt financing and massive datacenter projects among the emerging neoclouds have raised the risk profile. The interconnected nature of many deals, all leading back to a few companies, has ratcheted up some of the questions—and nerves—in the financial markets.

If AI revenues don't scale fast enough to support these investments, some of them may be scaled back or even dropped. In addition, the neoclouds and GPU clouds involve only one aspect of the AI innovation buildout. There will also be private AI datacenter buildouts, edge-enabled AI, and other forms of AI we don't even know about yet.

The Depreciation Question

We can all see what's motivating new datacenter buildouts, but what happens to today's datacenters as their GPUs age and depreciate? What if the construction schedules are not matched by an equal pace of adoption?

Unlike railroads or even fiber optics—capital intensive booms of the past that spurred massive investment—GPUs don't retain their value for decades. In fact, they depreciate fast—sometimes as fast as 4-5 years.

All clouds offering GPUs will need to deal with that soon. Older chips might remain in demand, since not all workloads require cutting-edge technology, but they can't command the prices they once did. That means they've depreciated faster than expected. Hyperscalers and CoreWeave use a six-year useful life when estimating depreciation on the chips. If NVIDIA's pace accelerates depreciation, there will need to be a reckoning in that accounting. A lot depends on how quickly supply catches up to demand.

Have the Hyperscalers Outsourced Risk?

The rush to build out AI clouds has also created some uncomfortable economics and interdependencies. Specifically, NVIDIA is frequently called out for essentially financing its own customers. The $100 billion handshake deal with OpenAI is the most overt example, but NVIDIA also invested in CoreWeave, and some of that money was applied to facilities housing OpenAI and running NVIDIA chips.

NVIDIA also runs a GPU cloud based partly on renting its own GPUs back from other clouds. This includes a $1.5-billion deal to lease 18,000 GPUs from Lambda over four years, and the $6.3-billion deal to absorb CoreWeave excess datacenter capacity. Both deals were announced in September 2025.

It's reminiscent of the vendor financing in the telecom sector circa 1999. Equipment vendors including Alcatel, Cisco, Lucent, and Nortel, funded their customers in amounts reaching into the billions. Most of those customers failed and/or were acquired as the telecom bubble collapsed. Note that the four vendors mentioned were all giants, but only one still exists as a standalone entity.

Regardless, there’s no legal or regulatory issue that applies to what NVIDIA is doing. The appetite for building outsized datacenter campuses clearly exists, and there's ample capital willing to feed it. AI infrastructure does seem headed for bubble status but there’s no way to predict how long it will last.

The hyperscalers appear to have recognized some of the risks of the interconnected dependencies leading back to NVIDIA, OpenAI, and some of the major neoclouds such as CoreWeave. By investing and outsourcing some of their GPU clouds to other players, the hyperscalers are reducing the risks to their own balance sheet. They are also implementing other hedging and diversification strategies. This includes developing their own AI chips and spreading the responsibility for building out new GPU cloud infrastructure to multiple partners.

Future Evolution of Neoclouds, Hyperscalers, and Alt Clouds

As the AI cloud industry grapples with these risks, it’s likely that there will be a fast evolution of business models and services. Neoclouds will likely evolve by becoming more specialized, potentially leading to consolidation, and they will serve a crucial, though perhaps smaller, role compared to hyperscalers.

The evolution of the industry will be highly dependent on the AI-specific needs of enterprises and consumers. The new generation of AI clouds will also have to adapt to engineering considerations such as energy, sustainability, cooling, water supplies, and proximity to the edge. The hypercale and alt clouds have an advantage of offering a wider variety of services and the capability to leverage their own infrastructure as AI economics improve. Some hyperscalers are developing their own silicon and cutting prices. We see an industry that should rapidly trend toward consolidation into a few major players, diversification into AI software, and a focus on additional service layers connected to commodity clouds.

Appendix: Selected AI and GPU Cloud Ecosystem Players to Watch

AMD (Nasdaq: AMD)

For more than 55 years, AMD has driven innovation in high-performance computing, graphics, and visualization technologies. AMD Instinct MI350 Series GPUs are built for generative AI and high performance computing (HPC) in data centers. Based on the fourth-generation AMD CDNA architecture, these GPUs deliver efficiency and performance for training massive AI models, high-speed inference, and complex HPC workloads such as scientific simulations, data processing, and computational modeling. Instinct accelerators are suitable for single-server solutions up to the world's largest, exascale-class supercomputers.

Arista Networks (NYSE: ANET)

Arista a leader in networking for AI with its Etherlink portfolio of switches powered by EOS and in AI for networking with its AVA (Autonomous Virtual Assist) for AIOPs, security and observability. A pioneer in building the world's largest data centers, Arista also delivers mission-critical AI centers of data that scale-up, scale-out and scale-across, supporting some of the largest clusters of AI accelerators and demanding workloads in training and inference. Arista's unifying approach, complete with open standards at every layer, optimizes performance and operations alike.

https://www.arista.com/

Cisco (Nasdaq: CSCO)

Cisco delivers the critical infrastructure to help organizations thrive in the AI era, offering products in networking, security, observability, and collaboration. It works as a trusted partner to hyperscale builders, neocloud providers, service providers, and enterprises. The company announced that it received more than $1 billion in AI infrastructure orders in fiscal 2025, which ended in July. Cisco's offerings include the NVIDIA Enterprise AI Factory validated design and AI PODs, modular elements that combine into a flexible, scalable architecture.

CoreWeave (Nasdaq: CRWV)

CoreWeave calls itself the "AI Hyperscaler," delivering elastic, NVIDIA GPU-dense clouds engineered for training and inference at scale. It rents ultra-specialized, pay-per-second clusters that spin up in minutes. CoreWeave launched in 2017 as Atlantic Crypto, an Ethereum mining operation. It rebranded as CoreWeave in 2019 and went public in 2025, raising $1.5 billion at launch, and it now enjoys a market cap of roughly $67 billion.

Crusoe

Crusoe was founded in 2018 as a cryptocurrency miner, using natural gas flare-offs as its source of electricity. The company's Crusoe Cloud is a scalable platform optimized for AI workloads. Crusoe is better known, however, for building the 1.2-GW Abilene, Texas, datacenter campus that will be operated by Oracle, which will lease the cloud capacity to OpenAI. In 2025, Crusoe announced it had another 20 GW worth of construction contracts in the pipeline.

DriveNets

DriveNets is a leader in high-scale disaggregated networking solutions. Founded in 2015, DriveNets modernizes the way service providers, cloud providers and hyperscalers build networks, streamlining network operations, increasing network performance at scale, and improving technological and economic models. DriveNets' solutions – Network Cloud and Network Cloud-AI – adapt the architectural model of hyperscale cloud to telco-grade networking and support any network use case – from core-to-edge to AI backend networks – over a shared physical infrastructure of standard white-boxes, radically simplifying the network's operations and offering telco-scale performance and elasticity at a much lower cost.

https://drivenets.com/

Eclypsium

Eclypsium's cloud-based and on-premises hardware supply chain security platform provides protection for critical software, firmware and hardware in enterprise infrastructure. Eclypsium helps enterprises and government agencies mitigate risks to their infrastructure from complex technology supply chains. Eclypsium provides solutions for AI data centers including hardware supply chain security for GPU clouds as well as network infrastructure devices.

https://eclypsium.com/

F5 (Nasdaq: FFIV)

F5 is a multi-cloud application services and security company committed to bringing a better digital world to life. F5 partners with the world's largest, most advanced organizations to secure and optimize apps and APIs anywhere—on premises, in the cloud, or at the edge. F5 enables organizations to provide exceptional, secure digital experiences for their customers and continuously stay ahead of threats.

https://www.f5.com/

Fluidstack

Fluidstack is a UK-based neocloud that claims to have 100,000 GPUs under management. This year, it signed a memorandum of understanding with the French government to build a 1-GW supercomputer for sovereign AI, comprising 500,000 GPUs. France has committed €10 billion ($10.4 billion) for Phase 1, due to come online in 2026. Fluidstack raised only $4.5 million in equity funding but, according to The Information, has approval to borrow up to $10 billion from Macquarie and others.

Lambda

Lambda, which calls itself the “superintelligence cloud,” was founded in 2012 by researchers looking for ways to reduce AI costs. Its cloud occupies datacenters in 11 US cities. Lambda has not yet occupied gigawatt-scale campuses, but it does gain access to the latest NVIDIA chips, since NVIDIA is an investor. Lambda raised a $380 million Series C in 2024 followed by a $480 million Series D in 2025.

Nebius (Nasdaq: NBIS)

Nebius was incorporated in The Netherlands in 2024, but its roots go back to 1989 as the infrastructure division of Russian search engine Yandex. Nebius is now independent and publicly traded. Nebius’s AI-native cloud platform, including proprietary software and hardware designed in-house, has been built for intensive AI workloads. Nebius provides AI builders with compute, storage, managed services, and tools for building, tuning, and running their models and applications. Its offerings include an end-to-end data preparation solution called TractoAI.

Nscale

UK-based Nscale was spun out of datacenter operator Arkon Energy in 2024. It offers a full-stack AI cloud platform with a focus on sustainability, including energy-efficient datacenters in Norway powered by renewable energy. By 2027, Nscale aims to operate hundreds of thousands of GPUs across multi-gigawatt green energy sites, striving to be a leading AI hyperscaler with major partnerships including NVIDIA, Microsoft, OpenAI, Lenovo, AMD, and Nokia.

NVIDIA (Nasdaq: NVDA)

NVIDIA is the world leader in AI and accelerated computing. It describes itself as the engine of AI, engineering the most advanced chips, systems, and software for the AI factories of the future. In addition to GPUs, which fuel industrial digitization across multiple markets, NVIDIA provides software libraries and development kits, AI frameworks, cloud-hosted AI platform services, edge computing platforms, and digital twin and simulation software.

Oracle (NYSE: ORCL)

Founded in 1977, Oracle's core mission is to help customers discover new insights in data. Its products include the flagship Oracle Database and integrated suites of applications. The Oracle Cloud provides autonomous infrastructure including computing, storage and networking as a service. In the neocloud context, Oracle has become a major infrastructure broker, leasing massive datacenter campuses and selling the capacity to frontier AI labs, particularly OpenAI.

TensorWave

TensorWave, founded in 2023, is a neocloud providing only AMD GPUs. The company raised $43 million in a Simple Agreement for Future Equity (SAFE) transaction in 2024, followed by a $100 million Series A in 2025 that AMD participated in. The company's deployments include a dedicated training cluster of 8,192 AMD MI325X GPUs in Tucson, Arizona, and it was an early MI355X buyer as well.

Vultr

Vultr delivers enterprise-grade AI and GPU cloud infrastructure featuring globally decentralized compute clusters, robust networking, top-tier security, and regulatory compliance. Developers can instantly access dedicated or on-demand AMD and NVIDIA GPUs—including latest-generation models such as the NVIDIA HGX B200 and AMD Instinct MI355X—for training, deployment, and edge-serving AI applications. Vultr removes vendor lock-in, offers transparent pricing, and integrates easily with tools and APIs. Advanced orchestration (via Run:ai and Kubernetes) ensures optimal resource use, lifecycle management, and rapid AI project development across regions. This approach enables secure, compliant, cost-effective scaling and redefines AI architecture as the new public cloud standard.

https://www.vultr.com/