

The Neocloud Economy

A sector study of the neocloud business model, competitive dynamics, financing structure and the pricing of compute

Central question. The three largest hyperscalers still control 63 per cent of cloud infrastructure. Why, then, is a group of companies barely four years into their present form taking on tens of billions of dollars of debt in 2026 to assemble GPU fleets?

Finding: the neocloud is best understood as a *capacity-financing structure* — a way for hyperscalers to expand GPU supply without carrying it on their own balance sheets. That structure works, but it rests on an economy squeezed between tenant concentration, fast hardware depreciation and a debt schedule that does not wait.

REPORT SCOPE Twenty-three analytical sections · market sizing and competitive map · four company studies · financing and credit-structure analysis · compute pricing indicators · power, financing and competition risk

EVIDENCE DISCIPLINE Every material quantitative claim is attributed to a named primary or third-party source, cited at the foot of each exhibit and listed in full in the references section. Announced targets are never presented as operating results.

NOT INVESTMENT ADVICE This report identifies capital-allocation categories and diligence tests. It contains no security recommendations, ratings, price targets or return expectations.

EXECUTIVE SUMMARY

1. What a neocloud is, and why it matters now

A "neocloud" is a non-hyperscaler operator that sells AI compute capacity by installing GPU or TPU clusters in leased colocation facilities rather than by building large-scale data centres of its own. CoreWeave, Nebius, Lambda and Crusoe all sit in this category, and their business model is the same in outline: debt-financed hardware, rented to hyperscalers and AI laboratories under multi-year capacity agreements.^[1]

\$400bn

Forecast neocloud market size by 2031, at a 58 per cent compound growth rate^[3]

\$1.75tn

Share of AI capex expected to be financed through credit markets over three years^[1]

9 of 40

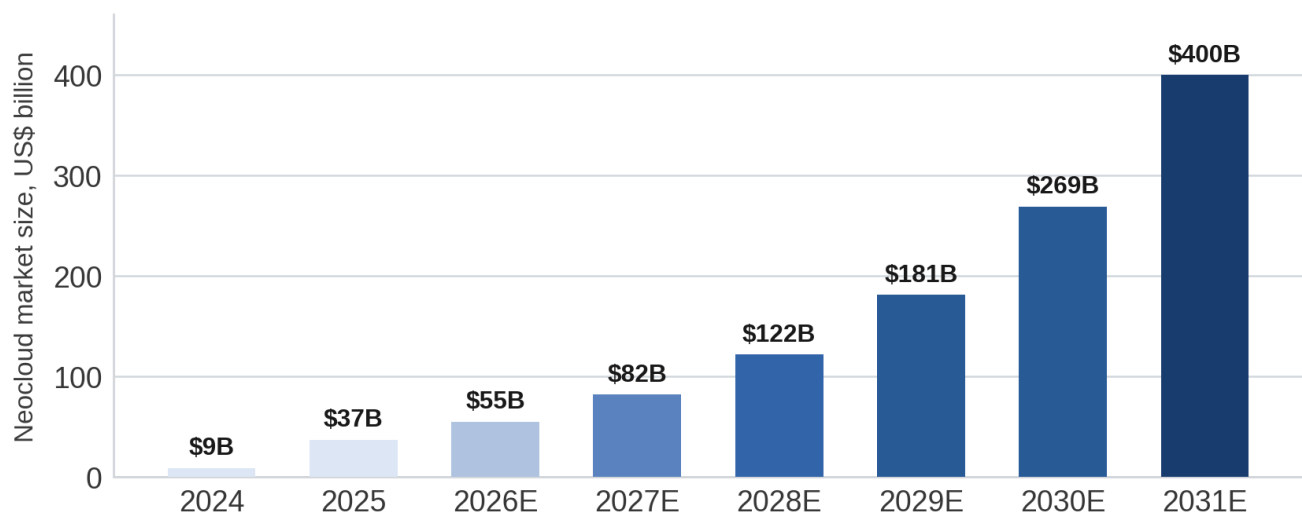
Neoclouds now among the forty largest cloud providers, Q2 2026^[4]

\$104bn

CoreWeave revenue backlog (RPO) as at 30 June 2026^[6]

Exhibit 1 · The growth path of the neocloud market

Research question: how large is this market and how fast is it growing? · **Units:** US\$ billion, annual revenue basis · **Period:** 2024–2031, forecast from 2026



Source: Synergy Research Group; full-year 2025 market size (about \$25bn) and the 2031 forecast (\$400bn at a 58 per cent CAGR) are the anchors, with intervening years interpolated by the author.^[3]

Takeaway: the market already grew 223 per cent in 2025. Synergy's 2031 projection assumes that this is momentum rather than a one-off correction; the sections that follow test where that assumption is most fragile.

The report proceeds along three axes. **Part I** sets out what a neocloud is, how large the market is and who competes in it. **Part II** examines four large operators — CoreWeave, Nebius, Lambda and Crusoe — one at a time. **Part III** turns to compute pricing, the power bottleneck, circular financing and regulatory pressure, closing with a risk register and an outlook.

RESEARCH METHOD

2. How this report was built

This report treats the neocloud as a leasing problem rather than a technology story. The organising chain runs: capital → hardware → capacity → contract → revenue → debt service. Each link is tested separately, because a failure at any one of them — a hyperscaler declining to renew, hardware losing residual value faster than modelled, a power bill that doubles — propagates through the whole structure.

Exhibit 2 · The neocloud value chain

Research question: what stages does capital pass through before it becomes revenue? · **Framework:** author synthesis, informed by the structure set out in Morgan Stanley's ecosystem study^[1]

STAGE ONE	STAGE TWO	STAGE THREE	STAGE FOUR	STAGE FIVE
Capital	Hardware	Contract	Revenue	Debt service
Equity and debt — asset-backed securities, private credit, bonds — raised on balance sheet or through a special-purpose vehicle.	GPUs and TPUs are purchased and installed in a leased colocation facility or an owned campus.	A multi-year capacity agreement is signed with a hyperscaler or a frontier AI laboratory.	Usage is billed per GPU-hour or per token, and cash begins to arrive against the installed fleet.	Revenue covers interest and principal; whatever remains is recycled into new capacity.

Takeaway: the most fragile link sits between stages four and five. The gap between operators with contracted revenue visibility (Nebius, with heavy customer prepayments) and those with less of it is the subject of Sections 10 and 16.

Source discipline

Company evidence is taken from earnings releases, exhibits filed with the Securities and Exchange Commission and shareholder letters. Market structure data comes from Synergy Research Group. Compute and token pricing evidence comes from published index providers — Silicon Data, Ornn and Compute Desk — and from the Morgan Stanley Wealth Management Global Investment Office study that compiles them. Where a figure is an estimate rather than a reported outturn, the text says so at the point of use.

WHAT IS DELIBERATELY OUT OF SCOPE

This report makes no security recommendation and sets no price target. It does not model individual company valuations. Its purpose is to describe how the neocloud model works, where its revenue actually comes from, and which observable indicators would confirm or break the case for its durability.

PART I — MARKET AND POSITIONING

3. What separates a neocloud from a hyperscaler and from colocation

The data-centre industry has historically resolved into two models. A colocation operator rents power and space, and the customer brings its own hardware. A hyperscaler owns the facility, the hardware and the software layer above them. The neocloud is a third category, a derivative of the hyperscaler cloud model that accelerated after 2016:^[1] the operator installs its own GPUs inside somebody else's building and rents them by the hour or by the token.

Exhibit 3 · Three data-centre models compared

Research question: how are capital, control and revenue risk distributed across the three models? · **Framework:** author synthesis of the conceptual structure in the Morgan Stanley study^[1]

Dimension	Colocation	Neocloud	Hyperscaler
Owns the facility	Operator	Usually leased	Usually owned
Owns the hardware	Customer	Operator, mostly debt-funded	Operator
Typical revenue per MW	\$2–3m ^[1]	\$10–23m ^[1]	\$22–50m ^[1]
Capital intensity	Moderate	Very high	Very high
Revenue model	Rent plus interconnection	Per GPU-hour or per token	Consumption-based (IaaS/PaaS/SaaS)
Principal risk	Occupancy	Tenant concentration and depreciation	Competition and return on capital

Source: Morgan Stanley Wealth Management Global Investment Office, "AI Data Centers: The Heart of the Ecosystem," 25 August 2026; revenue-per-megawatt figures compiled from company disclosure.^[1]

Takeaway: a neocloud earns three to ten times the revenue per megawatt of a colocation operator, but it earns it by financing the hardware itself — a risk the colocation operator never assumes.

Rent or own? The logic of the SPV structure

Rather than carry GPUs directly on the balance sheet, many neoclouds finance them through a special-purpose vehicle. Investors capitalise the SPV, the SPV buys the servers, and the servers are leased to a hyperscaler under a roughly five-year framework. Monthly lease payments cover the vehicle's operating costs and return capital and income to its investors. In effect the hyperscaler leases the cars rather than buying them, and the equipment returns to the neocloud or the SPV when the lease expires.^[1]

Two risks travel with that convenience, and both land on the neocloud and its investors rather than on the hyperscaler: servers may depreciate or become obsolete faster than modelled, reducing their residual value; and the hyperscaler may decline to renew, forcing the operator to find another customer or to sell the capacity into the spot market at a lower price.^[1]

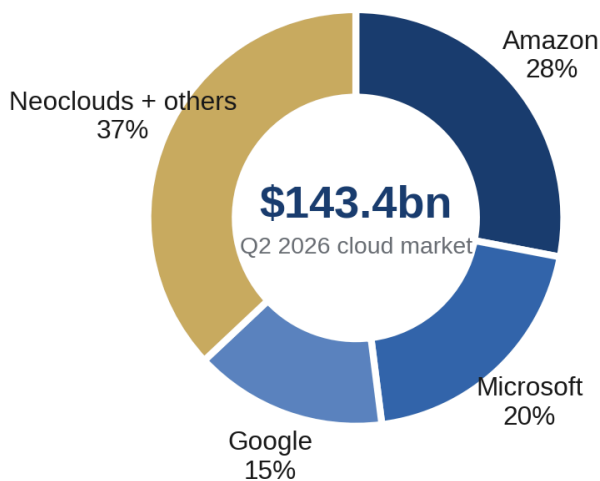
PART I

4. Market size: the smallest slice, growing fastest

The broader cloud infrastructure market reached \$143.4 billion in the second quarter of 2026, growing 43 per cent year on year — its highest rate in eight years.^[4] The engine of that growth is no longer traditional enterprise cloud. GenAI-specific cloud services grew 165 per cent, and have set the pace of the market for eleven consecutive quarters.^[4]

Exhibit 4 · Cloud market shares, Q2 2026

Units: per cent of total revenue · Period: Q2 2026



Source: Synergy Research Group.^[4]

Amazon, Microsoft and Google between them still hold 63 per cent of the public cloud market. But Synergy's analysis makes a second observation that matters more for this report: nine neocloud companies now rank among the forty largest cloud providers by infrastructure service revenue. CoreWeave, OpenAI, Oracle, Crusoe, Nebius, Anthropic and Nscale form the high-growth second tier.^[4]

That composition implies neoclouds have grown by enlarging the market rather than by taking share from the incumbents — at least so far. A separate Synergy analysis puts the pure neocloud category, excluding hyperscalers' own GPU-rental arms, at roughly \$25 billion in 2025 and approaching \$400 billion by 2031, a compound annual growth rate of 58 per cent.^[3]

Two qualifications are worth stating alongside that number. The first is definitional: Synergy counts CoreWeave, Crusoe, Core Scientific, Lambda, Nebius and Nscale as neoclouds, and classifies OpenAI and Anthropic separately as platform-centric providers. The second is that a 58 per cent compound rate sustained over six years has few precedents in infrastructure of any kind.

The size of the market is no longer the contested question. What is contested is how much of this growth rests on durable contracted demand and how much on one-off capacity agreements that will not renew.

PART I

5. The economics of an "AI factory"

The difference between a traditional data centre and an AI factory is not only scale; it is the nature of the asset. An AI factory behaves less like real estate and more like a semiconductor holding. Hardware accounts for roughly three-quarters of total project cost, and returns are acutely sensitive to GPU utilisation, to pricing, and to the hardware refresh cycle.^[1]

Exhibit 5 · Traditional data centre against AI factory

Units: US\$ per megawatt and per cent · Period: class averages as at 2026

Metric	Traditional data centre	AI factory
Revenue (US\$/MW)	\$2–3m	\$14–18m
Capital expenditure (US\$/MW)	\$10–14m	\$50–100m
EBITDA margin	50–60%	40–50%
Internal rate of return	10–13%	12–30%

Source: Morgan Stanley & Co. Research as of 24 March 2026.^[1]

Takeaway: the return range for an AI factory is wider, not simply higher — its floor is no better than a traditional data centre's. The difference sits entirely at the top of the range, which is another way of saying that the return depends on the utilisation and pricing assumption rather than on the asset class.

Training and inference produce different revenue formulas

TRAINING REVENUE

Total GPU hours × price per hour × utilisation, with limited batching available

INFERENCE REVENUE

Total GPU hours × price per token × utilisation, optimised through batching and scheduling

Inference work allows more revenue to be extracted from the same hardware through batching and scheduling, which is the economic foundation of the shift from "token-maxxing" to "token-managing" discussed in Section 13. It also means that two operators with identical fleets can report materially different revenue per megawatt depending on the mix of work they run.^[1]

PART I

6. The competitive map: four operators, four strategies

The category looks homogeneous from the outside and is strategically well differentiated from the inside. CoreWeave competes on scale and the disclosure that comes with being listed; Nebius on capital discipline and prepayment; Lambda on a brand built among researchers; and Crusoe on vertical integration that begins with energy.

Exhibit 6 · Four neoclouds compared

Research question: where does each operator sit across pricing, contract strength and capacity? · **Period:** published data as at August 2026

Company	Status	H100 price (US\$/hr)	Active / contracted power	Principal customers
CoreWeave	Listed (Nasdaq: CRWV)	\$6.16	1.5 GW / 4.2+ GW	Meta, Anthropic, OpenAI
Nebius	Listed (Nasdaq: NBIS)	\$3.85	~1 GW / 5 GW target	Microsoft, Meta
Lambda	Private, IPO in preparation	\$3.99	Undisclosed / 15 sites	Microsoft
Crusoe	Private, \$30.9bn valuation	\$3.90	~0.2 GW / 4.9 GW	Oracle/OpenAI, Microsoft, Jane Street

Sources: MarkTechPost compilation of published price lists and SemiAnalysis ClusterMAX ratings;^[5] valuations from Sacra and TechCrunch.^{[8][9]}

Takeaway: Nebius publishes the lowest H100 rate and posts the fastest growth (Section 8). CoreWeave prices 10 to 15 per cent above the group and is the only provider carrying a Platinum rating in SemiAnalysis ClusterMAX 2.0 — evidence that reliability, not scarcity alone, still commands a premium.^[5]

Note: published hourly rates apply to short-term and spot rental. Pricing under multi-year capacity agreements differs materially and is discussed in Section 12.

PART II — COMPANY STUDIES

7. CoreWeave: the leader in scale, and in leverage

Listed since March 2025, CoreWeave is the largest operator in the category and the one carrying the most visible debt. Second-quarter 2026 revenue rose 112 per cent year on year to \$2,575 million, and the revenue backlog reached approximately \$104 billion — a figure that excludes more than \$25 billion of net new customer commitments added in the opening weeks of the third quarter.^[6]

\$2.58bn

Q2 2026 revenue, up 112 per cent year on year^[6]

\$104bn

Revenue backlog (RPO), up 246 per cent

\$35.1bn

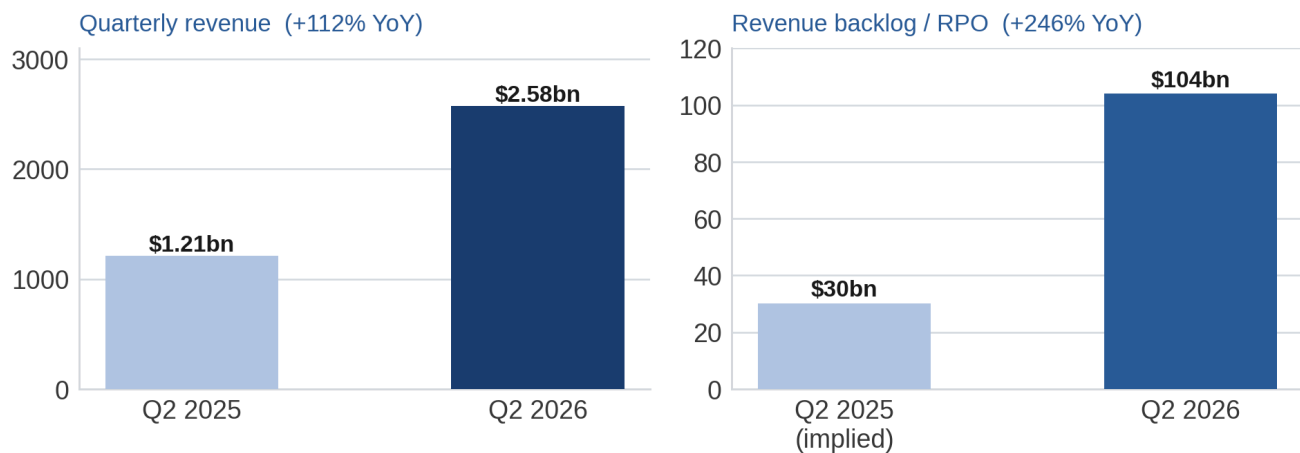
Total debt, recourse and non-recourse combined

(24)%

Net loss margin in the quarter^[6]

Exhibit 7 · CoreWeave revenue and backlog growth

Units: US\$ million and US\$ billion · Period: Q2 2025 to Q2 2026



Source: CoreWeave, Inc., second-quarter 2026 results release, filed with the Securities and Exchange Commission, 11 August 2026.^[6] The Q2 2025 backlog figure is implied by the disclosed 246 per cent growth rate and is an author calculation.

Takeaway: backlog is growing far faster than revenue, which says demand exceeds present capacity. It also says that servicing more than \$35 billion of debt depends on that backlog converting into billed revenue on schedule and in full.

"Our near-term capacity remains effectively sold out."

MICHAEL INTRATOR, CHIEF EXECUTIVE, COREWEAVE — SECOND-QUARTER 2026 EARNINGS CALL, 11 AUGUST 2026^[17]

Capital expenditure of \$6,422 million in the quarter, against a full-year 2026 programme estimated at \$35 to \$39 billion, explains both the scale of the build and the shape of the loss.^{[6][10]} Active power reached 1.5 gigawatts while total contracted power stands above 4.2 gigawatts — meaning a construction pipeline roughly three times the size of today's operating estate is still ahead of the company, and still to be financed.^[5]

PART II

8. Nebius: growth underwritten by prepayment

Nebius, formed from the former European assets of Yandex, posts the fastest growth in the category and — relative to its size — the strongest cash position. Second-quarter 2026 revenue rose 454 per cent year on year to \$582 million; annualised run-rate revenue grew 598 per cent to \$3.0 billion, against full-year guidance of \$7 to \$9 billion.^[7]

\$582m

Q2 2026 revenue, up 454 per cent year on year^[7]

41%

Adjusted EBITDA margin in the quarter

\$8bn

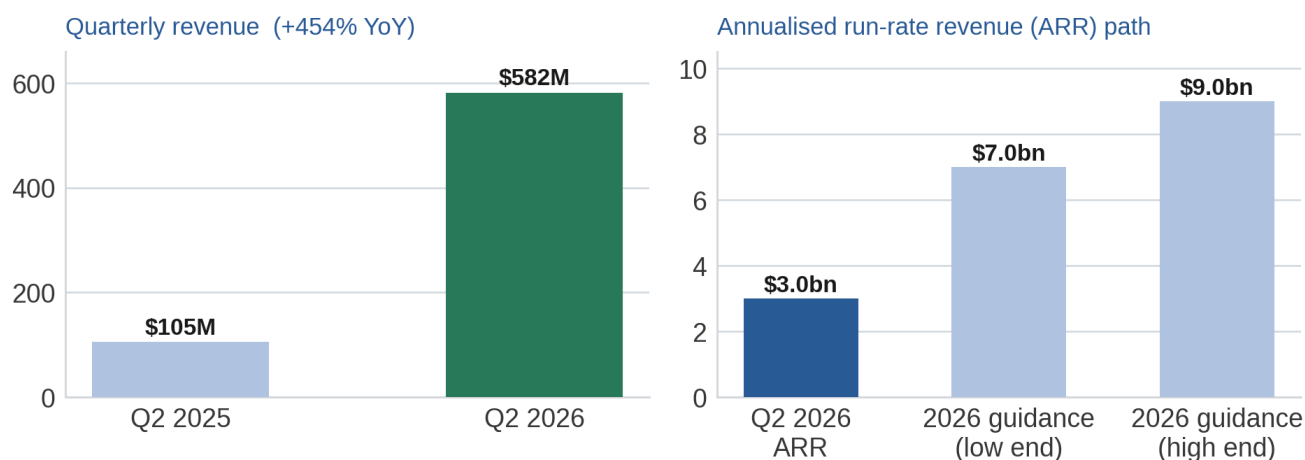
Cash position

\$40bn

Contracted backlog

Exhibit 8 · Nebius revenue growth and the ARR path

Units: US\$ million and US\$ billion · Period: Q2 2025 to full-year 2026 guidance



Source: Nebius Group N.V., second-quarter 2026 shareholder letter, 12 August 2026.^[7]

Takeaway: Nebius's advantage lies less in the growth rate than in how the growth is funded. Four new contracts — with Reflection, Cohere, a major US AI laboratory and a large quantitative trading firm, each above \$1 billion — carry prepayments covering 50 to 60 per cent of the associated capital expenditure.^[7]

That prepayment structure softens the debt-financing risk set out in Section 10. More than \$9 billion of expected customer prepayments in 2026, together with a \$775 million asset-backed facility secured in July at SOFR plus 250 basis points, cover a substantial share of capital expenditure before any of it has to be borrowed.^[7] Against a year-end target of 5 gigawatts of contracted power, connected power is expected to reach only 800 megawatts to 1 gigawatt by December — the gap between what is contracted and what is energised remains the central timing risk in this business.^[7]

PART II

9. Lambda and Crusoe: two routes to the public market

Lambda and Crusoe are the two largest operators still in private hands, and both are close to a listing. Each raised its valuation sharply during 2026, but they arrived there by different routes: Lambda through a brand built among researchers and developers, Crusoe through vertical integration that starts at the power source rather than at the rack.

Lambda

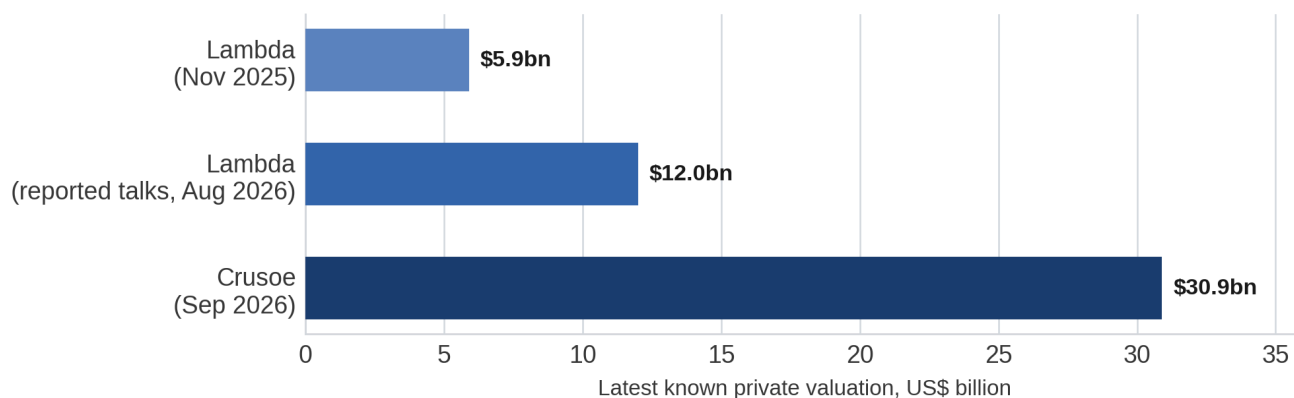
Calendar 2025 revenue reached approximately \$700 million, up 65 per cent year on year, and management has guided to more than \$1.5 billion for 2026.^[9] A Series E round in November 2025 valued the company at \$5.9 billion post-money; in August 2026 pre-IPO financing discussions were reported at a valuation above \$12 billion, though those talks were described as ongoing rather than concluded.^{[9][11]} A multi-billion-dollar agreement with Microsoft is the company's largest single contract, and also its largest single concentration.

Crusoe

Founded in 2018 as a crypto-mining operation powered by flared natural gas, Crusoe turned to AI infrastructure as compute demand accelerated. A \$3.9 billion Series F in September 2026 — co-led by Atreides Management, Mubadala Capital and Valor Equity Partners, with Nvidia, GIC, Founders Fund and the Qatar Investment Authority participating — set the valuation at \$30.9 billion.^[8] The company signed a \$13 billion, five-year cloud services contract with Jane Street and is working with Goldman Sachs and Morgan Stanley on a possible listing.^[8]

Exhibit 9 · Private valuations compared

Units: US\$ billion, latest completed round or reported discussion · Period: November 2025 to September 2026



Sources: Sacra (Lambda); TechCrunch (Crusoe); Forge Global (reported Lambda pre-IPO discussions).^{[8][9][11]}

Takeaway: Crusoe's valuation rose close to threefold in a year, and most of that increase came from the energy, construction and modular "AI factory" businesses rather than from GPU rental. That makes it structurally different from the other three names in this report.

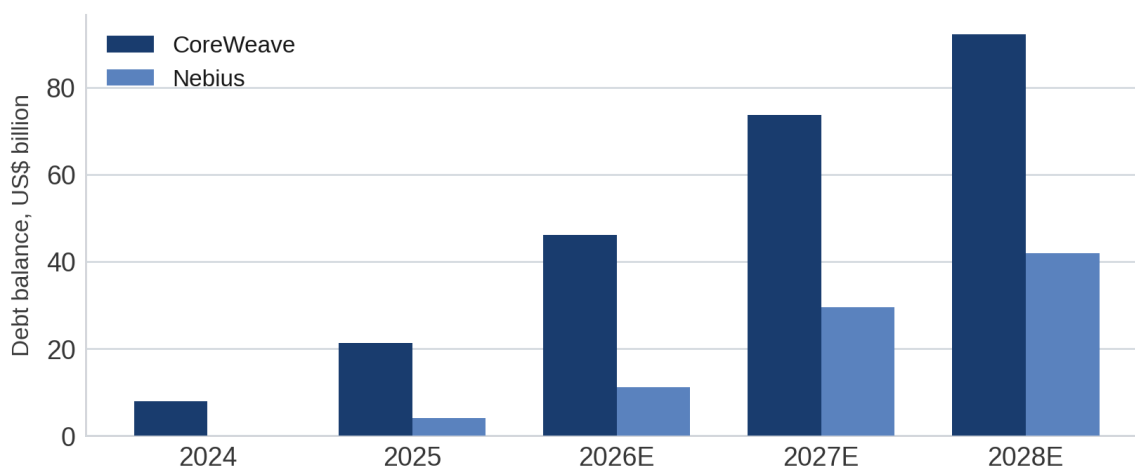
PART III — FINANCING AND RISK

10. Why neoclounds borrow so heavily

Neocloud cash flow cannot keep pace with the rate at which capacity is added, so almost all growth is debt-financed. Multi-year hyperscaler agreements are the collateral that gives lenders revenue visibility. In plain terms: these companies borrow to build a fleet before they have earned the cash to pay for it.^[1]

Exhibit 10 · Projected debt balances, CoreWeave and Nebius

Research question: how quickly does the debt burden accumulate? · **Units:** US\$ billion · **Period:** 2024–2028 estimated



Source: Morgan Stanley & Co. Research as of 15 April 2026.^[1]

Takeaway: CoreWeave's debt is projected to grow roughly twelvefold between 2024 and 2028. It tracks contracted backlog rather than accumulating independently of the business — but it does mean the equity case and the credit case have become one case.

THE TWO STRUCTURAL RISKS IN THE DEBT MODEL

Depreciation risk. Servers may lose value, or become obsolete, faster than the financing assumes — eroding the collateral pledged to lenders precisely when it is needed.

Renewal risk. If a hyperscaler declines to renew, the capacity must be placed elsewhere or sold into the spot market at a lower price — while the debt schedule is unchanged.^[1]

The scale of that collateral is visible from the other side of the contract. Microsoft's committed GPU leasing with neoclounds has passed \$60 billion — roughly \$23 billion with Nscale, \$9.7 billion with Iren, and multi-billion-dollar agreements with Lambda, Nebius and CoreWeave.^[19] A hyperscaler leasing at that scale is, in effect, underwriting the neocloud's debt without appearing on its balance sheet.

The concentration figures make this concrete. Microsoft, OpenAI and Meta together represented more than 60 per cent of CoreWeave's annual recurring revenue in 2025, a share projected to fall to roughly 36 per cent by 2028 as the customer base widens — but until then the company depends on a few very large buyers whose capital plans are revised annually.^[1]

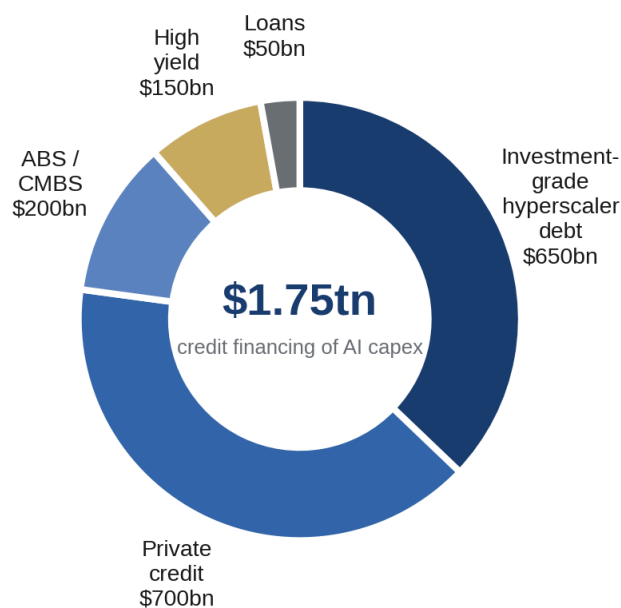
PART III

11. The \$1.75 trillion credit spine beneath AI capex

AI data-centre capital expenditure could approach \$3.2 trillion over the next three years. Analysts expect the majority of that investment — around \$1.75 trillion — to come from credit financing: investment-grade bonds, private credit, asset-backed securities, high-yield debt and loans.^[1]

Exhibit 11 · Composition of AI capex credit financing

Units: US\$ billion · Period: 2026–2028 outlook



Source: Morgan Stanley Wealth Management Global Investment Office as of 29 May 2026.^[1]

The five largest US hyperscalers' outstanding debt represents about 4 per cent of the Bloomberg US Corporate Index, against roughly 20 per cent of the S&P 500's market capitalisation. Year-to-date issuance has exceeded \$180 billion, including more than \$50 billion outside the US dollar market.^[1]

Hyperscaler leverage has risen from 0.9 times to around 1.8 times, still below the broader non-financial corporate average of 2.3 times.^[1] But AI-related debt trades with wider spreads than its ratings imply: instead of trading like their AA-rated peers, many of these issues now trade more like A-rated debt, particularly at longer maturities. That is the credit market registering a view about the quality and duration of the growth, not merely its quantity.^[1]

Demand for new issuance has softened somewhat, reflected in lower subscription ratios. Insurance companies are the traditional incremental buyer of thirty-year corporate paper, and their appetite has weakened — one reason long-dated spreads have widened.^[1]

WHAT TO WATCH

Investment-grade issuance for 2026 is tracking at \$250 to \$300 billion, and some estimates put next year's pipeline as high as \$400 billion. AI financing has also begun to play a significant role in adjacent credit markets — non-dollar investment grade, high yield, leveraged loans and private credit — which is how a sector-specific problem would become a broader one.^[1]

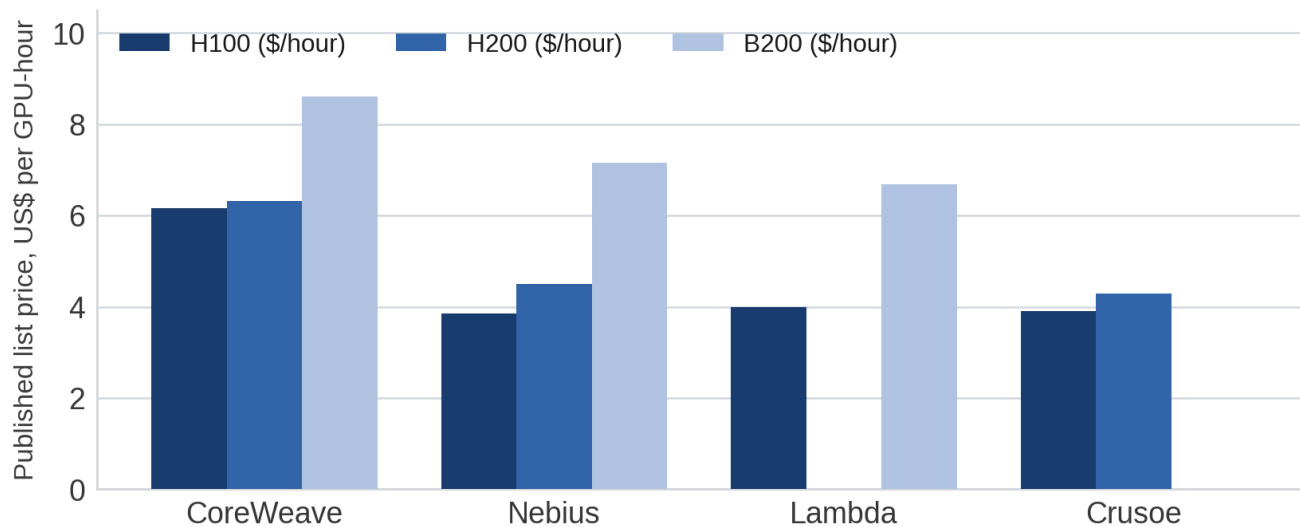
PART III

12. Compute pricing: a new layer of transparency

The GPU rental market was opaque for most of its short history: prices were negotiated privately rather than published. That changed across 2025 and 2026. Silicon Data, Ornn and Compute Desk now publish live indices for GPU rental rates and token prices, and those indices are tracked on Bloomberg.^{[1][2]}

Exhibit 12 · GPU pricing by provider

Research question: how far apart are providers on the same silicon? · **Units:** US\$ per GPU-hour, published list price · **Period:** August 2026



Source: MarkTechPost compilation of published provider price lists.^[5]

Takeaway: the spread on H100 reaches 60 per cent — \$6.16 at CoreWeave against \$3.85 at Nebius. Scarcity has not produced a single clearing price, which means brand, reliability and cluster quality still carry a measurable premium.

Prices for access to Nvidia's Blackwell and Vera Rubin parts are "setting new highs", while even older inventory commands rental rates not seen in years.

MICHAEL INTRATOR, CHIEF EXECUTIVE, COREWEAVE — SECOND-QUARTER 2026 EARNINGS CALL^[17]

That is the seller's account of the same market the indices measure, and the two are not in conflict: list rates for the newest silicon can rise while the blended token price falls, because they price different things — scarce current-generation capacity on one side, the cost of serving a million tokens on the other.

Ornn's Compute Price Index family publishes transaction-based reference prices for H100, H200, B200, B300 and RTX5090. The CME plans to launch cash-settled futures on H100s and B200s based on the Silicon Data indexes — another sign that compute is becoming a standardised, tradable and therefore hedgeable market.^{[1][2]}

PART III

13. How open-weight models compress neocloud pricing

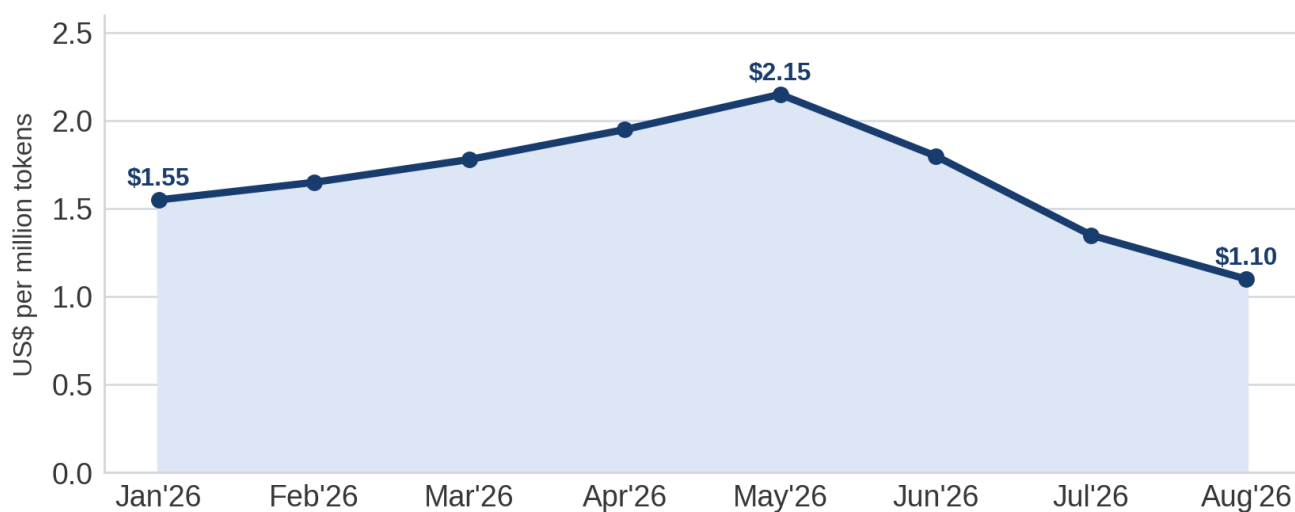
AI models fall into two categories — closed-weight, which stay under the provider's control and are accessed by subscription or API, and open-weight, which a customer can download and run on its own systems. The distinction bears directly on neocloud revenue: a customer running an open-weight model can buy compute and run it anywhere, while a customer on a closed-weight model pays both a licensing fee and a token bill.^[1]

THREE OBSERVATIONS FROM MORGAN STANLEY & CO. RESEARCH

1) More than 50 per cent of enterprise usage is now routed to open-weight models. 2) Open-weight models introduce challenges for token economics and put downward pressure on frontier-model prices. 3) Chinese model creators have deployed open-weight models unusually fast, facing fewer power and infrastructure bottlenecks; US frontier models carry greater exposure to project delays and to regulatory and political backlash.^[1]

Exhibit 13 · Token expenditure index, redrawn

Units: US\$ per million tokens · Period: January to August 2026



Source: Silicon Data LLM Token Expenditure Index, approximated from the series published in the Morgan Stanley study; Ormn's OCPI series shows comparable softening.^{[1][2]}

Takeaway: token pricing peaked above \$2.00 in May 2026 and retreated to around \$1.10 by August — consistent with the open-weight shift described below.

How Chinese competition affects price

Chinese open-weight models arrive with a substantial cost advantage: the cost of one million input plus one million output tokens is a fraction of comparable US frontier models. The gap matters more because Chinese developers have continued to improve despite export controls on Nvidia's highest-performance parts; domestically designed accelerators from Huawei and Cambricon may over time produce a more independent Chinese AI infrastructure stack.^[1]

Meta's recent decision to re-embrace open-weight models points the same way: weaker frontier-model demand would put the addressable demand behind current capital expenditure in question.^[1]

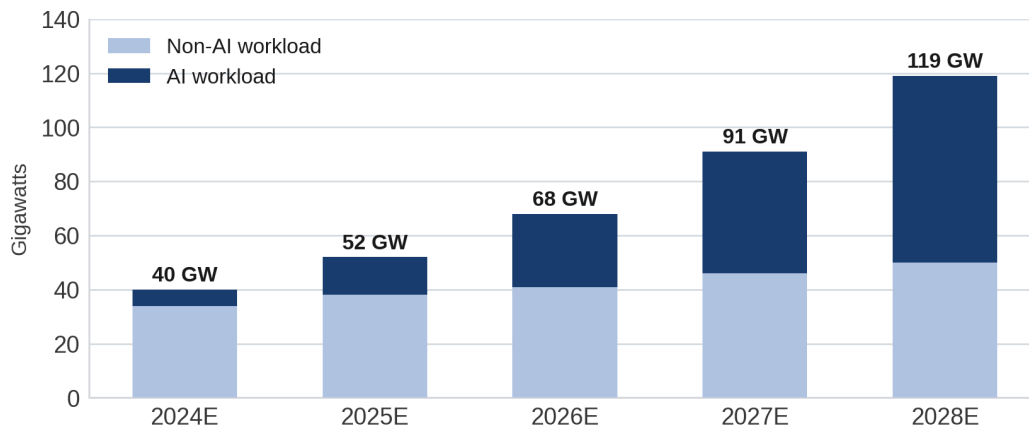
PART III

14. The power bottleneck: the hardest physical constraint

AI data centres become enormous consumers of electricity the moment they are energised. Morgan Stanley & Co. Research anticipates 81 per cent annualised growth in AI data-centre power demand, against 10 per cent for traditional ones — more than eight times the rate, compounding.^[1]

Exhibit 14 · US data-centre power demand, AI and non-AI workloads

Units: gigawatts · Period: 2024E–2028E



Source: Morgan Stanley & Co. Research as of 10 June 2026.^[1]

Takeaway: AI workloads were roughly 15 per cent of total data-centre power in 2024 and are projected to reach 58 per cent by 2028 — from near zero to the equivalent of all traditional data centres combined, in three years.

A one-gigawatt AI data centre costs an estimated \$30 to \$50 billion, roughly 75 per cent of it IT hardware; the 33.2 gigawatts planned across 2026 and 2027 carry a price tag near \$1.7 trillion.^[1] Nvidia's chief executive puts the unit cost higher — \$50 to \$60 billion per gigawatt.^[20] The Electric Power Research Institute estimates incremental demand of about 37 gigawatts by 2030 from high-confidence projects; the United Kingdom's entire generation capacity is about 33 gigawatts.^[1]

THE POWER VIEW

New AI data centres are expected to struggle to secure grid connections, so owners increasingly build behind-the-meter: off-grid plants on gas pipelines, modular turbines or storage. These can move power costs from \$0.10–\$0.12 per kilowatt-hour to \$0.20–\$0.22.^[1]

Power is 15 to 20 per cent of operating costs, so a doubling compresses margins directly while the capital behind it stays fixed and up-front.^[1]

"The biggest issue we are now having is not a compute glut, but it's power... I may actually have a bunch of chips sitting in inventory that I can't plug in."

SATYA NADELLA, CHIEF EXECUTIVE, MICROSOFT^[19]

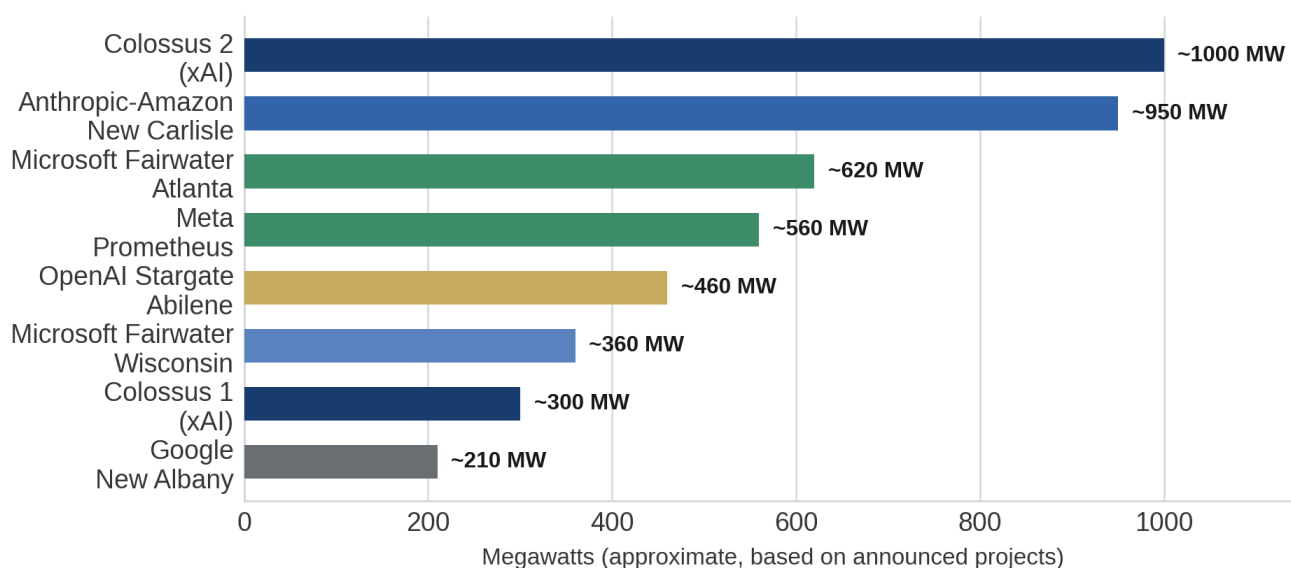
PART III

15. Mega-campuses: capacity concentrating in a handful of projects

Long-term power demand estimates remain early. Many of the largest projects have entered only their initial phases of development and will ultimately reach gigawatt scale. Data compiled by Epoch.AI shows that the biggest AI data-centre projects now carry, as individual campuses, power demands approaching those of a mid-sized city grid.^[1]

Exhibit 15 · Power capacity of the largest AI data-centre campuses

Research question: in how few projects is capacity concentrating? · **Units:** megawatts, approximate · **Period:** announced and under-construction projects as at July 2026



Source: Epoch.AI as of 8 July 2026, redrawn with approximate readings from the published series.^[1]

Takeaway: eight campuses alone account for several gigawatts of demand. At that scale a project is only feasible with dedicated generation, a long-dated grid agreement, or both — which explains why neoclouds are increasingly entangled with energy companies, and why Crusoe, which began as one, is structurally advantaged.

Geography is becoming a cost variable in its own right. Countries with higher power costs, such as those in Europe and North Asia, face challenges participating in the AI build-out owing to limitations from regulation, natural resources and infrastructure. Power availability will likely support continued leadership from the United States and China in the AI capex boom — a geopolitical wrinkle in what otherwise looks like a purely industrial problem.^[1]

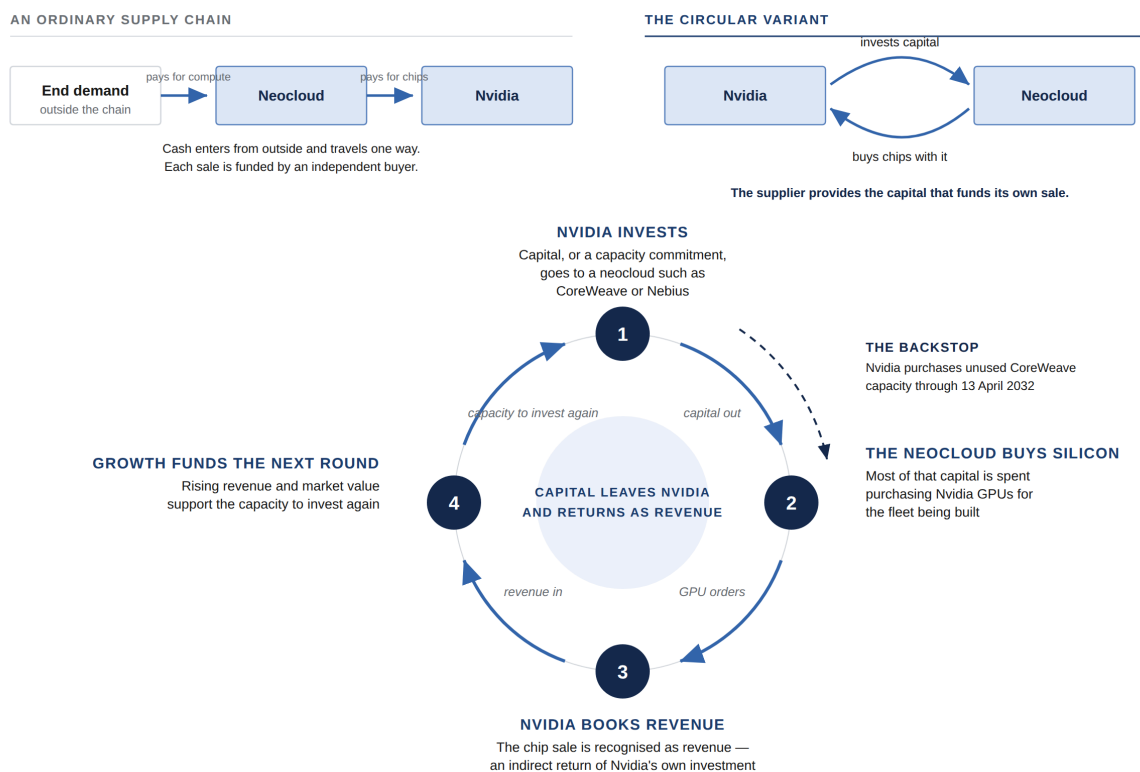
PART III

16. Circular financing: Nvidia as both supplier and investor

The risk that drew most investor attention in 2026 was structural rather than operational: Nvidia invests capital into neoclouds, those neoclouds use the capital largely to buy Nvidia processors, and the loop registers as growth on both sides of the transaction.^[10]

Exhibit 16 · The anatomy of the circular financing loop

Research question: how does a supplier come to fund its own sale? · **Framework:** author synthesis, compiled from public reporting^[10]



Sources: reporting on Nvidia's investments in, and capacity commitments to, CoreWeave and Nebius.^[10] The upper panels contrast an ordinary supply chain with the circular structure and are the author's construction.

Takeaway: the structure lowers bankruptcy risk — Nvidia's balance sheet functions as a backstop — while obscuring how much neocloud revenue reflects independent end demand and how much reflects vendor financing. Those are different qualities of revenue and they should not be valued identically.^[10]

\$35bn

CoreWeave balance-sheet debt, now tracked by lenders including pension funds^[10]

\$5.75bn

Nebius convertible note offering

2032

Expiry of Nvidia's CoreWeave capacity purchase commitment

(41)% / (30)%

CoreWeave and Nebius shares below their all-time highs, September 2026

The counterargument deserves weight. Structural parallels to the dot-com cycle are real, but Nvidia holds a balance sheet with ample capacity to backstop these commitments, which makes a wave of bankruptcies less likely than the comparison implies — even if it does nothing to protect equity holders from a re-rating.^[10]

PART III

17. Is Nvidia alone? AMD, Broadcom and the shape of supplier support

Nvidia is not the only supplier taking a financial position in its own customers, but the structures differ in ways that matter more than the headlines suggest — and both comparisons below sharpen what the loop in Section 16 does and does not represent.

AMD: warrants in place of capital

In October 2025 AMD agreed with OpenAI to deploy 6 gigawatts of Instinct accelerators, granting OpenAI warrants over up to 160 million AMD shares — roughly 10 per cent of the company — at one cent per share, vesting in tranches as capacity is purchased and as AMD's own share price clears defined thresholds.^[13] In February 2026 AMD signed a practically identical agreement with Meta: another 6 gigawatts, another 160 million warrants, with the first gigawatt of custom MI450 silicon shipping in the second half of 2026.^[12] Together the two commitments represent 12 gigawatts.

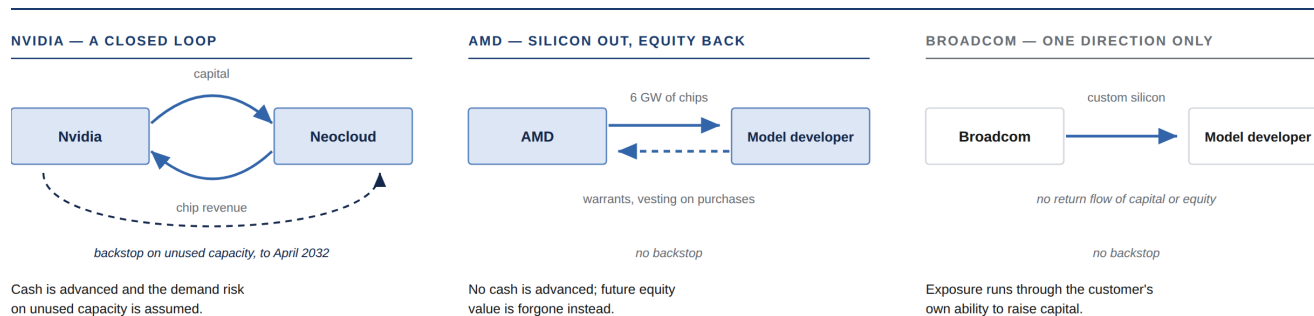
This is not what Nvidia does. AMD advances no cash; it forgoes future equity value, which makes the arrangement closer to a volume discount denominated in shares than to vendor financing. Nor does it assume the customer's demand risk.

Broadcom: co-design without an equity loop

Broadcom's 10-gigawatt collaboration with OpenAI on custom accelerators is a design and supply relationship rather than a financing one; the first chip from the programme was unveiled in June 2026.^[14] The \$18 billion financing gap reported there in May 2026 makes the distinction plain: Broadcom is not funding its customer, so a customer that struggles to raise capital threatens Broadcom's schedule rather than its credit.^[15]

Exhibit 17 · Three shapes of supplier support

Research question: what flows back to the supplier, and what does not? · **Period:** arrangements announced October 2025 – September 2026



Sources: reporting on Nvidia's neocloud arrangements;^[10] The Register and AMD investor relations on the Meta agreement;^[12] OpenAI on the AMD partnership;^[13] Broadcom and CNBC on the OpenAI collaboration.^[14]

Takeaway: only the first structure closes. AMD's warrants travel back as equity rather than as cash and carry no obligation to absorb unused capacity; Broadcom's relationship has no return flow at all.

PART III

18. What would bring an ASIC vendor into the loop

The three structures answer one question differently: who carries the loss if the capacity is never used? That is the variable worth tracking, and it separates the suppliers more cleanly than the gigawatt figures in their announcements do.

Exhibit 18 · Three structures of supplier support compared

Research question: what does each supplier actually put at risk? · **Period:** arrangements announced October 2025 – September 2026

Supplier	Structure	Counterparty	What is put at risk	Exposure to neoclouds
Nvidia	Capital or capacity commitment, plus a purchase backstop to April 2032	Neoclouds (CoreWeave, Nebius)	Cash, and the demand risk on unused capacity	Direct
AMD	Warrants over up to 160m shares at \$0.01, vesting against purchases	Model developers (OpenAI, Meta)	Future equity value, not cash	Indirect
Broadcom	Co-design and supply of custom accelerators, no equity loop	Model developer (OpenAI)	Schedule and capacity reservations, not credit	None at present

Sources: The Register and AMD investor relations; ^[12] OpenAI; ^[13] Broadcom and CNBC; ^[14] reporting on the programme's financing gap; ^[15] reporting on Nvidia's neocloud arrangements. ^[10]

Takeaway: the decisive variable is the counterparty. AMD and Broadcom have contracted with model developers; the documented arrangement that pairs capital with a capacity backstop for a *neocloud* remains Nvidia's alone as at the date of this report. Section 16 therefore describes one supplier's structure, not an industry norm.

Could an ASIC vendor enter the loop later? The route runs through the threshold identified in Section 22: the diffusion of custom silicon. Broadcom sits one step removed from the fleets neoclouds operate; should those fleets start carrying custom accelerators at scale, the financing question migrates to whoever underwrites them.

THREE THINGS WOULD HAVE TO BECOME TRUE FIRST

1) Neoclouds would have to host custom accelerators at scale rather than GPUs, which today they largely do not. 2) The vendor would have to accept balance-sheet risk in place of schedule risk — a financing vehicle neither Broadcom nor its peers operates at present. 3) Custom silicon would need a secondary market: a part designed for one customer is weaker collateral than a GPU that any operator can rent out, and lenders price that difference.

PART III

19. Regulatory and political pushback is building

Data centres have recently faced increased regulatory scrutiny from state governments, in response to concerns around electricity demand, transmission costs and water usage. The most consequential measures fall into three categories: restrictions on new development, requirements that data centres bear more of their infrastructure costs, and tighter approval and disclosure requirements.^[1]

Exhibit 19 · Notable regulatory measures at state and federal level

Period: in force or announced as at 2026

Jurisdiction	Measure	Effect
Virginia	Data centres required to pay for transmission infrastructure built to serve them, plus an energy-consumption tax	Direct cost increase
Oregon	POWER Act, a "pay-your-own-way" regime effective July 2026	Average 29 per cent rate increase for large data centres served by Portland General Electric
Texas (ERCOT and PUC)	Governor-ordered audit of all data-centre projects	Review of power and water use, public subsidies, on-site generation, community impact and ownership
Federal (FERC)	Show-cause orders issued to six grid operators on large-load interconnection, co-location and cost allocation	Pressure to standardise connection rules

Source: Morgan Stanley & Co. Research, state and federal regulatory developments compiled as of 5 December 2026.^[1]

Takeaway: Texas has historically been considered among the most favourable regulatory regimes for business and energy development. That it has ordered an audit is the clearest evidence that the shift towards tighter power regulation is broad rather than local.

Voter anxiety — about both the physical presence of data centres and the effect on household power costs — is likely to garner popular support and may play heavily in this year's midterm election cycle. For an operator planning a three-year build, that implies regulatory friction is more likely to increase than to ease over the life of the current capital programme.^[1]

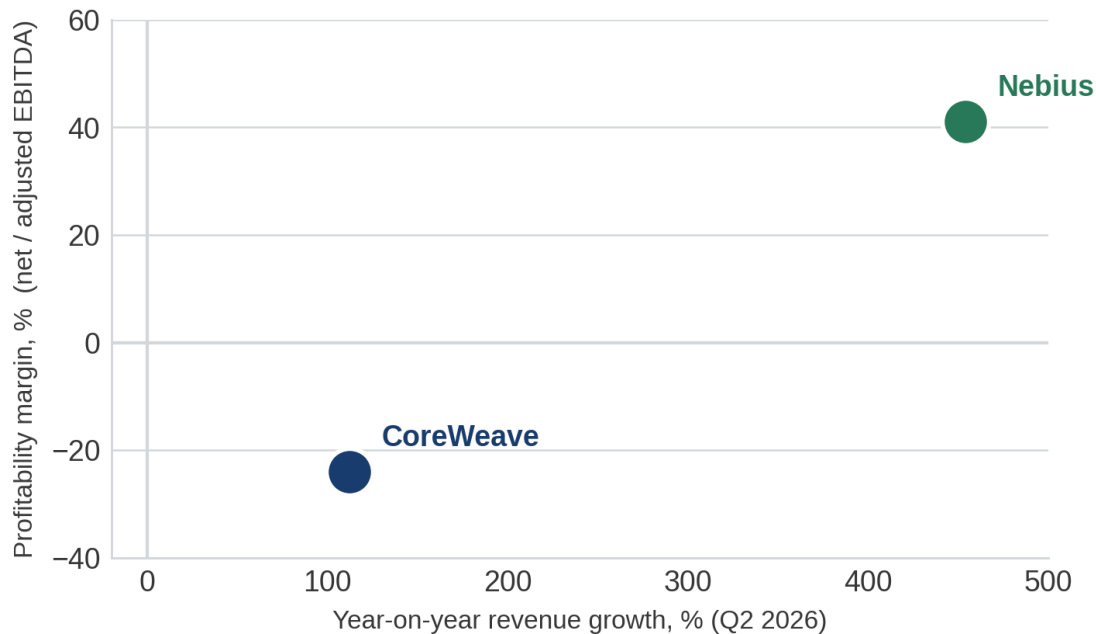
PART III

20. Reading growth and profitability together

Placing the two listed neoclouds on the same axes shows how early this sector remains in financial terms. The construction below is an original calculation derived from the company data presented in Sections 7 and 8.

Exhibit 20 · The growth–profitability scorecard

Research question: which operator is converting growth into profit? · **Units:** per cent · **Period:** Q2 2026, most recent reported quarter



Source: author calculation from CoreWeave and Nebius second-quarter 2026 results releases.^{[6][7]} Note: GAAP net margin is used for CoreWeave and adjusted EBITDA margin for Nebius; the two measures are not directly comparable.

Takeaway: Nebius sits in the quadrant with both faster growth and a positive margin. CoreWeave's growth is strong but its negative net margin shows that scale has not yet converted into profit — a natural consequence of a heavy capex cycle, and a difference worth tracking quarter by quarter.

"If we stop growing at the rate we're growing, this company would become enormously profitable within three months. We're making the strategic decision to get to scale."

MICHAEL INTRATOR, CHIEF EXECUTIVE, COREWEAVE — CNBC INTERVIEW, APRIL 2026^[18]

The claim cannot be verified from outside, but it does describe what the chart measures: a negative margin presented as a decision rather than a constraint.

THE HONEST LIMIT OF THIS METRIC

The comparison rests on a single quarter and ignores the balance sheet entirely (Section 10). CoreWeave's negative margin is driven largely by capital expenditure and interest expense; Nebius's positive margin is a product of its prepaid contract structure. Neither measures durable profitability — both measure today's cash-flow picture.

PART III

21. Risk register: what to watch, and where

The table below gathers the risks treated across this report into one place, and names for each of them an observable indicator that would show the risk crystallising rather than merely persisting.

Exhibit 21 · Risk register and monitoring indicators

Framework: author synthesis, drawing on the sources cited in Sections 10 to 19

Risk	Why it matters	Indicator to monitor
Power bottleneck	Behind-the-meter supply can double the cost per kilowatt-hour ^[1]	Connected megawatts against targeted megawatts
Tenant concentration	Dependence on a small group of very large buyers	Revenue share of the three largest customers
Hardware depreciation	GPU residual value may fall faster than financed	Interval between accelerator generations
Circular financing	Vendor financing obscures independent end demand	Share of revenue and funding linked to Nvidia
Open-weight price pressure	Erodes pricing power on frontier-model capacity	Direction of the token expenditure index
Regulatory and political pushback	New connections delayed, costs pushed onto operators	State legislation count; midterm election agenda
Debt service pressure	Leverage priced wider than the credit rating implies	Option-adjusted spread on AI bonds against the IG index

Source: synthesis of all sources cited in Sections 10 to 19 of this report. ^{[1][2][4][5][6][7][8][9][10]}

Takeaway: only two of these seven risks are within an operator's control. The rest — power, regulation, silicon cycles, the pricing of open-weight models — are set elsewhere, which is the defining feature of the neocloud position.

The sceptic's case, stated plainly

The sharpest public challenge to this sector's economics has come not from a competitor but from the chief executive of a large enterprise software company. Following Nvidia's second-quarter results in August 2026, Nikesh Arora of Palo Alto Networks argued that the capital structure behind the build-out holds only while the market stays euphoric.

"In two years from now you will be able to buy a neocloud for less than they raise at today... When supply demand finds balance, equity will normalise. Equity to fund capex only works in a euphoric market."

NIKESH ARORA, CHAIRMAN AND CHIEF EXECUTIVE, PALO ALTO NETWORKS — AUGUST 2026^[16]

He had made the related point on his own company's June earnings call: when data-centre growth starts to taper off or plateau, "this thing is going to come to roost."^[16] Arora nonetheless draws a distinction that the register above implies rather than states. Operators that reach sufficient scale, lock in major customers, secure power and add a software layer — he calls these "neoscalers" and names Nebius as one — are a different proposition from those that rent GPUs by the hour and do nothing else. On that reading the risk is not that the category fails, but that it separates.^[16]

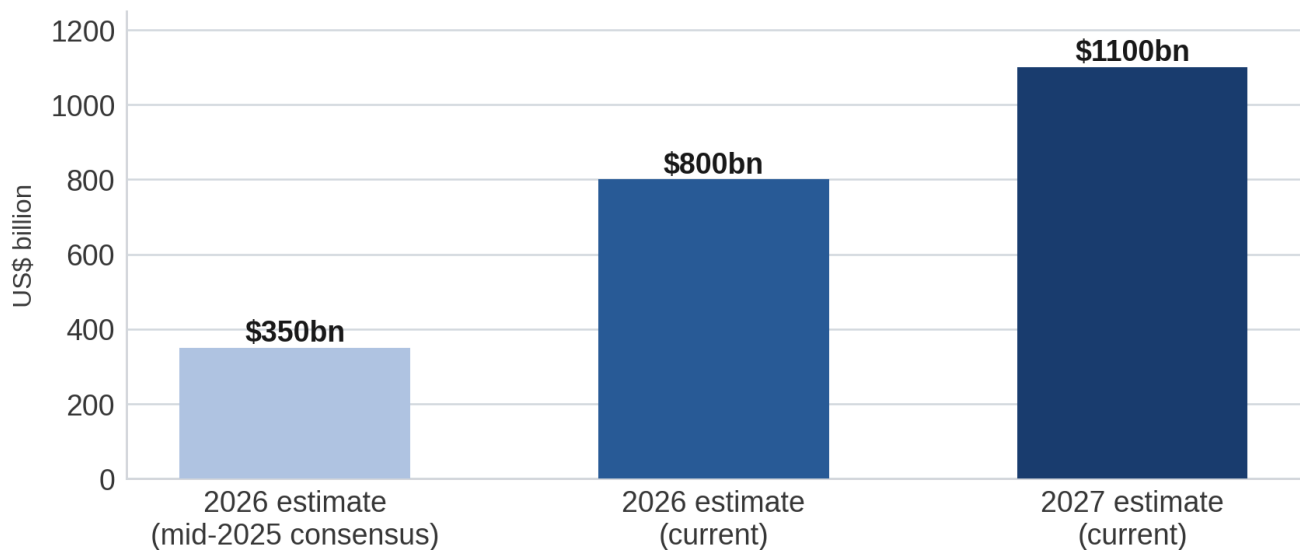
PART IV — OUTLOOK

22. Why the forecasts keep being revised upward

One of the most striking features of this capex cycle is that consensus estimates have consistently lagged and then been revised sharply higher. In mid-2025, sell-side estimates reflected hyperscaler capex expectations of \$350 billion for 2026. A year later those estimates had ballooned to \$800 billion, with spending forecast to hit \$1.1 trillion in 2027.^[1]

Exhibit 22 · The re-rating of consensus hyperscaler capex estimates

Units: US\$ billion · Period: mid-2025 consensus against current estimates, August 2026



Source: Morgan Stanley Wealth Management Global Investment Office, August 2026.^[1]

Takeaway: a 129 per cent upward revision in a single year may mean the market still underestimates demand. It equally shows how fragile the estimates themselves are. Revisions of that magnitude cannot continue indefinitely; the question is when, and how sharply, they stop.

Three thresholds to watch before 2027

- **The IPO window.** Lambda and Crusoe would give the category its first fully audited, comparable financial statements, making comparisons of the kind in Section 20 far more reliable.
- **The spread of custom silicon.** Application-specific chips could weaken any single chipmaker's monopoly and materially lower the hardware cost of the next gigawatt of capacity.^[1]
- **The third-quarter debt-service test.** CoreWeave's and Nebius's Q3 2026 results will show how much backlog converted into billed revenue on time — the first concrete test of the risks in Section 10.

CONCLUSION

23. The bottom line

The neocloud model answers a real and large need: it lets hyperscalers expand GPU supply quickly without carrying the assets on their own balance sheets. As at the second quarter of 2026 the scale of that answer is no longer in question. CoreWeave alone carries a \$104 billion revenue backlog, Nebius is growing at 454 per cent a year, and nine neoclouds now rank among the forty largest cloud providers in the world.^{[4][6][7]}

Scale should not be mistaken for durability. What this report describes is a sector growing very fast, financed without much margin for error, and dependent on a handful of very large customers and on a single supplier that is also one of its investors.

Two paths, and the likely middle

THE FAVOURABLE CASE

Demand keeps growing, prepaid structures of the Nebius type become standard practice, custom silicon lowers the hardware bill, and public listings improve disclosure. Debt service proceeds uneventfully because backlog converts on schedule.

THE ADVERSE CASE

Open-weight models erode pricing power further, regulatory friction slows energisation, a large hyperscaler declines to renew, and one operator whose debt service exceeds its revenue triggers a broader repricing of the category's credit.

The likely outcome sits between the two. The sector consolidates, with the strongest balance sheets and the most diversified customer bases surviving; pricing grows steadily more transparent and more commoditised as index and futures markets mature; and Nvidia's backstop commitments mean the correction, when it comes, is selective rather than systemic. For investors and decision-makers the question is no longer whether this market will grow. It is which part of that growth rests on durable contracted demand, and which part rests on capacity that will not be renewed.

This report was compiled from publicly available sources as at 21 September 2026. Company figures reflect the most recently published quarterly results and are subject to change. Nothing in this document constitutes investment advice.

APPENDIX

References

#	Source
[1]	Morgan Stanley Wealth Management, Global Investment Committee, "AI Data Centers: The Heart of the Ecosystem," Special Report, 25 August 2026.
[2]	Ornn, Ornn Compute Price Index (OCPI), ornn.com, accessed September 2026.
[3]	Synergy Research Group, "Neocloud Market Forecast to Approach \$400B by 2031, Driven by Surging AI Infrastructure Demand," 2026.
[4]	Synergy Research Group, "Q2 Cloud Market Passes \$143 Billion; Highest Growth Rate in Eight Years," 2026.
[5]	MarkTechPost, "Best GPU Neoclouds 2026: CoreWeave, Nebius, Lambda, Crusoe, and Groq Ranked by Published Pricing and Contracted Power," 23 August 2026.
[6]	CoreWeave, Inc., "CoreWeave Reports Strong Second Quarter 2026 Results," exhibit filed with the Securities and Exchange Commission, 11 August 2026.
[7]	Nebius Group N.V., Second Quarter 2026 Shareholder Letter, 12 August 2026.
[8]	TechCrunch, "Crusoe raises \$3.9B to build massive data centers and small modular 'AI factories'," 17 September 2026.
[9]	Sacra, "Lambda Labs: revenue, valuation and funding," 2026.
[10]	The Motley Fool, "Should CoreWeave and Nebius Group Investors Be Worried About Circular Financing?," 3 September 2026.
[11]	Forge Global, "Lambda IPO Timeline and Financing Details," 2026 (reported pre-IPO valuation discussions).
[12]	The Register, "Meta signs 6 GW chips-for-stock deal with AMD," 24 February 2026; AMD Investor Relations, "AMD and Meta Announce Expanded Strategic Partnership to Deploy 6 Gigawatts of AMD GPUs," 2026.
[13]	OpenAI, "AMD and OpenAI announce strategic partnership to deploy 6 gigawatts of AMD GPUs," October 2025.
[14]	Broadcom Inc., "OpenAI and Broadcom announce strategic collaboration to deploy 10 gigawatts of OpenAI-designed AI accelerators," 2025; CNBC, "OpenAI unveils first chip as part of Broadcom deal," 24 June 2026.
[15]	Winbuzz, "OpenAI Broadcom Deal Hits Financing Snag in AI Chip Buildout," 8 May 2026.
[16]	Remarks by Nikesh Arora, chairman and chief executive of Palo Alto Networks, posted 26 August 2026 and on the company's June 2026 earnings call, as reported by Stocktwits and 24/7 Wall St., August 2026.
[17]	Fortune, "CoreWeave CEO Michael Intrator cites 'sold out' capacity as revenue more than doubles and backlog swells to \$104 billion," 11 August 2026.
[18]	Michael Intrator, interview with CNBC, April 2026, as reported by Stocktwits.
[19]	DatacenterDynamics, "Microsoft's neocloud spending surpasses \$60bn," 2026, including remarks by Satya Nadella.
[20]	Remarks by Jensen Huang at the G20 Summit, as reported September 2026.