PitchBook

Ali Javaheri
Analyst, Emerging Technology
ali.javaheri@pitchbook.com

Brendan Burke
Senior Analyst, Emerging Technology
brendan.burke@pitchbook.com

Ben Riccio
Associate Analyst, Emerging Technology
ben.riccio@pitchbook.com

EMERGING SPACE BRIEF

AI Neoclouds

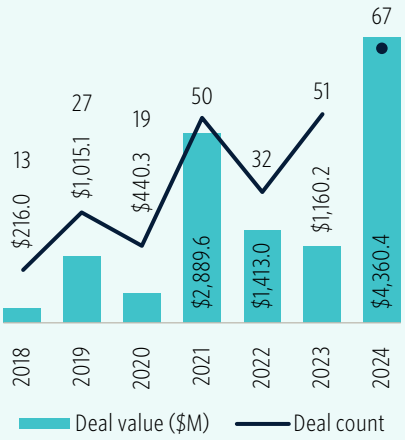
Originally published November 27, 2024

pbinstitutionalresearch@pitchbook.com

Trending companies

CoreWeave · Lambda · Crusoe · SambaNova Systems · groq · cerebras

AI neoclouds VC deal activity



Source: PitchBook • Geography: Global
As of October 31, 2024

For access to more of this data and PitchBook's Emerging Spaces tool, access a free trial link here.

Overview

"Neoclouds" is a term coined by infrastructure market research firm SemiAnalysis, referring to cloud providers that offer graphics processing unit (GPU) compute rental.[1] This business model can involve both GPU ownership and brokerage. Besides coining this term, SemiAnalysis has taken a lead on more conventional market researchers for the firm's depth of investigation into novel semiconductor architectures and data collection on infrastructure performance. We believe the term should also include non-GPU cloud services and legacy cloud providers that develop net new clusters for AI processing. Based on our review of the market, we believe the advanced AI cloud market is congealing into a standalone segment that can complement the market leadership of hyperscalers by focusing on the needs of high-growth large language model (LLM)-native customers.

Background

The growth of AI neoclouds originated with the early demand for high-performance infrastructure initially driven by cryptocurrency mining. Recognizing that their GPUs suited the intensive needs of AI model training and deployment, crypto miners began repurposing and expanding their clusters to meet the demands of modern AI models, laying the groundwork for a new market of cloud services focused on high-performance compute. In practice, these rigs were configured for crypto mining, requiring different software optimizations and power requirements to carry out the massively parallel operations of neural network training. Of 187 companies in our AI neocloud emerging space, 15 are also tagged in the cryptocurrency vertical, showing the continued overlap between the spaces.

During the AI boom of 2023, GPUs became scarce, pushing customers past hyperscalers to alternative cloud providers. The leading-edge NVIDIA H100 experienced price increases as much as three times its initial launch cost. Simultaneously, VCs granted unicorn valuations to companies with high AI model download counts, driving demand for additional GPU access. These trends encouraged aggregators of GPUs to repurpose their clusters for AI developers. However, initial GPU brokers faced constraints from older hardware, lack of

1: "AI Neocloud Playbook and Anatomy," SemiAnalysis, Dylan Patel and Daniel Nishball, October 3, 2024.

advanced networking, and limited availability. The initial signals of demand encouraged the forefather of this trend, NVIDIA, to support the company's channel partners.

NVIDIA has significantly supported distributed AI computing by partnering with both hyperscalers and specialized neocloud providers to expand access to its high-performance hardware, including GPUs as well as other essential components such as network accelerators and fast storage solutions. NVIDIA has announced 13 cloud partners and made VC investments in nine companies via its multiple CVC arms, including some overlap for companies such as Applied Digital and CoreWeave. Further, 39 data center operators have become colocation partners with NVIDIA, leveraging the company's DGX platform for AI workloads. These collaborations enable a range of cloud hosts to offer not only GPU compute but also a tailored combination of software and networking resources optimized for AI and machine learning. This vendor support allows neoclouds to provide flexible, high-throughput infrastructure, supporting industries that require intensive computational power and scalable resources.

On the PitchBook platform, we have made this space trackable in multiple ways. First, we tag companies via a bottom-up process to identify those cloud providers making NVIDIA GPUs and other advanced AI-specific hardware available. Second, our emerging space automatically labels all new companies featuring description keywords, including GPU cloud, GPU compute, GPU computing, and GPU rental. This automation integrates new companies fitting these descriptions in real time. These combined approaches include updated lists of market participants as legacy cloud operators pivot into the space.

Technologies and processes

GPU compute rental occurs through diverse business models:

AI hyperscalers: Large-scale cloud providers that offer extensive AI infrastructure and services, capable of handling massive data processing and computational needs.

AI neoclouds: Next-generation cloud platforms specifically designed to cater to AI workloads with optimized architectures and specialized capabilities beyond traditional cloud services.

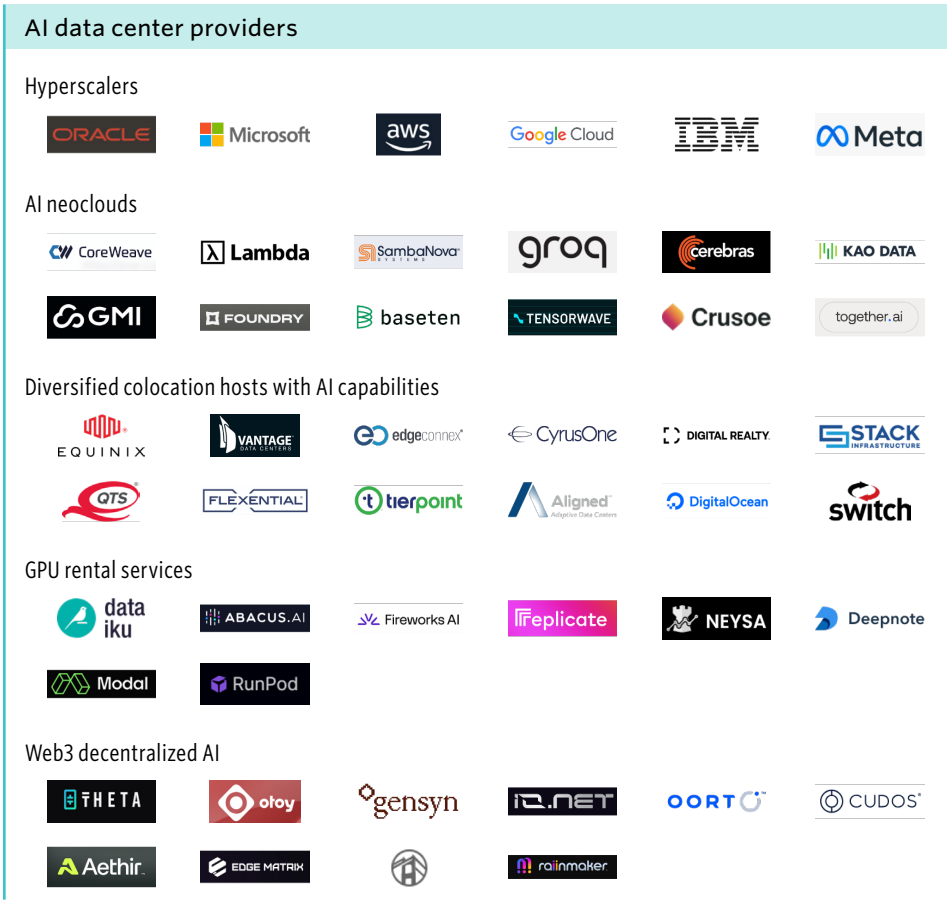
Diversified colocation hosts with AI capabilities: Data center providers that offer colocation services with additional AI infrastructure support, allowing businesses to deploy a range of workloads along with enhanced AI processing capabilities.

GPU rental services: Aggregators that rent out high-performance GPUs to customers via third parties, enabling users to select among diverse hardware instances and lowering upfront costs.

Web3 decentralized AI: Cryptocurrency-supported decentralized infrastructure,

governance models, and data sources to design, train, and run AI models. This technical innovation was recently covered in our crypto analyst note Token Incentives for Decentralized AI Operations.

Q4 2024 AI data center market map



Source: PitchBook
Note: This is a non-exhaustive list of leading vendors. For the full list of constituents, see our platform market map here.

AI neoclouds can stand out on both pricing and performance benchmarks. From the start, neoclouds aggregated GPUs and offered them at competitive prices. VC subsidies and aggressive vendor financing from NVIDIA enabled low rates, but this trend is stabilizing as GPU costs remain high. If these GPU clusters lack reliability, then total costs can remain high, as customers must waste time and compute cycles on failed training runs. As a result, customers push neoclouds to improve their underlying software optimization, lowering the total cost of usage and improving customer satisfaction.

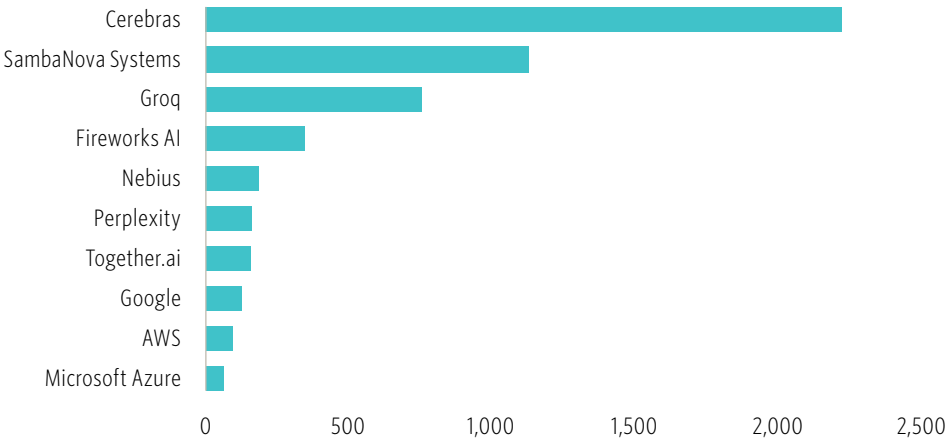
Hardware optimization has been a megatrend driving LLM cost reductions, as engineering in the background of NVIDIA's CUDA software library improves utilization and throughput of AI model operations. NVIDIA has demonstrated the

value of hardware optimization software via M&A. The company acquired Deci, OctoAI, and Run:AI for more than $1.5 billion total to bring cutting-edge open-source frameworks to deploy models via containers, which flow downstream to their hyperscaler customers, who rely on NVIDIA software for GPU scheduling.

Startups stand out for bringing their own scheduling approaches to market. Common GPU scheduling frameworks include Slurm (simple Linux utility for resource management) for GPU workload queuing, which interfaces directly with bare metal, and Kubernetes, which deploys AI models via container instances and can scale out across cloud environments. In 2023, market leader CoreWeave launched a proprietary scheduler called Sunk, which refers to Slurm on Kubernetes, and enables users to deploy the more batch-based Slurm jobs via more scalable Kubernetes containers. CoreWeave previously stood out for supporting the more generalist Kubernetes framework but improved its deployment methods to adapt to leading research approaches. Baseten commercializes the open-source Truss framework to containerize workloads, showing the novel innovations that startups can bring.

Custom hardware can avoid the need for increasingly complex GPU scheduling frameworks. The fastest APIs for AI output come from the unique hardware stacks of Cerebras, Groq, and SambaNova Systems. While these companies sell their proprietary chips, they also offer API services that resemble clouds, leveraging unique configurations of their custom hardware. These approaches have caused neoclouds to stand out in tokens per second for leading open-source models such as Meta's Llama series, according to independent testing by Artificial Analysis.[2] They remain competitive on price, which we believe serves as a loss leader to sell higher-value chip licensing deals, yet encourage large customers to adopt architectures that may prove to be best-in-class for inference in the long run if they can scale up their services with additional funding and chip shipments. For production workloads, customers will still go with more reliable cloud services that can support greater scale and leading closed-source models.

Speed by API provider (Llama 3.1 8B inference median output tokens per second)



Source: Artificial Analysis LLM API Providers Leaderboard • Geography: Global • As of November 15, 2024

2: "LLM API Providers Leaderboard—Comparison of Over 100 LLM Endpoints," Artificial Analysis, n.d., accessed November 19, 2024.

The race to offer NVIDIA H200 GPUs, made available earlier this year, has widened the gap between the most well-funded startups and newcomers to the space. Currently, CoreWeave, Crusoe, Together AI, and GMI offer H200 GPUs for reservation, while Lambda offers them only through private cloud deployment. Groq and Cerebras develop and deploy their own application-specific processors to provide top inference speeds for pretrained models. As a result, companies can specialize in either training or inference based on the quantity of chips available. There may be a need for rapidly deployed small clusters for fine tuning small models and then larger clusters of 10,000 or more H200s for foundation model training. The fine-tuning use case relies on custom software such as Axlotl and can use smaller GPUs such as RTX instead of NVIDIA's latest. For this reason, not all VC in neoclouds will be spent on the latest generation of NVIDIA chips, affecting NVIDIA's revenue mix.

Available semiconductors for leading neoclouds

AI neocloud	Available semiconductors
CoreWeave	HGX H100 & H200, Blackwell: B200 & GB200, A100, A40, A5000, and RTX 6000
Lambda	H200 through private cloud only, H100 SXM/Pcle, A100 SXM/Pcle, A10, A6000, Tesla V100, RTX 6000
Crusoe	H200 & H100, MI300x, A100, A40
Groq	Groq LPU
Cerebras	Cerebras WSE-3
Together	H200, H100, A100, A40, A5000
GMI	H200, H100
Fireworks AI	H100, A100
Foundry	H100 , A100, A40, A5000

Source: PitchBook analysis of company disclosures

Applications

AI neoclouds optimize for LLM innovation yet can benefit from diverse customer bases to insulate themselves from potential AI winters. Key customer types include:

LLM training and inference: LLMs require significant computational resources to process their vast numbers of parameters during training and to deliver real-time responses during inference. Neoclouds provide the flexibility to handle these workloads, offering high-performance compute clusters capable of managing the significant memory, storage, and bandwidth demands of LLMs. Customers training custom foundation models bring maximum utilization of GPU clusters, making them anchor customers of neoclouds, although high-throughput applications may increase inference needs in the future.

Key AI partnerships for leading neoclouds

AI neocloud	AI partnerships
CoreWeave	EleutherAI Microsoft Mistral AI NovelAI
Lambda	Nous Research
Crusoe	Cartesia
Groq	Aramco Digital Meta
Cerebras	G42 Jasper AI
Together	Meta Salesforce Snowflake

Source: Company disclosures

Healthcare: Medicine remains a specialized domain for AI training with strict privacy requirements. Neoclouds can configure specialized cloud instances that maintain data governance. Cerebras became the first generative AI (GenAI) partner of leading medical center Mayo Clinic for custom model training.

Visual effects: Visual effects professionals require GPU access and benefit from virtual workstations that interface directly with advanced hardware in real time. Neoclouds can offer specialized workflow solutions for these professionals by integrating GPU access—a cornerstone of CoreWeave's business along with cryptocurrency mining.

Scientific research and high-performance computing: Fields such as genomics, climate science, and molecular modeling rely on AI neoclouds to process extensive datasets and run simulations. By providing access to powerful and scalable GPU resources, AI neoclouds enable researchers to accelerate data analysis and simulations, driving progress in scientific discovery without large capital investments in hardware.

Limitations

Carbon emissions: Neoclouds face stringent emissions regulations with which data centers must comply, including limits on greenhouse gas outputs and reporting standards for sustainability. AI workloads generate significant heat, necessitating advanced cooling systems. Traditional cooling methods often consume additional energy, contributing to a center's carbon footprint. As a result, data centers often participate in carbon offset programs to balance out emissions they cannot eliminate by investing in reforestation, renewable energy projects, or carbon-capture technologies.

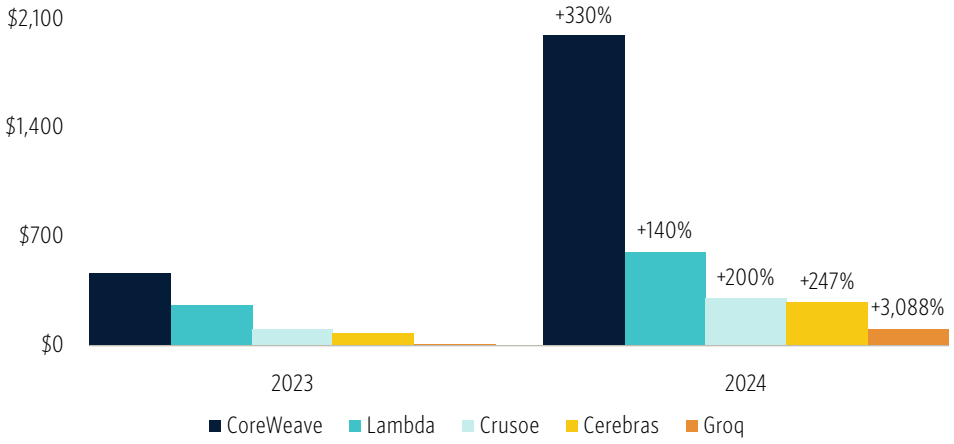
Power availability: Power remains a significant constraint to hyperscalers, with data center power demand likely to double in the coming decade. Meta recently failed to receive a permit for a nuclear plant to power an AI data center because of environmental permitting reasons including the presence of rare bees. xAI installed gas turbines onsite to match its peak power loads, although the company skipped environmental permitting that may delay its projects. CoreWeave demonstrates that power permitting can be a skill, given the founding team's background in commodity trading. The company has secured 850 megawatts of power supply, nearly as much as Microsoft is planning for its cutting edge 1 gigawatt data center. These strategies may separate companies over the next three years.

Hardware depreciation: Cutting-edge semiconductors may depreciate more rapidly than scheduled due to release of new chip architectures and rapid progress in underlying base models. NVIDIA offers bundled software and hardware packages that promise to maintain the relevance of legacy hardware. Even so, access to new phases of hardware will incur high upfront capital costs to AI neoclouds and lessen the market value of their existing hardware, challenging debt repayment.

Recent deal activity

As with NVIDIA, outlier revenue growth for AI neoclouds fuels rapid valuation growth. Coreweave and Lambda Labs have leveraged their best-in-class hardware and comprehensive software suite to take advantage of the surging demand for AI computing. Coreweave reportedly will earn $2 billion in revenue in 2024, after achieving just $30 million two years prior. Lambda Labs, which offers public and private cloud deployment, forecast $600 million in revenue for 2024. Cerebras and Groq, some of the only providers to not use NVIDIA GPUs, are on pace to achieve $272.8 million and $108.4 million in 2024 revenue, respectively. Together, a startup utilizing third party GPUs, has reached $100 million in revenue in 10 months of operation. These companies have raised significant rounds in rapid succession, uncovering an underserved market with expectations for future high growth.

Most recent revenue estimates for leading AI neocloud companies



Source: PitchBook and company disclosures • Geography: Global • As of November 1, 2024

VC deals fuel the GPU land grab and increasingly incorporate debt to expand physical footprints. We have tracked seven megadeals in 2024 after only three in 2023, with rapid valuation growth the norm. Strategic investors make outsize contributions to these rounds, with non-NVIDIA leaders including Cisco, Samsung, and Salesforce. We have tracked 31 AI neocloud deals incorporating debt in 2024, up from 19 in 2023, including CoreWeave's $7.5 billion megatranche. Companies differ in capital needs based on their GPU purchasing plans, with some companies outsourcing usage to third-party colocation hosts such as Together and needing to raise less as a result. Chip companies including Groq can use high deal values to build out custom data centers and take capital expenditure risk on behalf of their customers. These VC deals tend to be completed around 10x forward revenue, showing the difference in business models between these companies and software-focused AI startups.

Key recent pure-play neocloud VC deals

Company	Close date (2024)	Deal size ($M)	Post-money valuation ($M)	Lead investor(s)
Crusoe	October 29	$500.0	$3,000.0	Founders Fund
Groq	August 5	$640.0	$2,800.0	BlackRock, Cisco Investments, Samsung Catalyst Fund, Type One Ventures
Lambda	August 1	$800.0	N/A	Bossa Invest
CoreWeave	May 1	$1,100.0	$19,000.0	Coatue Management, Magnetar Capital
Celestial AI	March 26	$175.0	$1,175.0	US Innovative Technology Fund
Together	March 13	$106.0	$1,250.0	Salesforce Ventures
Groq	February 28	$46.0	$1,146.0	Cisco Investments, Samsung Catalyst Fund, Type One Ventures
Lambda	February 16	$320.0	$1,520.0	US Innovative Technology Fund

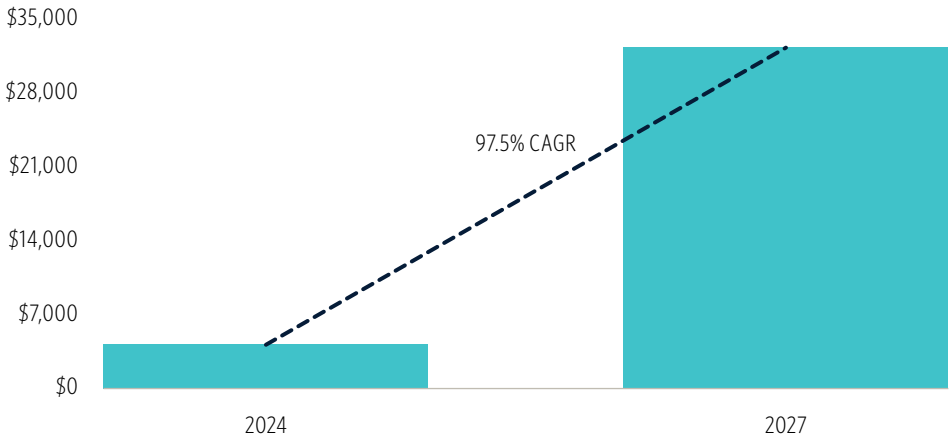
Source: PitchBook • Geography: Global • As of November 1, 2024

Secondary markets drive valuations higher and prepare companies for IPO. Astera Labs showed that NVIDIA-related data center tech companies can achieve successful IPOs in the current market environment. CoreWeave plans to list for an IPO, driving pre-IPO investment. The company faces high secondary market demand ahead of this IPO, most recently completing a secondary transaction with blue chip public investors. Cerebras has already filed for an IPO, and its case will be bolstered by revenue diversification through cloud services.

Market outlook

We believe the neocloud market will reach at least $4 billion in end-user spending in 2024 and can grow quickly even with market leadership from Amazon and Microsoft. This top-down estimate based on an overall GenAI cloud market of $9.2 billion may prove conservative, as CoreWeave claims to be on track for $2 billion in revenue in 2024 and Amazon has also generated a multi-billion-dollar business. This market is on pace to grow quickly in coming years, and we estimate it will reach $32 billion by 2027. At that time, the incremental investments in cutting-edge hardware may pay off, although sustained capital expenditures will be required to maintain market leadership.

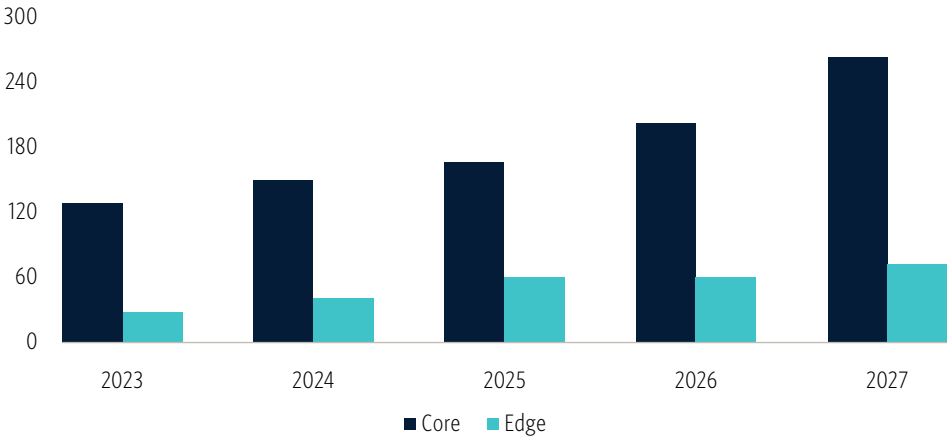
GenAI cloud services spending estimate ($M), excluding Microsoft



Source: PitchBook analysis of IDC data • Geography: Global • As of September 30, 2024

This high spending growth supports plans for datacenter construction to restore high growth after a global decline in 2022. Centralized cloud-service-provider datacenters produce the majority of new build spending, yet edge data centers will grow more quickly over the next three years, supporting the cloud-edge strategies of hyperscalers as well as content delivery specialists, including telecom operators and inference data centers. This forecast takes into account power scarcity and supply chain constraints, assuming some adaptation in terms of energy efficiency.

Datacenter new build and retrofit spending ($B) by type



Source: IDC • Geography: Global • As of October 31, 2024

Quantitative perspective

For a deeper dive into the data and to explore additional insights, visit the PitchBook Platform or request a free trial.

187 companies	**779** deals	**1,307** investors	**$423.1B** capital invested
128 deals (TTM) *64.1% YoY*	**$51.0M** median deal size (TTM) *134.5% YoY*	**$265.0M** median post-money valuation (TTM) *163.9% YoY*	**$76.0B** capital invested (TTM) *99.8% YoY*

As of October 31, 2024

Top VC-backed AI neocloud companies by total raised

Company	Total raised ($M)	Last financing value ($M)	Last financing date (2024)	Last financing deal type	HQ location	Year founded
CoreWeave	$9,089.8	$1,300.0	November 13	Secondary transaction - private	Roseland, US	2017
Lambda	$1,692.5	$800.0	September 30	Late-stage VC	San Jose, US	2012
Crusoe	$1,408.1	$500.0	October 29	Late-stage VC	Denver, US	2018
SambaNova Systems	$1,138.0	N/A	August 1	Late-stage VC	Palo Alto, US	2017
Groq	$1,048.6	$640.0	August 5	Late-stage VC	Mountain View, US	2016
Dataiku	$851.9	N/A	N/A	Accelerator/incubator	New York, US	2013
Moore Threads	$800.2	$274.6	November 15	Late-stage VC	Beijing, China	2020
Cerebras	$723.0	N/A	April 2	IPO	Sunnyvale, US	2016
Kao Data	$473.4	$261.9	January 24	Debt refinancing	Harlow, UK	2014
WEKA	$415.1	$140.0	May 15	Late-stage VC	Campbell, US	2013

Source: PitchBook • Geography: Global • As of October 31, 2024

Top VC-backed AI neocloud companies by Exit Predictor Opportunity Score

Companies	Opportunity score	Success probability	M&A probability	IPO probability	Total raised ($M)	HQ location	Year founded
HIGHRESO	99	95%	17%	78%	$30.1	Tokyo, Japan	2007
Neysa Networks	99	91%	62%	29%	$49.5	Mumbai, India	2023
Aethir	98	91%	82%	9%	$23.0	Singapore, Singapore	2021
io.net	98	96%	93%	3%	$35.0	New York, US	2019
Salad	97	92%	91%	1%	$29.6	Salt Lake City, US	2018
Kindo	96	86%	78%	8%	$27.6	Los Angeles, US	2022
CUDOS	95	89%	88%	1%	$23.4	London, UK	2017
Foundry	95	84%	74%	10%	$79.1	Palo Alto, US	2022
GMI	95	85%	78%	7%	$142.0	San Jose, US	2023
TensorWave	93	84%	81%	3%	$45.3	Las Vegas, US	2023

Source: PitchBook • Geography: Global • As of October 31, 2024
Note: Probability data based on PitchBook VC Exit Predictor Methodology.

Top VC-backed AI neocloud companies by active patents

Company	Active patent documents	Total raised ($M)	HQ location	Year founded
SambaNova Systems	189	$1,138.0	Palo Alto, US	2017
Cerebras	132	$723.0	Sunnyvale, US	2016
Numenta	116	$45.4	Redwood CIty, US	2005
Groq	81	$1,048.6	Mountain View, US	2016
LiquidCool Solutions	70	$27.7	Rochester, US	2006
Cerio	55	$91.5	Ottawa, Canada	2012
Crusoe	17	$1,408.1	Denver, US	2018
GRAID Technology	6	$20.3	Santa Clara, US	2020
SiliconArts	5	$10.1	Seoul, South Korea	2010
Juice Labs	2	$1.8	San Francisco, US	2020

Source: PitchBook • Geography: Global • As of October 31, 2024

Top AI neocloud investors

Investor	Investments	Primary investor type	HQ location	Year founded
Y Combinator	12	Accelerator/incubator	Mountain View, US	2017
NVIDIA	9	Corporation	Santa Clara, US	2016
Celesta Capital	8	VC	San Mateo, US	2005
Foundation Capital	7	VC	Palo Alto, US	2016
Intel Capital	7	CVC	Santa Clara, US	2006
Sequoia Capital	7	VC	Menlo Park, US	2012
Andreessen Horowitz	6	VC	Menlo Park, US	2018
Dawn Capital	6	VC	London, UK	2020
FirstMark Capital	6	VC	New York, US	2010
SV Angel	6	VC	San Francisco, US	2020

Source: PitchBook • Geography: Global • As of October 31, 2024

Recommended reading

- "AI Neocloud Playbook and Anatomy," SemiAnalysis, Dylan Patel and Daniel Nishball, October 3, 2024.
- "Assessing Potential US Nuclear Plant-Data Center Relationships," Morningstar, Travis Miller, September 26, 2024.
- "Industry Update: Data Center Implications Across Industrials and Software," Davidson Research, D.A. Davidson, et al., July 1, 2024.
- "Private Credit & Middle Market Weekly Wrap," PitchBook/LCD, October 17, 2024. (CoreWeave research on page 8)
- "Semiconductors: 2024 Q3," Morningstar Equity Research, Brian Collelo and Jivyaa Vaidya, October 7, 2024.