The Oil of the AI Age? Compute Futures and the Neocloud Economy

Before we begin: the terms used in this report

A handful of technical terms recur throughout the piece. Each is also explained where it first appears; for anyone who wants them all in one place:

Compute — The processing capacity AI needs in order to run. Think of it like a factory's electricity: without it, nothing can be produced.

GPU — The specialised chip that does AI work. NVIDIA models such as the H100 and B200 are the subject of this piece.

GPU-hour — The price of using one GPU for one hour. This market's "kilogram," its "barrel," its "kilowatt-hour."

Neocloud — A company whose only business is renting out GPUs. CoreWeave and Nebius are this piece's two examples. Like a hotel: they rent GPUs instead of rooms.

Hyperscaler — The giant cloud companies — Amazon (AWS), Microsoft (Azure), Google. Renting out GPUs is only one of the things they do.

Spot price — Today's price, right now.

Index — A single average number produced from scattered prices. Like the CPI: the prices of thousands of products reduced to one figure.

Futures — An agreement to "fix today the price six months from now." Bought and sold on an exchange.

Hedge — Taking out insurance against price risk. Done by buying or selling a futures contract.

Forward curve — The full set of prices formed for the same good 3 months, 6 months, 1 year and 3 years out. A map of the market's expectations about the future.

Backwardation — The forward curve sloping downwards: the future is priced cheaper than today.

Contango — The opposite: the future is priced more expensively than today.

Basis risk — The risk that the price your insurance is based on differs from the price you actually face.

Take-or-pay — A contract of the "you pay whether you use it or not" type. Like a gym membership.

Depreciation — Expensing the value of a machine by spreading it across the years. The accounting answer to the question "how many years does this GPU take to lose its value?"

EBITDA — Profit before interest, tax and depreciation are deducted. Shows how much cash a company's core operations generate.

Leverage — A small change producing a large outcome. *Operating leverage*: with costs fixed, profit erodes fast when revenue falls. *Financial leverage*: debt makes that erosion hit the shareholder in compounded form.

Backlog — The total value of work a company has signed but not yet delivered.

Scarcity premium — The part of the price that cost cannot explain. The excess paid purely because "there is little of the good and many who want it."

Basis point — One hundredth of a percentage point. 100 basis points = 1%. Interest rate differentials are discussed in this unit.

CapEx — Capital expenditure: investment in lasting assets such as factories, machines and GPUs.

Offtake backstop — A commitment by the producer that it will itself buy any goods that cannot be sold. It is what allows a small company to raise credit.

Circular financing — A company giving money to the customers who will buy its own product. It raises the question of how much of the demand is real.

MW / GW — Megawatt / gigawatt. The electrical power a data centre draws. 1 GW = 1,000 MW ≈ the consumption of a mid-sized city.

1. Why is compute now a "commodity"?

On 11 August 2026, CME Group — the world's largest derivatives exchange — put out a product announcement. It contained two new futures contracts: **Silicon Data H100 Rental Index Futures** and **Silicon Data B200 Rental Index Futures**. They begin trading on 5 October 2026, on NYMEX. Each contract represents one month of GPU rental cost.

Two details in that announcement say more than the announcement itself.

The first is where the product was filed. CME did not shelve these contracts next to equity indices or rates products; it put them **under the energy desk**. The executive who made the announcement was Pete Keavey, CME's Global Head of Energy and Environmental Products — meaning the business unit that owns the product is the energy desk. In other words, CME has institutionally classified computing power as something that belongs on the same shelf as power and natural gas. Keavey's own sentence confirms it: "Compute has become the currency of the AI era, and this market will bring transparency to the current and future costs that AI developers and hyperscalers need to hedge."

The second is what the contract is written on. Nobody expects a crate of chips to arrive at their door at expiry. What is bought and sold is not the chip itself but **one hour of the chip's rental.** At expiry the two sides settle the price difference in cash.

That distinction is bigger than it looks. In oil the unit of measure is the barrel; in power it is the megawatt-hour. In artificial intelligence it is the **GPU-hour**: the cost of running one graphics processor for one hour.

> **What exactly is a futures contract?**
>
> Picture a hazelnut farmer. The crop comes in during September, but he has no idea what the price will be then. A chocolate factory faces the same uncertainty — just from the other direction: if the price spikes, its input costs blow out.
>
> If the two agree today, both can relax: "$2 a pound in September." The farmer is protected if the price falls, the factory if it rises. When that agreement is standardised and made tradable on an exchange, it is called a **futures contract**.
>
> That is exactly what starts on 5 October — except for GPU-hours instead of hazelnuts.

This report starts from that piece of news, but the news is not its subject. The real question is this: when a futures market is created for a good, what happens to the economics of the companies that produce and sell it? We know how that transition played out in oil, natural gas, power and agricultural commodities. We do not know how it will play out in compute — but there is enough data to make an informed guess.

How compute became a scarce factor of production

Training and running AI models depends on a resource as fundamental to economic value as electricity: parallel computing capacity. Training a language model requires tens of thousands of GPUs running uninterrupted for weeks; inference, unlike training, is a *continuous* load that grows linearly with the number of users.

Between 2023 and 2026 the supply side of that capacity went through three generations. When the H100 (Hopper, 2023) launched, it simply could not be sourced. The H200 arrived as a higher-memory version of the same architecture and found demand in particular for inference on large open-weight models. In 2025–2026 the Blackwell generation — the B200 and its "Ultra" variant, the B300 — came online; according to TrendForce, **more than 70% of NVIDIA's high-end GPU shipments** in 2026 will be Blackwell, while the next generation, Rubin, carries slippage risk.

Every generational transition does two things: it pushes the new chip's price up and the old chip's rental price down. That is a dynamic with no equivalent in classical commodities — and, as later sections will show, the fundamental reason compute is *not* quite oil.

Renting instead of buying

There are two ways to get access to GPU capacity. The first is to buy the hardware: an H100 runs $25,000–40,000 a unit, and an 8-GPU B300 system sits in the $300,000–350,000 range. On top of that come the data centre, power, cooling and networking. The second is to rent: to buy GPU-hours at an hourly rate.

Almost every AI company chose the second route, because the first is both capital-intensive and puts technological obsolescence risk straight onto the balance sheet. That preference created a new kind of company sitting in between: the **neocloud**. CoreWeave, Nebius, Lambda and Crusoe do exactly one thing — commit billions of dollars of GPU capital expenditure and rent that capacity out by the hour. Unlike the hyperscalers (AWS, Azure, Google Cloud), they do not sell a broad portfolio of software services; what they sell is essentially a single line item: the GPU-hour.

That makes them companies whose business model is tied to the price of a single commodity. Exactly like an independent refiner or a natural gas producer.

An analogy may help: think of neoclouds as **hotels**. They spend billions building the property, then rent the rooms out by the night. The only difference is that the rooms are GPUs and they are sold by the hour rather than

the night. Hyperscalers, by contrast, are vast complexes that also run a restaurant, a gym and a conference centre alongside the hotel — room revenue is only part of the take.

And that price is anything but stable

The most striking feature of the rental market is that, for a long time, *there was no single price at all*.

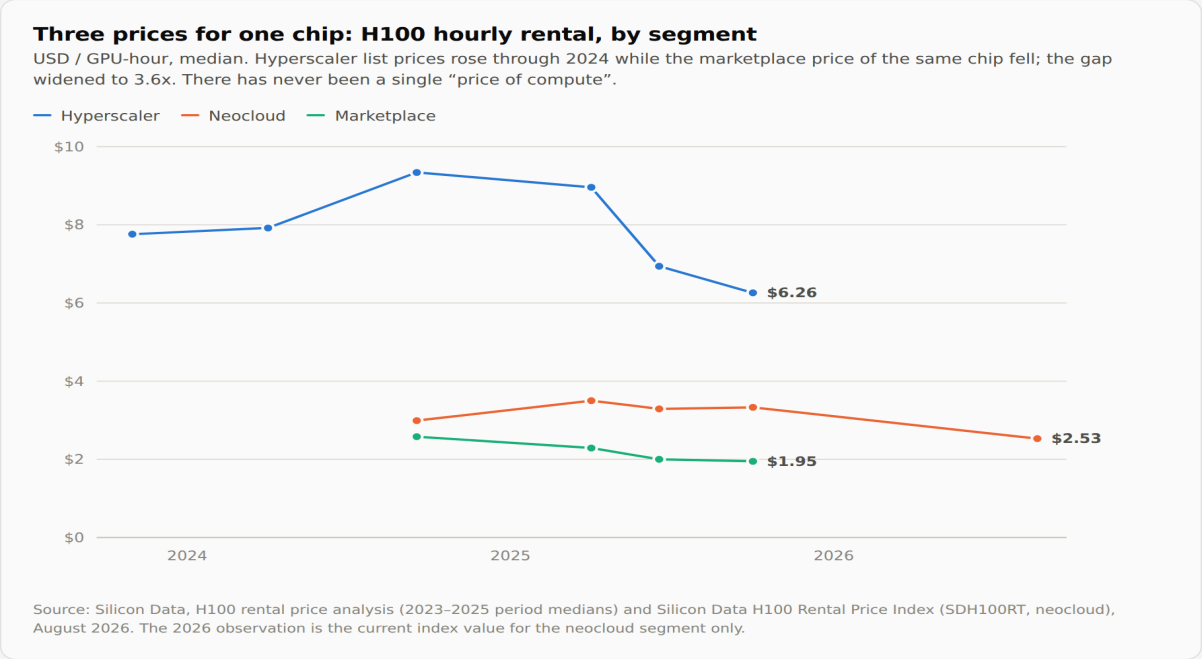

Chart 1 — H100 hourly rental price by segment

Silicon Data's segment-level data shows the same chip trading at diametrically opposed prices at the same moment. In the second half of 2024 the hyperscaler median peaked at $9.34 per GPU-hour while the same H100 was renting on marketplaces for $2.58 — a 3.6x spread. Hyperscaler list prices behaved stickily and drifted higher while the competitive segment fell.

The time series itself is at least as instructive. Silicon Data's H100 index (SDH100RT) stood at $3.06 in September 2024 and fell to $2.36 by June 2025 — a 23% correction in nine months. The 44% cut AWS made to its H100 pricing in June 2025 repriced the entire market. Then direction reversed: SemiAnalysis's one-year rental index bottomed at $1.70 in October 2025 before rising to $2.35 by March 2026 — **a 38% move off the low**, with 15–20 points of it happening between late January and February alone. Silicon Data's H100 index was $2.72 per GPU-hour on 19 July 2026; in August it is approaching $3.

From the March 2023 peak ($4.70, with spot trades above $8) to the marketplace level at the end of 2025 ($1.95), that is a fall of roughly 58%, followed by a 38% recovery in five months. A commodities analyst shown that price series with the source label covered up might have taken it for natural gas or shipping freight.

And that was precisely the problem. Neither buyer nor seller had any instrument for managing the volatility. A neocloud was entering into billions of dollars of capital commitments without knowing what price it would rent out, in 2028, the GPUs it would take delivery of in 2026. An AI company was building next year's compute budget on an assumption. In the words of Silicon Data CEO Carmen Li: "Compute futures give the market something it has never had: a public, tradable reference price for the resource every AI system runs on."

The question this report asks

From here on, the question this study tries to answer is:

> **Can compute futures change the economic structure of AI infrastructure — and, in particular, the risk/return profile of the neoclouds?**

In looking for an answer, we will follow this chain:

AI demand → GPU demand → compute supply/demand balance → GPU rental prices → the compute forward curve → neocloud revenue and margins → valuation → the effect on the AI ecosystem.

The last link is there deliberately. The first seven answer a financial question: what is a company's stock worth? But the price of compute is not a revenue line for a handful of firms — it is the input cost of everything built with AI. The final link asks what it means for the whole sector that this cost is about to acquire a market price.

Every link in the chain is measurable. From October 2026, for the first time, the last three will have a market price too.

2. The birth of compute futures

Two exchanges, two indices, one race

To read compute futures as a single product announcement is to miss half the story. What is actually happening is a **benchmark race** — and it began within the space of a week in May 2026.

On 12 May 2026 CME Group said it would launch futures on GPU rental price indices; its partner was Silicon Data. Seven days later, on 19 May, ICE (Intercontinental Exchange — owner of the NYSE) announced it would do the same; its partner was a startup called **Ornn** — which a month later, in June 2026, closed a $33 million seed round led by a16z.

This is no coincidence. In commodity markets the real power sits not with the exchange that lists the contract but with the **reference price** the contract settles against. The rivalry between Brent and WTI was far more a benchmark contest than an exchange contest; the winning reference defined the world's oil price for thirty years. The same thing is happening in compute: whichever index becomes the standard, the "compute curve" used to value AI infrastructure assets will be built on top of it.

	CME Group	ICE
Index provider	Silicon Data	Ornn
Index	Silicon Data Rental Index (e.g. SDH100RT)	Ornn Compute Price Index (OCPI)
GPUs covered	H100, B200 (A100 available in the index)	H100, H200, B200, RTX 5090
Data basis	Prices observed across 50–100 platforms	Executed trades only ("printed transactions")
Settlement	Cash, USD	Cash, USD
Contract unit	One month of rental	Arithmetic average of the daily index over the contract month (Asian-style)
Listing	NYMEX	ICE
Launch	5 October 2026	Subject to regulatory approval

The single most important technical difference between the two announcements is one line of that table: **what the index is built from.**

Silicon Data collects roughly 150,000 verified price records a day from 40–50 countries and 50–100 platforms; it tracks on-demand, spot and reserved contracts of 1–60 months separately, normalises for machine specification, rental term, geography and performance, discards statistical outliers, and publishes a single daily "like-for-like" hourly rate. It maintains separate indices for hyperscalers and neoclouds — which is precisely why the spread we saw in the Section 1 chart can be measured at all.

> **Why is an index needed at all?**
>
> Think about the CPI basket. There is no single "price" in an economy; it differs by store and by city. The statistics agency collects prices for thousands of goods, weights them, and produces one number. Nobody actually shops at exactly that price, but everyone talks in terms of that number — wage settlements, rent increases and interest rate decisions all hang off it.
>
> GPU rental is the same. The same chip rents for $2 on one platform and $11 on another on the same day. Before a futures contract could be anchored to anything, that dispersion had to be reduced to a single number. That is the job Silicon Data and Ornn do.

Ornn makes a different claim: OCPI is "the first compute index built exclusively from executed transactions." Prices coming off a platform used by more than 400 data centre operators, investors and AI companies are published on the Bloomberg Terminal.

The distinction looks like a technicality, but it is the oldest fight in derivatives markets. An observed price (a quote) gives broad coverage but may be an offer that never traded; an executed trade (a print) is more resistant to manipulation, but when volume is thin it stops being representative. The collapse of LIBOR was exactly a failure of the first type; "thin print" manipulation in commodity markets is a failure of the second. Which of the compute indices proves the more robust will be settled by the first big wave of volatility.

How exactly will the contracts work?

The two contracts CME will list on 5 October are:

- **Silicon Data H100 Rental Index Futures**
- **Silicon Data B200 Rental Index Futures**

Both will trade on NYMEX, denominated in dollars. There is no physical delivery; at expiry the two sides simply settle the price difference.

The calculation method embeds an interesting design choice. The contract settles not against the price on the **last day** of the month but against the **average of every day** through the month. The reason is simple: compute cannot be stored. Like electricity, it is consumed the moment it is produced. For a company renting GPUs every day for a month, the true cost is not the price on the final day but the average of thirty days. The contract is built to match.

So how big is one contract? CME says "one month of rental"; exact figures have not yet been published. For a rough sense: a month is about 730 hours and a GPU-hour is $2.72. So a single contract represents a notional of **around $2,000**.

For comparison, a crude oil contract is 1,000 barrels — $91,000 at today's price. So the compute contract is about one-fortieth the size of the oil one. That looks like a deliberate choice: large companies can buy hundreds of contracts, and a small investor can buy them one at a time.

One caveat: as of July 2026, **no GPU futures venture had received final CFTC approval**; every announced contract carried "pending regulatory review" status. CME's 5 October date comes with that qualification.

Who will hedge?

A futures market survives only if there are natural position-holders on both sides. In compute, they are clear:

Natural buyers (long hedgers) — protecting against price increases. The AI companies renting GPUs. A company that knows it will consume 2 million H100-hours next year buys futures to protect its budget: if the spot price rises, the futures gain covers the cost. Structurally, this is identical to an airline hedging jet fuel.

Natural sellers (short hedgers) — protecting against price declines. Neoclouds, data centre operators, GPU fleet owners. They hold a capital-intensive asset with an economic life well short of ten years, and that asset's return depends entirely on future rental prices. By selling futures they can convert "an eroding rental stream into a fixed one."

The critical point here is that the two groups are not symmetric. **Operators that finance their fleets with debt will probably be the market's most persistent sell side** — because for them hedging may become not a choice but a lender's requirement. That is the subject of Section 5.

This two-sided structure is the sine qua non of a commodity market. Airlines do exactly this with jet fuel: because they publish ticket prices months ahead, a rise in oil hurts them, so they insure themselves with futures. On the other side stands the oil producer, who fears a fall in price.

The third group is speculators: macro funds, equity traders with indirect exposure to AI capex, participants hunting relative value between GPU generations (the H100 vs B200 spread), calendar spread traders, and algorithms arbitraging between the exchange and spot. In commodity markets this is usually the group that carries liquidity; hedgers only set the net position.

The oil analogy — and the first crack

The parallel with classical commodity futures is almost one-for-one:

	Oil	Compute
Physical unit	Barrel	GPU-hour
Spot market	Physical cargoes, regional differentials	Rental platforms, regional differentials
Reference price	Brent / WTI	SDH100RT / OCPI
Derivative	Futures curve	Futures curve
Use	Hedging, price discovery, financing	Hedging, price discovery, financing

CME's decision to put the product on the energy desk is the institutional admission of that parallel. But it is precisely here that the analogy shows its first crack.

In oil, a barrel is a barrel. There are known, standard discounts for API gravity and sulphur content, and they hold steady for decades. In compute, an H100-hour is not always an H100-hour. In the words of Dave Friedman's note on basis risk, **"an H100-hour is not the same economic good as a B200-hour — nor is the Rubin-class capacity that will swap in eighteen months from now."** Region, interconnect quality (InfiniBand or Ethernet), uptime guarantees, storage, and above all *cluster scale* — an 8-GPU node and a 512-GPU interconnected cluster are economically different goods. A startup looking for a 512-GPU co-located cluster cannot hedge effectively with a broad index.

This is called **basis risk**, and in compute it is structurally larger than in classical commodities. A futures contract transfers the risk of the index price; not the risk of the price at which a given company sells its *own* GPU to its own customer in its own region.

The second crack runs deeper: oil does not become technologically obsolete in 18 months. We will take up what that difference means in detail in Section 9, but there is an early indicator already. According to Ornn's data, the Blackwell spot rental price jumped from $2.75 to $4.08 — **48%** — between mid-February and mid-April 2026. Over the same period the H100 index sat around $2.5. So compute is not a single commodity; it is a *family* of commodities, each with its own supply curve and its own obsolescence calendar, and each an imperfect substitute for the others. The oil equivalent would be discovering a new, more efficient type of hydrocarbon every eighteen months.

Why now?

One last question: why is this market being built in 2026 rather than 2023?

Three conditions all had to be met. First, **a spot market you can price**: in 2023, renting an H100 was a negotiation, not a price. Second, **scale**: in the words of Ornn CEO Kush Bavaria, compute has "grown into a trillion-dollar market, but still lacks the pricing infrastructure that other major commodities rest on." Big Tech's 2026 capital expenditure estimates were running around $650 billion at the start of the year; with guidance revised upward, the range under discussion today is $690–770 billion. Third, and most important, **volatility**: without the roughly 58% two-year decline and the subsequent five-month 38% recovery we saw in Section 1, nobody would have needed a hedging product.

Commodity futures markets have always been born in the same order: first scarcity, then volatility, then hedging demand, and finally the contract. Compute reaches the last step of that sequence in October 2026.

3. The neocloud business model: where exactly does the money come from?

To understand what compute futures will change, you first have to see how these companies make money — and where they lose it. Because in this business model, profit does not sit where it appears to sit.

The equation

On paper, a neocloud's revenue is the product of four variables:

> **Active power (MW) × Utilization × GPU-hour price × Time = Revenue**

Behind that sits a capital side as well:

> **GPU capex → Capacity → Revenue → Gross margin → EBITDA → Interest and depreciation → Bottom line**

It looks simple. But the Q2 2026 numbers show that every term in that equation is tied to a different risk.

Two companies, same product, different scale

	CoreWeave (CRWV)	Nebius (NBIS)
Q2 2026 revenue	$2.60bn (+112% YoY)	$582m (+454% YoY)
Annual revenue run-rate — (last quarter annualised)	~$10.4bn at quarter-end; 2026 exit target $18.5–19.5bn	$3.0bn at quarter-end; 2026 exit target $7–9bn
Gross margin	66% (down 8 points)	77% (up 6 points)

	CoreWeave (CRWV)	Nebius (NBIS)
Adj. EBITDA — (earnings before interest, tax and depreciation)	$1.54bn (59%)	$236m (41%)
Adj. operating income	$128m (5%)	—
GAAP bottom line	−$626m	—
Q2 capital expenditure	$9.4bn (2026: $35–39bn)	~$5.7bn (2026: $20–25bn)
Active power	1.5 GW (51 data centres)	Year-end target 0.8–1.0 GW
Contracted power	4.2 GW (11 August)	5.0 GW target (was 3.5 GW in March)
Backlog	$129bn	Microsoft ~$17bn; up to $27bn with Meta
Net interest expense (quarterly)	$640m	—
Recourse debt — (debt for which the company is liable with all its assets)	$31.4bn (debt/equity 7.39)	—
GPU depreciation period	**6 years**	**5 years** (from 2026; previously 4 years)

At first glance the two do the same thing: raise capital, buy GPUs, connect power, sell by the hour. But the last row of that table contains the distinction the rest of this report turns on.

What actually determines revenue?

The answer is more interesting than the textbook version. There are eight drivers, and their weights differ enormously.

1. Power you can energise — the binding constraint. In 2026, what limits neocloud growth is not GPU supply but electricity. Against CoreWeave's 4.2 GW of contracted power, its active power is 1.5 GW; that is, revenue-generating capacity is one-third of what has been committed. Nebius management says it outright: *"However fast we bring capacity online, we can sell it."* On this picture the driver of near-term revenue growth is not price but **energisation speed**.

2. Contract structure — where revenue is severed from price. The **overwhelming majority of CoreWeave's revenue comes from committed contracts**; the company does not disclose an exact percentage, but per its 2025 annual report the weighted average term of those contracts is **roughly five years**. They are largely take-or-pay: the customer pays whether or not it uses the capacity. This is a crucial distinction — **CoreWeave's revenue this year is almost entirely independent of the spot GPU price.** Spot only hits revenue when contracts come up for renewal. Which means the risk is concentrated not today but in 2028–2030.

> **What does "take-or-pay" mean?**
>
> It is like a gym membership. If you signed up for the year, you pay the dues even if you never set foot in the place. From the gym's point of view this is a magnificent business model: its revenue is independent of whether members actually exercise.
>
> Most of CoreWeave's contracts work that way. The customer pays for five years whether or not it uses the GPUs. That is why, whatever today's GPU price does, CoreWeave's revenue this year does not move.
>
> But bear this in mind: every subscription has an end date. And when it renews, it renews **at that day's price.**

3. The term premium. Nebius, with a transparency rare in the sector, disclosed its pricing: medium-term deals of 1–3 years generate $20–25 million per MW; **short-dated capacity of up to 6 months goes for $40–50 million per MW** — twice as much. The reason is simple: a customer that wants capacity today rather than waiting eighteen months is paying for urgency. It is the same logic that makes a last-minute plane ticket expensive. Commodity markets even have a name for it: *convenience yield*, the premium on having the thing to hand.

4. Price itself — and right now it is going up. A capacity auction Nebius ran in the second quarter cleared at "the highest price we have seen for the Blackwell generation, **15% above** the highest price we had previously charged." It is the same story as the H100 index's 38% recovery off the low in Section 1.

5. Utilization. On long-term contracted capacity the risk of sitting idle passes to the customer — they are paying anyway. But on capacity sold by the hour, every idle GPU-hour is revenue that never comes back. Like an empty seat on a departing flight.

6. GPU generation mix. Put B200s in the same MW and you generate more revenue than with H100s. But the old generation does not die either: CoreWeave signed a contract running **out to 2029** for A100s launched in 2020, and according to CEO Mike Intrator, H100s coming off their initial contracts were re-rented at **95% of the original price.** The reason is interesting: older data centres were designed for about 20 kW per rack; Blackwell's 120–140 kW requires building from scratch. So renting out an old GPU beats leaving the facility empty.

7. Customer concentration. CoreWeave's largest relationship is Meta: a $21 billion agreement running to 2032, on top of a prior $14 billion commitment. At Nebius, Microsoft is ~$17 billion. These contracts guarantee revenue but also tie it to a single counterparty.

8. Cost of capital — and far more decisive than revenue. CoreWeave's quarterly net interest expense is $640 million; annualised, ~$2.56 billion. Adjusted operating income in the same quarter was $128 million. So **interest expense is five times adjusted operating income.** This is not an operating company; it behaves like a leveraged infrastructure fund.

Where does the money go?

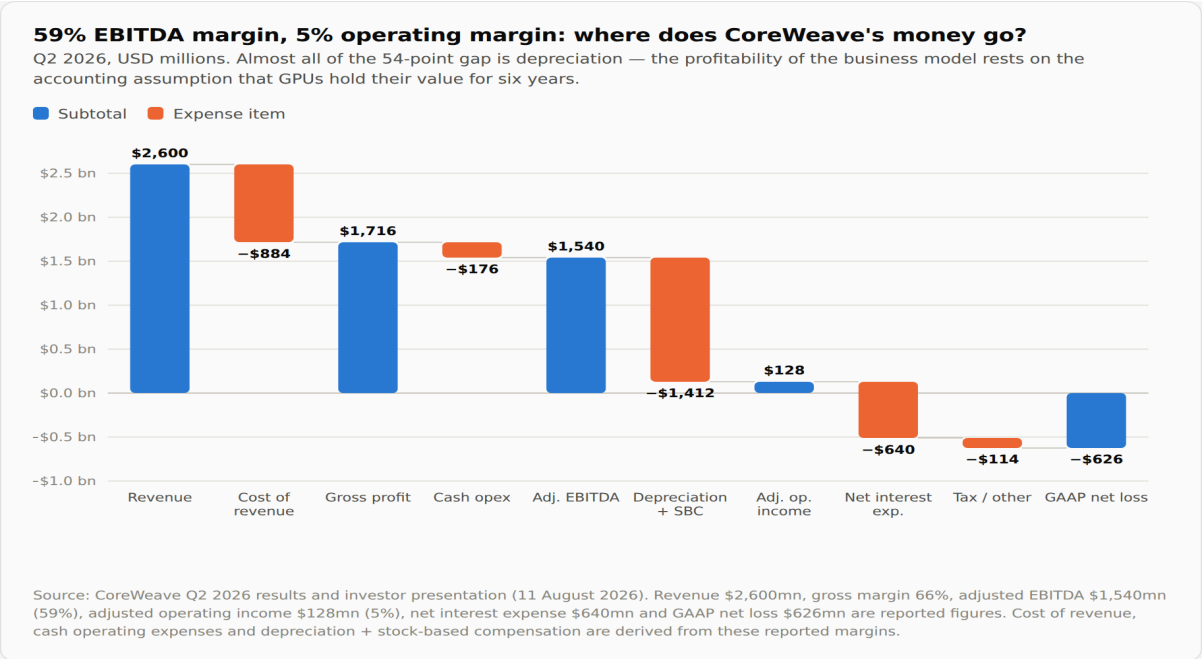

Chart 2 — CoreWeave margin ladder, Q2 2026

This chart may be the single most important visual in the report. CoreWeave's adjusted EBITDA margin is 59% — a figure any software company would envy. In the same quarter its adjusted operating margin is 5%. Almost the entirety of those 54 points is **depreciation**.

An aside: who created this sector?

So far we have treated the neoclouds as data. But there is a question worth stepping back to ask: **why do these companies exist?**

The answer is not, as commonly assumed, "entrepreneurs spotted a gap." CoreWeave was a small crypto mining outfit in 2017. Nebius came out of the break-up of the Russian search engine Yandex. Neither walked naturally into the centre of AI infrastructure.

NVIDIA carried them there.

In 2023, at the peak of the GPU shortage, NVIDIA had a simple option in front of it: sell every chip it made straight to Amazon, Microsoft and Google and be done with it. Demand already exceeded supply many times over; selling required no effort.

NVIDIA did not do that. Instead it used the scarcity to **build a new customer layer from scratch.**

The method had three legs:

1. Allocation priority. New-generation chips went to the neoclouds first. CoreWeave and Nebius were "among the first to take delivery" of new platforms like Blackwell Ultra and Rubin. That gave small companies the one thing that draws large customers: hardware nobody else had.

2. Capital. NVIDIA became a direct shareholder — roughly a $900 million stake in CoreWeave plus a $2 billion investment on top; in Nebius, first a $33 million stake, then another $2 billion.

3. And most critically: an offtake backstop. NVIDIA signed an agreement with CoreWeave **running to April 2032 with an initial value of $6.3 billion**: if CoreWeave cannot find a buyer for capacity it fails to sell, NVIDIA will buy that capacity.

That third item is not a detail; it is the keystone. When you walk into a bank and say "I am going to buy billions of dollars of GPUs, lend me the money," the first question is: *what happens if you can't sell it?* NVIDIA's guarantee removes the question. So NVIDIA did not merely sell chips — **it stood as guarantor so that small companies could raise credit.**

In July 2026 that arrangement became a formal programme. NVIDIA's cloud partnership programme gives new cloud providers a capacity guarantee; in exchange it takes either share options or a revenue share (that share steps down over the life of the contract). If the GPUs sit idle, NVIDIA either finances the rent or buys the capacity back. Among the programme's disclosed participants are Sharon AI, which is planning 40,000 GB300 chips, and Firmus, which is building a 360 megawatt facility in Indonesia.

The hyperscalers financed their own competitors

The real artistry here is in the next step.

As NVIDIA built out this new layer, Amazon, Microsoft and Google — struggling to source chips themselves — **were left with no choice but to sign enormous contracts with these small companies.** Microsoft committed roughly $17 billion to Nebius. Meta signed a $21 billion contract with CoreWeave running to 2032 — on top of the earlier $14 billion.

Consider the result. Companies that did not exist four years ago now own a vast infrastructure paid for with the money of the largest technology companies in the world. And that infrastructure will, in time, compete with exactly those companies.

In other words, the hyperscalers paid the founding capital of their own future competitors.

From NVIDIA's point of view the arithmetic is flawless: the customer base widened, dependence on a single group of buyers fell, and its customers were turned into each other's customers. A textbook case of a monopoly using its monopoly power to strengthen its own negotiating table.

> **Two readings of the same event**
>
> Critics of this strategy give the same facts a different name: **circular financing.** NVIDIA gives money to companies that will buy its chips; those companies use that money to buy NVIDIA's chips; NVIDIA books it as revenue. The question of how much of the demand is real and how much is a self-feeding loop is a legitimate one.
>
> Both readings describe the same phenomenon. The difference lies in whether a real customer stands at the end of the loop. So far one has: Meta, Microsoft and OpenAI pay real money. But that does not mean the question is closed — only that it has not yet been tested.

Keep this story in mind. Later in the report, when we discuss why the futures market may work this time, we will come back to precisely this point.

Then comes $640 million of interest expense, which wipes out the $128 million of operating income and turns it into a $626 million GAAP net loss.

The conclusion that follows is this: **the neocloud business model rests on an accounting assumption about how quickly you write the book value of GPUs down to zero.** If the assumption is close to reality, the 5% operating margin is real. If it is not, that margin never existed.

> **Why does depreciation matter this much?**
>
> Say you borrow and buy a $200,000 truck and lease it out to a corporate fleet. You collect $60,000 a year in rent, and fuel and maintenance run $10,000. Is your profit $50,000?
>
> No. The truck loses value every year. If it is scrap in ten years, then $20,000 a year is melting away and your real profit is $30,000. But if it is scrap in five years, $40,000 a year is melting away and your profit drops to $10,000.
>
> **Your assumption about how many years the vehicle lasts can double your profit or halve it.** With nothing whatsoever changing in the business itself.
>
> CoreWeave says "six years" for its GPUs. Nebius says "five years" for the same chips. Critics like Michael Burry say "far less." That is exactly the size of the argument.

Six years or five?

This is precisely why CoreWeave depreciating its GPUs over **6 years** and Nebius depreciating the same GPUs over **5** is not a technical accounting detail but two different claims about the business model. The difference is more instructive still: Nebius used 4 years until 2026 and raised it to five this year, citing "usage patterns and existing occupancy commitments" — meaning the assumption is not fixed but a moving target.

Choosing six years has two effects. First, annual depreciation expense falls and operating income rises. Second — and more important on the credit side — **the book value of the pledged GPUs stays higher for longer.** CoreWeave uses its GPU inventory as debt collateral; slower depreciation means higher collateral value.

The counterargument is strong: if the overwhelming majority of revenue comes from take-or-pay contracts averaging five years, then the customer is effectively bearing the depreciation and interest cost. The A100s contracted out to 2029 and the H100s renewing at 95% of price support that thesis too.

The counter-counterargument comes from Michael Burry: he argues that by extending GPU useful life assumptions to five or six years, the hyperscalers understated depreciation by **$176 billion** over the 2026–2028 period — while NVIDIA ships a new architecture every year. One Google architect, meanwhile, has estimated the service life of a data centre GPU at one to three years.

It is no accident that this argument remains unresolved. What is needed to resolve it is a **market price** for the future level of the GPU-hour — and until now no such thing existed. In October 2026 it will.

What does the market say?

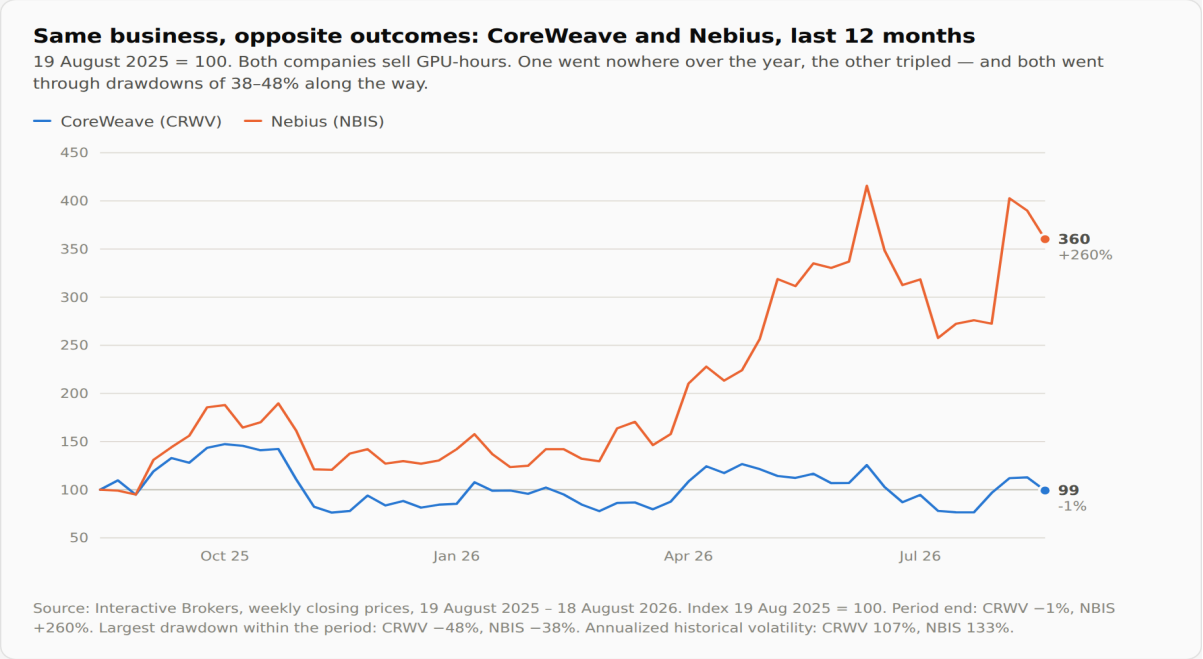

Chart 3 — CoreWeave and Nebius, indexed 12-month performance

Over the last twelve months CoreWeave's stock has gone nowhere (−1%); Nebius is up 260%. Same sector, same product, same customer base. But both took hard drawdowns along the way: CoreWeave 48% from its peak, Nebius 38%.

The volatility figures are more striking still: CRWV's annualised historical volatility is **107%**, NBIS's **133%**. For comparison, the S&P 500's long-run average sits in the 15–20% band; even in crude oil's most turbulent years it is 60–80%. These stocks trade less like an infrastructure company than like a long-dated commodity option.

Let us make those volatility numbers concrete: 107% annual volatility means that losing half its value or doubling within a year counts as **ordinary** for the stock. These are priced not like infrastructure companies but like bets.

The main reason for the divergence is the balance sheet. CoreWeave's credit default swap spreads had widened to roughly **850 basis points** ahead of earnings; on a standard recovery assumption, that implies a five-year default probability of around 50% (not a figure published by market data providers, but a derived reading). Nebius, by contrast, collects prepayments on 70% of its contracts, and those prepayments cover 50–60% of the associated capital expenditure — meaning the customer supplies half the financing. According to Nebius management, the payback period on second-quarter contracts' capital and operating costs has fallen to **22 months**; the prior range was two to three years.

The difference between the two business models comes down to this: CoreWeave grows on debt, Nebius on customer prepayments. Both are tied to the price of the same commodity, but when that price falls, the first has an interest calendar it has to meet.

The impossibility of measurement

One final caveat — and it connects directly to the report's central thesis.

The "revenue per MW" figures above are **not comparable** across companies. Nebius discloses $20–25 million per MW for medium-term contracts. If you try to derive the same metric from CoreWeave's data (dividing its annualised $10.4 billion of revenue by 1.5 GW of active power) you get roughly **$6.9 million** per MW. The three-to-four-fold gap is not a genuine performance difference; it is a difference in definition. Is "MW" IT load or facility power? Is "revenue" annual, or the total over the life of the contract? Did the capacity come online in stages through the quarter?

An investor who wants to compare the economics of these companies today has no standard unit. Where the oil investor has cost of production per barrel and the power investor has marginal cost per MWh, the compute investor has only the metrics the companies define for themselves.

That is precisely the first problem a reference price — and the futures curve built on top of it — would solve.

4. The real risk: volatility in the price of compute

Let us sharpen the question. The H100 index sits close to $3 today. What happens if it is $1.80 twelve months from now? Or $4?

The finding of this section is this: **the answer lies somewhere other than where it first appears to.** Because the 2026 data does not corroborate the simple thesis that "a new GPU launches and the price of the old GPU falls."

First the thesis: the rental-price squeeze

The classic narrative runs as follows. NVIDIA ships a new architecture every year, the new chip does the same work more cheaply, the hourly price of the old chip falls, and a fleet bought against a six-year depreciation schedule starts generating revenue below its book value.

The thesis had empirical support. As we saw in Section 1, H100 rental fell roughly 58% from its March 2023 peak ($4.70) to the marketplace level at the end of 2025 ($1.95); even within the marketplace segment alone, the decline from mid-2024 to end-2025 was 24%. Silicon Data's own index fell from $3.06 in September 2024 to $2.36 in June 2025 — 23% in nine months.

On the CoreWeave side, the numbers make the thesis more uncomfortable still. The company **bought $10.3bn of equipment in order to generate $5.1bn of revenue in 2025.** Its property, plant and equipment went from $11.9bn to $30.6bn in a single year, and the 2025 depreciation charge rose to $2.45bn — nearly three times the prior year's. In 2023 the company had extended the useful life of its GPUs from five years to six; that decision lifted profit in the year it was made, but 2025 still closed with an operating loss.

The structural mismatch here is clear: **depreciation is tied to a fixed schedule, while rental revenue is tied to a semiconductor roadmap the company does not control.** Every new chip generation resets how much the old fleet can charge.

Then the counter-evidence: prices are rising in 2026

The problem is that the 2026 data shows the opposite.

- H100 rental was just under $2 in January 2026; by August it is approaching $3.
- The B200 index rose **24.4%** in the first quarter of 2026 — and almost all of it came in a single month: from $4.43 to $5.48 over the course of March, or **23.6%**. At the centre of the move was NVIDIA's GTC event on 13–20 March.
- A100 and H200 rental rates are also above where they started the year.
- CoreWeave signed a contract running to 2029 for A100s that shipped in 2020; H100s coming off their first contracts were re-leased at **95% of their original price.**

So where technological obsolescence was supposed to crush prices, six-year-old chips are holding theirs. Why?

The answer is not in the silicon. It is in **electricity.**

Older data centres were built on the assumption that each rack would draw roughly 20 kilowatts. New Blackwell racks want 120–140 kilowatts — six or seven times as much. You cannot put the new chip in the old building; the electrical distribution, the cooling, all of it falls short. In most cases a new chip means a new building.

Under those conditions, renting out an old GPU beats leaving the building empty. Because the bottleneck has shifted from the chip to the power, the older generation is holding its value.

The second reason sits on the demand side: as models become more efficient, the same hardware does more work, which extends the economic life of older chips.

So where is the risk?

The data that reconciles the two theses is in the chart below.

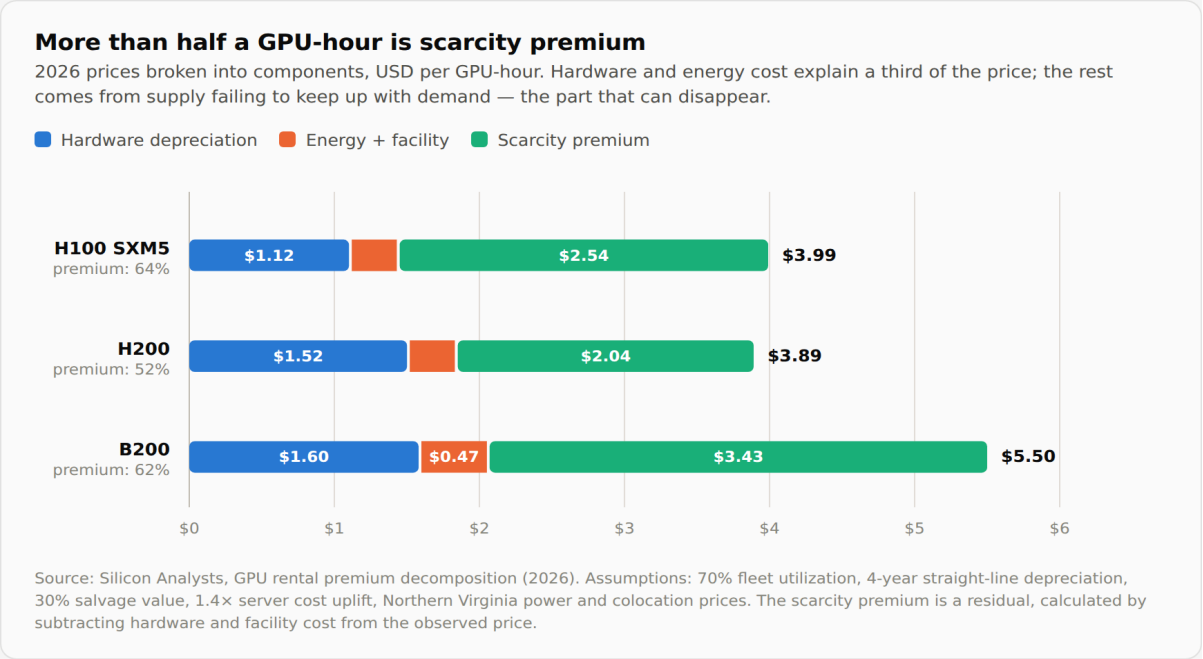

Chart 4 — Anatomy of a GPU-hour price

Break the price of a GPU-hour into its components and the picture that emerges reframes the entire debate. Assuming 70% utilization, four-year straight-line depreciation and a 30% residual value:

	Reference price	Hardware	Energy + facility	Scarcity premium
H100 SXM5	$3.99	$1.12	$0.33	**$2.54 (64%)**
H200	$3.89	$1.52	$0.33	**$2.04 (52%)**
B200	$5.50	$1.60	$0.47	**$3.43 (62%)**

Hardware depreciation accounts for **28–39%** of the price. Electricity and facility costs are — counterintuitively — only **8%** of it (electricity alone is about 1.5%). The remaining 52–64% is a residual that corresponds to no cost line at all: the **scarcity premium.**

> **What is a scarcity premium?**
>
> Think of a concert ticket. Printing the ticket, renting the venue, paying the artist — add it all up and the cost comes to, say, $50. But the ticket is changing hands on the resale market at $400.
>
> The $350 in between corresponds to no cost whatsoever. It is paid purely because **a lot of people want in and there are few tickets.** That is a scarcity premium.
>
> The critical point is this: if the venue doubled in size, that $350 would evaporate and the $50 of cost would remain. And you would not have to do a single thing to cut costs — the premium melts away on its own.
>
> A GPU-hour works the same way. More than half the price is scarcity premium. And that means **the whole of the neoclouds' profit.**

One caveat: this decomposition applies to the reference price points Silicon Analysts uses ($3.99 for the H100 SXM5 — the on-demand level for a particular cohort of providers). At the index level — $2.72 — the same cost structure puts the premium's share at 47%. The conclusion does not change: **whichever price point you look at, half the price or more corresponds to no cost line at all.**

That gives the risk its correct address. The neoclouds' profit margin does not come from the hardware; it comes from **supply failing to keep up with demand.** And that premium is sensitive not to technological obsolescence but to the capacity balance.

The consequence is easy to compute. If the H100's scarcity premium halves — even with hardware and energy costs completely unchanged — the hourly price falls **32%** at the $3.99 reference point and **23%** at the $2.72 index level. If only a third of the premium melts away, the declines are 21% and 16% respectively.

And the neocloud balance sheet cannot absorb it

Now combine that sensitivity with the margin ladder from Section 3.

CoreWeave's second-quarter 2026 revenue was $2,600m and adjusted operating income $128m, implying a cost base of $2,472m. Costs are largely fixed in the short run — depreciation, interest, leased data centre commitments, headcount. On that basis:

Change in revenue	Adjusted operating income
Today	+$128m
−4.9%	**0 (breakeven)**
−10%	−$132m
−20%	−$392m
−30%	−$652m

A revenue decline of roughly 5% wipes out adjusted operating income entirely. And that is the line *before* $640m of quarterly interest expense.

This table should be read with care: the overwhelming majority of CoreWeave's revenue is protected by take-or-pay contracts, so a decline like that will not materialize tomorrow. But the weighted-average contract term is around five years. Which makes this a sensitivity not of today but of the **renewal date.** And every new capacity contract signed today is signed at that day's price.

And then there is a measurement problem

One final layer. Every calculation above proceeds as if there were a single number called "the H100 price." There is not.

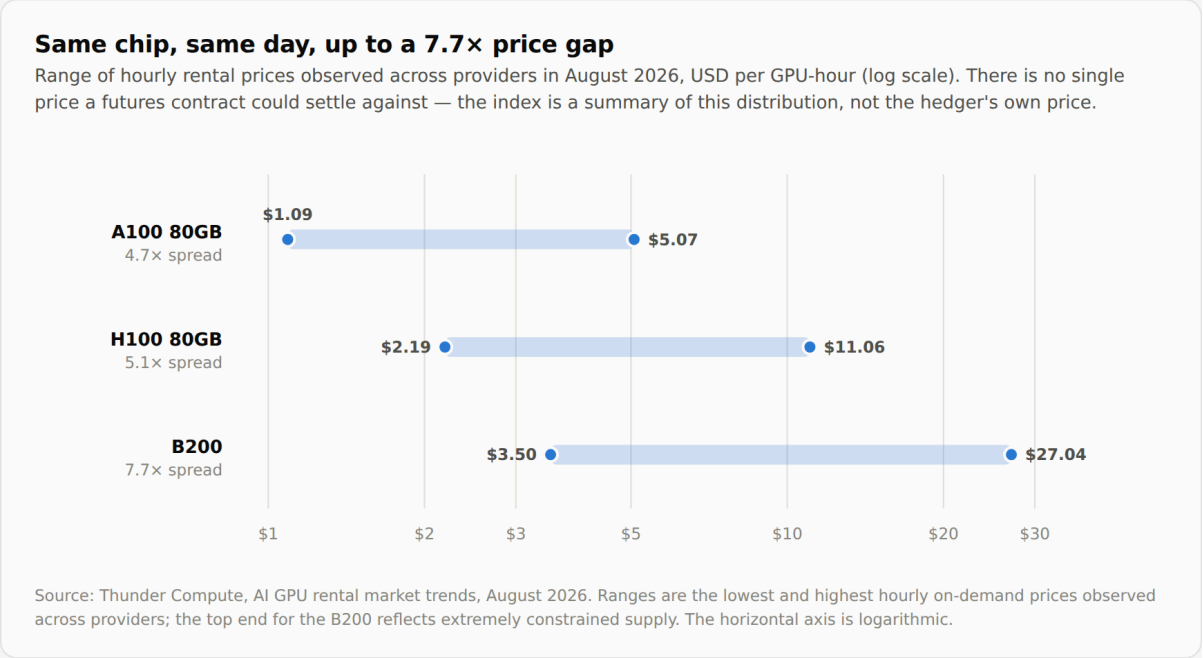

Chart 5 — Same chip, different provider, different price

On-demand price ranges observed across providers in August 2026: $1.09–5.07 for the A100 (4.7×), $2.19–11.06 for the H100 (5.1×), $3.50–27.04 for the B200 (7.7×).

That dispersion says two things.

First, it explains what an index is for. Producing a single reference price in a market this scattered is a genuine service — the normalization work described in Section 2 is precisely this.

Second, and more uncomfortably: **the index is not a price any company actually faces.** A hyperscaler selling at $11 on Azure and an operator selling at $2.19 on a marketplace cannot hedge with the same index. The index can go up while one of them sees its own price go down.

In derivatives markets this is called **basis risk**, and in compute it is extraordinarily large. That is the real subject of the next section: how much of this risk do futures actually transfer?

Section summary

The question of this section was "what happens if GPU prices fall?" The answer the data gives has three layers:

1. **Prices fell hard in 2023–2025, but they are rising in 2026.** The simple "technological obsolescence crushes price" thesis did not work, at least not in this cycle — because the binding constraint is not the chip but the power.
2. **But 52–64% of the price is scarcity premium.** Today's profitability comes not from the value of the hardware but from an imbalance. This is precisely the line item that will melt away as power capacity comes online.
3. **And balance sheets are in no condition to absorb that melt.** A 5% revenue decline zeroes out CoreWeave's operating income; interest expense is already five times operating income.

The depreciation debate — six years, five years, or less still — is really the third point dressed up in accounting clothes. And until now nobody has had a future price with which to settle it.

From October 2026 they will.

5. How could compute futures change the neoclouds?

This section holds the central question of the piece. And the answer lands somewhere slightly different from the most common hypothesis.

The simple hypothesis and why it falls short

The standard narrative is built like this:

> **Today:** buy GPUs → commit billions of dollars of capital → have no idea what price you will rent them at in the future. **After futures:** buy GPUs → sell part of your future rental revenue today → gain revenue visibility.

It sounds reasonable. But the data in Section 3 breaks the frame: **the overwhelming majority of CoreWeave's revenue is already contracted.** It has pre-sold that revenue through take-or-pay agreements with a weighted-average term of around five years. At Nebius, 70% of contracts are prepaid, and the prepayment covers 50–60% of the associated capital expenditure.

In other words, most of these companies had built their own hedges through **bilateral contracts** long before a futures market existed. The classic function of a commodity futures market — letting a producer price future output today — is already present here.

So what exactly do futures solve?

The real address of the risk: where the contract ends

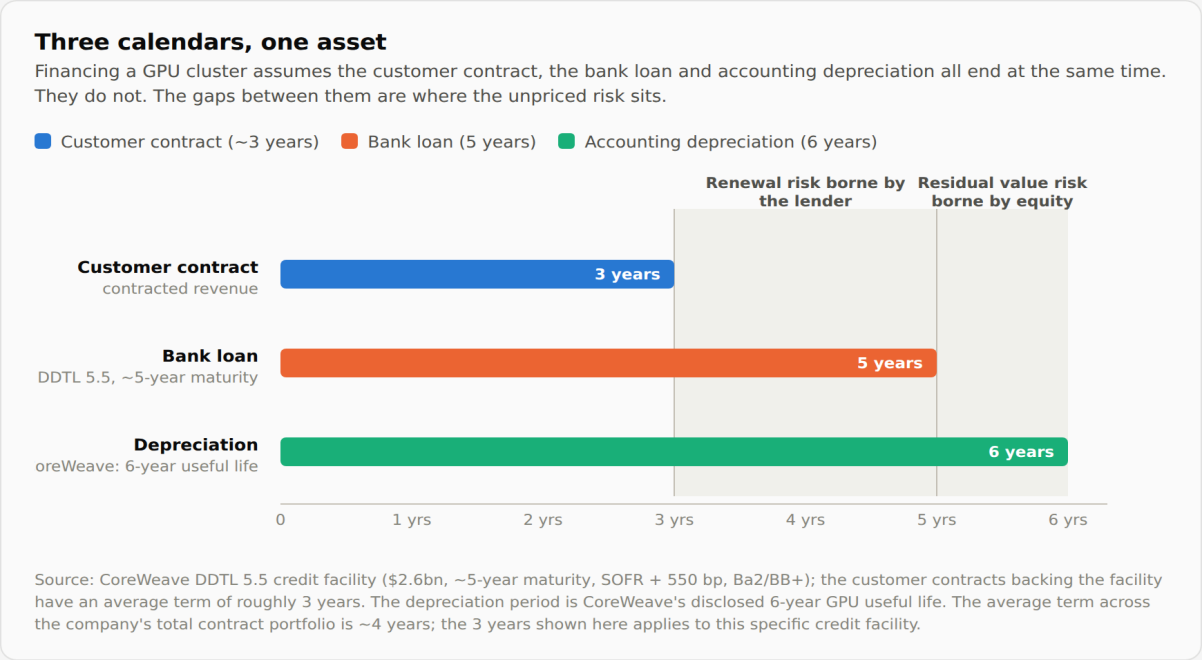

Chart 6 — Three different calendars, one asset

The $2.6bn DDTL 5.5 credit facility CoreWeave closed in June 2026 shows the heart of the problem. The loan runs roughly five years. The customer contracts backing it have an average term of about **three years.** Accounting depreciation, meanwhile, is spread over six.

Three different calendars, one cluster of GPUs. And the gaps in between belong to someone:

- **Years 3–5:** the customer contract has ended, the loan is still outstanding. The capacity has to be re-leased during this period — and at what price is unknown. The party bearing that risk is **the lender.**
- **Years 5–6:** the loan is repaid but the asset is still on the books. The party bearing this residual-value risk is **the shareholder.**

> **Why is maturity mismatch dangerous?**
>
> Imagine you buy an apartment and rent it out. You took a ten-year mortgage from the bank and signed a three-year lease with the tenant.
>
> The first three years are fine: rent comes in, the mortgage gets paid. But in the fourth year the tenant moves out and your mortgage still has seven years to run. What will market rents be on that day? You do not know. Neither does your bank — but getting its money back depends on exactly that.
>
> CoreWeave's situation is this, point for point. A five-year loan, a three-year customer contract, a six-year book life for the GPU. Three different calendars, one asset.

The credit agreement says so explicitly: when the initial customer contract expires, CoreWeave may renew it or lease the capacity to another customer — subject to criteria set out in the loan agreement.

So the financing market is placing a bet on where GPU rental prices will be in 2029. Deliberately. And it is pricing that bet.

The market has already put a price on this risk

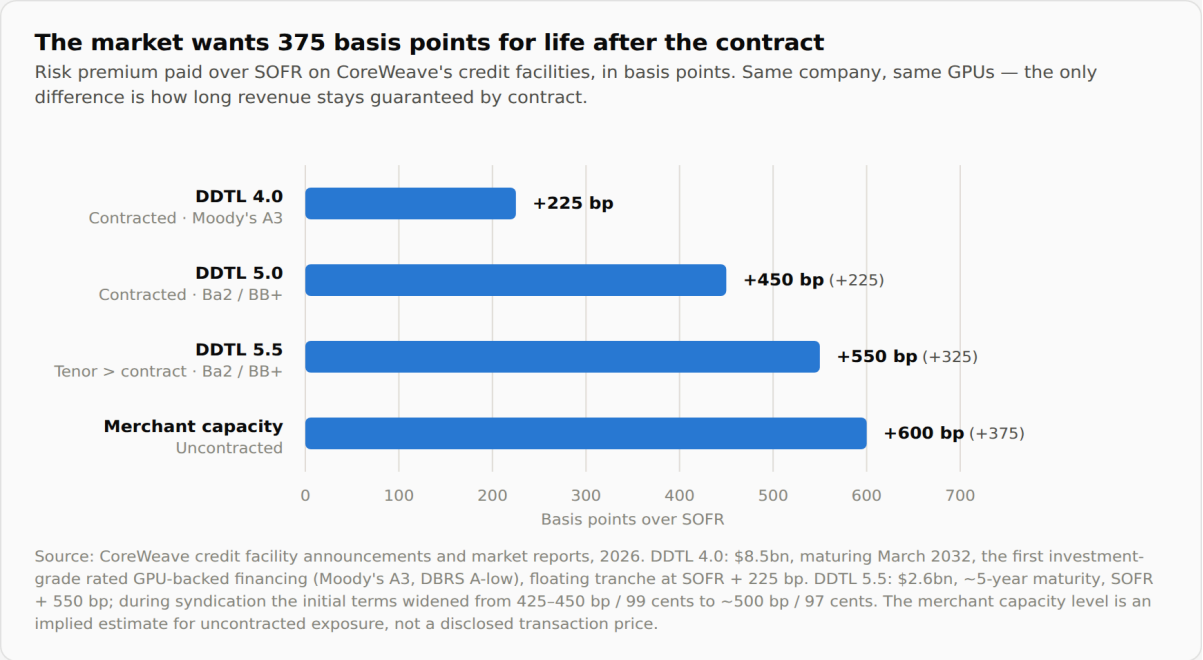

Chart 7 — The credit spread ladder

Look across CoreWeave's various credit facilities and a very clear ladder appears. Same company, same GPUs, same type of collateral — the only thing that changes is how long the revenue is locked under contract:

In the table below, "SOFR" is the reference rate for dollar interest and "bp" is a basis point — one hundredth of one percent. SOFR + 225bp means "2.25% above the reference rate."

Facility	Structure	Spread
DDTL 4.0 ($8.5bn)	Investment-grade customer, fully contracted. Moody's A3	SOFR + **225bp**
DDTL 5.0	Sub-investment-grade customer, contracted. Ba2/BB+	SOFR + **450bp**
DDTL 5.5 ($2.6bn)	Maturity extends beyond the contract term. Ba2/BB+	SOFR + **550bp**
Merchant capacity	Uncontracted exposure (implied)	SOFR + **~600bp**

DDTL 4.0 is the **first investment-grade-rated transaction** in GPU-backed financing — Moody's A3, DBRS A (low), maturing March 2032, with Blackstone Credit & Insurance as anchor investor. When contractual protection is complete, the market treats a GPU as an investment-grade asset.

When that protection is removed, the price moves 375 basis points higher.

What happened while this loan was being marketed to banks is instructive in its own right. CoreWeave initially wanted SOFR + 425–450 basis points; lenders would not take it, and the facility closed at SOFR + 550 basis points, and at a discount to par at that. The effective cost was roughly 10.4%. That **100–125 basis point widening** between ask and close had nothing to do with GPU scarcity — it was about **contract stability and delivery risk.**

The real thesis that follows

What compute futures contribute to neocloud economics is not revenue visibility. Revenue visibility is already in the contracts.

The contribution is making the period after the contract ends measurable.

Today a credit committee cannot look at the question "what will these H100s rent for in 2029?" and produce a number. Because it cannot, it demands a wide margin of safety — 375 basis points. Once a forward curve exists, three things become measurable:

1. **The value of uncontracted capacity.** The future value of capacity left over once a deal ends stops being an estimate and becomes a market price. (The industry calls this "merchant capacity": capacity that has not been pre-sold and will be sold on the open market.)
2. **Whether a contract is still attractive.** A fixed-price agreement signed today turns into an expensive agreement from the customer's point of view if market prices fall — and the risk of non-renewal rises. Much as a tenant on a fixed rent wants to move once market rents drop. A forward curve lets you see that coming.
3. **Residual value.** The depreciation debate — six years or five — acquires an objective reference point.

And most important of all: a risk that can be measured **can be hedged**, and a risk that can be hedged **gets cheaper.**

What that is worth in money

A rough calculation gives the order of magnitude. Start with the narrowest definition: on DDTL 5.5's $2.6bn alone, the **325 basis point gap to DDTL 4.0 is worth $84m a year.**

At the upper bound: CoreWeave's recourse debt is $31.4bn. A 100 basis point compression across all of it would be worth **$314m a year** — roughly **a third** of the company's full-year 2026 adjusted operating income guidance ($960–1,150m). In reality only a slice of the debt would reprice, so the effect sits somewhere between those two numbers; but that is the order of magnitude.

Across the sector the picture is larger still: there is **more than $20bn of GPU-backed debt** in the market — Fluidstack with borrowing capacity of up to $10bn, Lambda with $500m securitized, Crusoe with $425m. The lenders include BlackRock, Blackstone, PIMCO, Carlyle, JPMorgan and Macquarie.

So compute futures' biggest contribution to the neoclouds may not be on the revenue side at all, but on the **cost of capital.**

And lenders could make it mandatory

A natural consequence follows: if hedging lowers the spread, lenders can turn it into a **covenant.** A clause reading "you will hedge this much of your uncontracted capacity with futures" could find its way into credit agreements.

That has a second-order effect, and it is critical for market structure: such a requirement creates **a natural and continuous sell side.** The biggest problem facing any newly launched commodity futures market is the chicken-and-egg liquidity trap — hedgers want tight spreads before they trade, market makers want hedger flow before they quote tightly. Credit-driven mandatory hedging flow is one of the rare mechanisms capable of breaking that loop.

For now this is a forecast, not an established practice. But it is most likely what will determine whether compute futures are stillborn or become the standard.

The limits of hedging

Now to the other side of the equation. Futures are not a magic wand. They have five serious limits.

1. Basis risk. The chart in Section 4 showed it: on the same day, for the same H100, the price range across providers was $2.19–11.06. The index is a normalized summary of that dispersion; it is not a price any company actually faces. An operator selling at $11 on Azure and one selling at $2.19 on a marketplace cannot be protected by the same contract. What is more, as noted in Section 2, an H100-hour is not always an H100-hour: region, interconnect quality, uptime guarantees and cluster scale create economically distinct goods. A customer looking for a 512-GPU co-located cluster and an 8-GPU node are pooled into the same index.

> **What is basis risk?**
>
> Say you own a gas station in rural Wyoming and you want protection against a fall in prices. But the only insurance product available to you is linked to the **national average US gasoline price.**
>
> Most of the time it works: the two move together. But they can decouple on exactly the day you need them not to — the national price can rise while your price in Wyoming falls. On that day you lose money on your business and on your insurance at once.
>
> That is **basis risk**, and in compute it is extraordinarily large: the same chip rents for $2.19 and for $11.06 on the same day.

2. Liquidity risk. The contracts launch on 5 October and are not yet trading — as of today they do not appear in the product list of a major broker like Interactive Brokers. And CME and ICE are entering simultaneously with competing indices. Fragmentation widens the bid-ask spread; a wide spread deters the hedger; without hedgers, liquidity does not come. This is called the "liquidity trap," and two exchanges launching the same product at the same time increases that risk rather than reducing it.

3. Generation mismatch. The contracts are written on the H100 and the B200. But a neocloud's fleet is mixed: A100, H100, H200, B200, B300 and, soon, Rubin. As we saw in Section 4, these generations do not move together — the B200 index rose 24.4% in the first quarter of 2026 while the H100 took a different path. Hedging a B300 fleet with an H100 contract is like hedging natural gas with a heating oil contract: the correlation is there, but it can break at exactly the moment you need it.

4. Utilization risk — and this is the most fundamental limit. Futures hedge a **price.** A GPU-hour you cannot sell has no price. When demand collapses, a neocloud's problem is not that rental is $1.80 instead of $2.72; it is that nobody is renting at all. Futures offer no protection in that scenario; worse, an operator whose capacity sits idle while prices rise loses money on the futures it sold without seeing any increase in revenue. In short: **futures transfer price risk, not utilization risk.**

There is a striking detail here. Futures do not solve utilization risk — but for CoreWeave, somebody else already has: **NVIDIA.** The $6.3bn agreement discussed in Section 3 does exactly this; NVIDIA buys the capacity that cannot be sold.

So the neoclouds have two distinct risks, and each is covered from a different place: **price risk** from the market (via futures), **utilization risk** from the supplier (via NVIDIA's guarantee). That is an unusual structure — and it raises a question: if NVIDIA does not maintain that guarantee, who takes on the risk?

5. Counterparty and market-structure risk. As of July 2026 no GPU futures venture had received final CFTC approval. There is also an unresolved question on index methodology: Silicon Data builds its index from observed prices, Ornn only from executed transactions. In thin volumes, a settlement window open to manipulation could do to compute what happened to LIBOR. And if in the early years most positions are concentrated among a handful of players, the settlement price itself could become contentious.

Section summary

Compute futures could change neocloud economics as follows:

Expected effect	Realistic assessment
Revenue visibility ↑	**Limited** — 96% of revenue is already contracted
Cash flow volatility ↓	**Limited, but meaningful at renewal**
Margin predictability ↑	**Moderate** — partial, because of basis risk
Financing risk ↓	**High — this is the real channel**
Cost of capital ↓	**High** — 100bp is 12–15% of annual operating income
Capital planning ↑	**High** — residual value becomes measurable
Idle capacity risk ↓	**None**

In short: the real value of compute futures for the neoclouds is not in the income statement but on the **right-hand side of the balance sheet.**

And the precondition for it is the existence of a forward curve. The good news is that, in a sense, that curve has already begun to form. That is the subject of the next section.

6. The compute forward curve

The most valuable output of a commodity's financialization is not the contract itself but the **curve** the contracts form. In oil, Brent's forward curve carries far more information than the spot price alone: the market's expectations about the supply-demand balance, inventory levels and geopolitical risk all live there.

In compute this curve is not trading yet — the contracts launch on 5 October. But a preliminary version **already exists**, and it can be read.

First, two concepts to separate

There are two distinct objects in the data Silicon Data publishes, and confusing them is a common mistake:

Term rate. The hourly price of a fixed-duration rental contract you could sign today. A 36-month term rate means "you will pay this much per hour on average over the next three years." This is a price actually quoted in the market.

Forward rate. The price calculated as applying at a specific moment in the future. It is derived mathematically from term rates. The month-36 forward rate answers the question "what will an H100 rent for per hour three years from now?"

The difference between the two is the same as the difference between a bond's yield curve and its forward rate curve. The term rate is an average; the forward rate is marginal. And the marginal is always more extreme.

An everyday analogy: imagine you rent a house for three years at $3,000 a month. That is the **term rate** — the average of the three years. But it does not mean rent will still be $3,000 in the third year; the first year is probably cheaper and the last year more expensive. The third year's rent itself is the **forward rate.** Averages always soften the ends.

The data today

The term rate structure published by Silicon Data as of 19 July 2026:

GPU	Spot	36-month term rate	Discount
B200	$5.62	$5.17	**−8.0%**
H100	$2.72	$2.38	**−12.5%**
A100	$1.65	$1.40	**−15.2%**

The curve slopes downward for all three GPUs. In commodity language this is **backwardation**: the future is priced more cheaply than today.

> **Backwardation and contango**
>
> Tomatoes are $2 a pound in August. If you struck a deal today for January delivery, what would it cost?
>
> More — because tomatoes are hard to come by in winter. That is **contango**: the future is dearer than today.
>
> Now consider the reverse: drought has pushed tomatoes to $6 a pound today, but everyone knows next year's harvest will be plentiful. Then a deal struck today for future delivery is done more cheaply. That is **backwardation.**
>
> Backwardation means the market is saying "today's scarcity is temporary." That the curve points this way for all three GPUs shows the market expects today's bottleneck to ease.
>
> **Data note.** These term rate levels come from Silicon Data's subscription portal and are not publicly verifiable. What is more, the company's 20 April 2026 announcement introducing its forward curve states that "long-term GPU contracts do not carry a meaningful discount to on-demand pricing, and in some cases trade at a premium" — which sits in tension with the consistent backwardation shown here. All of the analysis below rests on the accuracy of this data set; the reader needs to treat it as an assumption. Real, verifiable price discovery begins on 5 October.

And the ordering here is no coincidence. The newest generation carries the smallest discount (−8%), the middle generation sits in the middle (−12.5%), and the oldest carries the largest (−15.2%). **The market is pricing technological obsolescence monotonically.** The older a GPU, the cheaper its future. Classic commodities have no equivalent; there is no such thing as a 1972-vintage barrel of oil.

But is that discount large?

The real question is whether −12.5% is a lot or a little. Without a reference point it means nothing. So let us compare it, on the same day, against a comparably structured classic commodity.

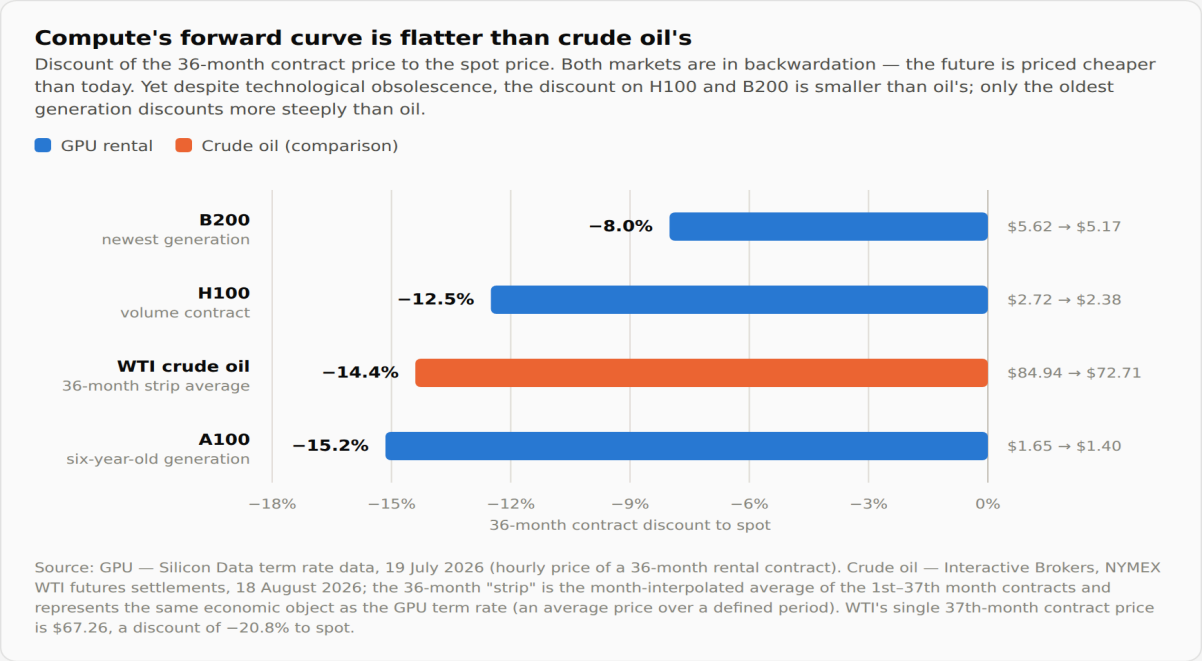

Chart 8 — Discount of the 36-month contract to spot: GPUs versus crude oil

The WTI crude forward curve as of the 18 August 2026 close is also in backwardation. But to make the comparison properly you have to match apples with apples: a GPU's 36-month term rate means "the average price over 36 months." The oil equivalent is not the price of a single delivery month but the **36-month strip average** — the average of every delivery month from 1 through 37.

On that basis WTI's 36-month strip is $72.71 against a spot price of $84.94: **−14.4%.**

The result:

> **The B200 (−8.0%) and the H100 (−12.5%) carry smaller discounts than crude oil (−14.4%). Only the six-year-old A100 (−15.2%) has a steeper discount than oil.**

This is perhaps the most counterintuitive finding in the piece. Given that compute is supposedly "a commodity that goes obsolete in 18 months," you would expect the forward curve to be far steeper. It is not. For the reason we saw in Section 4: the binding constraint is not silicon but electricity. And in a world where an A100 can be put under contract out to 2029, the market values the future of the old chip more highly than one might suppose.

For the record: WTI's single month-37 contract price is $67.26, or −20.8% against spot. That is a forward price for one specific month — not directly comparable to a GPU term rate, but it shows how steep the oil curve really is.

If we derive the forward rates

Getting from term rate to forward requires an assumption. Under the simplest one — that price declines linearly over the term — the month-36 forward rate becomes `2 × term − spot`:

GPU	Derived month-36 forward	Versus spot
B200	$4.72	−16.0%
H100	$2.04	−25.0%
A100	$1.15	−30.3%

It is possible to test how reliable that derivation is: apply the same linear assumption to the oil curve and you get $60.48 for month 37; the actual price is $67.26. So the linear assumption places the far end 10% below the actual price — expressed as a discount it says −28.8%, where the truth is −20.8%. Too steep. The reason is that real forward curves are convex: the decline is fast in the early years and then slows.

The figures above should therefore be read as a **conservative upper bound.** For the H100 the true expectation probably lies somewhere between −25% and −12.5%.

And this settles the depreciation debate

This was the question we left unresolved in Sections 3 and 4: does a GPU lose its value over six years, five years — or much faster?

The forward curve gives that question a **market-priced** answer for the first time. Look at the rate at which the H100's rental revenue erodes over three years:

Measure	Annual decline in rental price
On a term rate basis (−12.5% / 3 years)	**−4.4% / year**
On a derived forward basis (−25% / 3 years)	**−9.1% / year**
Implied by CoreWeave's 6-year depreciation	−11.7% / year
Implied by 5-year depreciation (Nebius)	−14.0% / year
Implied by 4-year depreciation (Nebius's pre-2026 policy)	−17.5% / year

The erosion rate the market is pricing is **slower than the six-year depreciation schedule even in the worst case.** And markedly slower than the five- and four-year schedules.

Does that vindicate CoreWeave's accounting choice? Partly. Two caveats are essential:

First, **a rental price is not an asset value.** A GPU's economic value is the present value of the rental stream it will generate over its remaining life; even if price falls slowly, value falls fast if the remaining life is shortening.

Second, and more importantly, **a term rate is an asking price, not a traded futures price.** Today's curve is derived from the prices providers are asking for long-term contracts. A seller's own pricing is not the same thing as a price formed on an exchange where buyer and seller meet. Real price discovery begins on 5 October.

What to watch in the curve from here

Once the curve begins trading, three shapes will point to three different economies:

Backwardation (today's state). The future is cheaper. There are two possible reasons — either the market expects supply to catch up, or it is pricing technological obsolescence. The ordering across generations (B200 < H100 < A100) suggests the second explanation dominates.

A shift into contango. The future is dearer. This would mean the market expects the compute shortage to *deepen* — the scenario in which the power constraint binds harder in 2029 than it does today. That is the best possible news for the neoclouds: renewal price risk disappears.

A humped curve. Rising in the near term, falling in the long term. This would indicate that a specific generational transition is being priced — Rubin coming online, for example. The emergence of that shape would make the collective expectation about which chip disrupts supply on which date legible.

And a deeper possibility: **the spread between generations could itself become an indicator.** The gap between the 12-month forwards for the B200 and the H100 is the market's expectation of how fast Blackwell supply normalizes. A narrowing spread means new-generation supply is easing; a widening one means the bottleneck persists.

In time, this curve could become one of the earliest warning indicators in the AI sector. In the next section we will connect it to valuation; in the eighth, to the rest of the chain.

7. From the price of compute to valuation

Now let us close the chain. In Section 6 we saw what the market expects for the price of compute. The question of this section is: **if that expectation is realized, what happens to the stock?**

First, how the model should not be built

The standard method runs like this: cut revenue by 20%, hold the profit margin constant, translate the result into a valuation. For neoclouds this gives the wrong answer — and the reason matters.

A shock to GPU rental prices **does not change costs.** The electricity bill, the data centre rent, depreciation and interest all stay the same. The only thing that changes is the revenue each GPU-hour brings in. Holding the margin constant therefore systematically understates the impact of the shock.

The right setup is this: **hold costs fixed and apply the shock to revenue alone.** Only then does the real effect become visible.

To make it concrete: a shop with $100,000 of fixed monthly costs and $120,000 of monthly revenue makes a profit of $20,000. If revenue falls 20% to $96,000, profit does not go to zero — it **turns into a loss.** Because costs are still $100,000. A 20% fall in revenue produces a 200% swing in profit. This is called operating leverage, and it is where the real story of neocloud economics lives.

The second correction comes from the finding in Section 3: most revenue is contracted. So the shock does not hit at full force today. Which is why we build two scenarios:

- **Near term:** 40% of revenue is price-exposed (newly signed capacity plus the slice that renews annually).
- **Full renewal:** 100% of revenue is price-exposed — the post-2029 world in which every contract has rolled.

The model base

	CoreWeave	Nebius
2027E revenue (midpoint of 2026 exit ARR)	$19.0bn	$8.0bn
Adj. EBITDA margin (Q2 2026)	59%	41%
Base EBITDA	$11.21bn	$3.28bn
Capacity-linked fixed cost	$7.79bn	$4.72bn
Enterprise value (EV)	$97.74bn	$69.89bn
Net debt	**$46.03bn**	**$2.16bn**
Share count	551.5m	274.1m
Implied 2027E EV/EBITDA	**8.7×**	**21.3×**

The last line is a finding in its own right. We are looking at how many times its annual profit a company is worth — like the ratio of a shop's sale price to its annual profit. The market is paying **2.4 times** as much for each dollar of Nebius's profit as it does for CoreWeave's. Some of that gap can be explained by growth rates (Nebius is growing 454%, CoreWeave 112%) — but the balance sheet difference we saw in Section 3 is also in the price: one is growing on customer prepayments, the other on $46bn of net debt.

The model reproduces current share prices (CoreWeave $93.8 calculated / $93.0 actual; Nebius $247 / $247), so its internal consistency is established.

Results

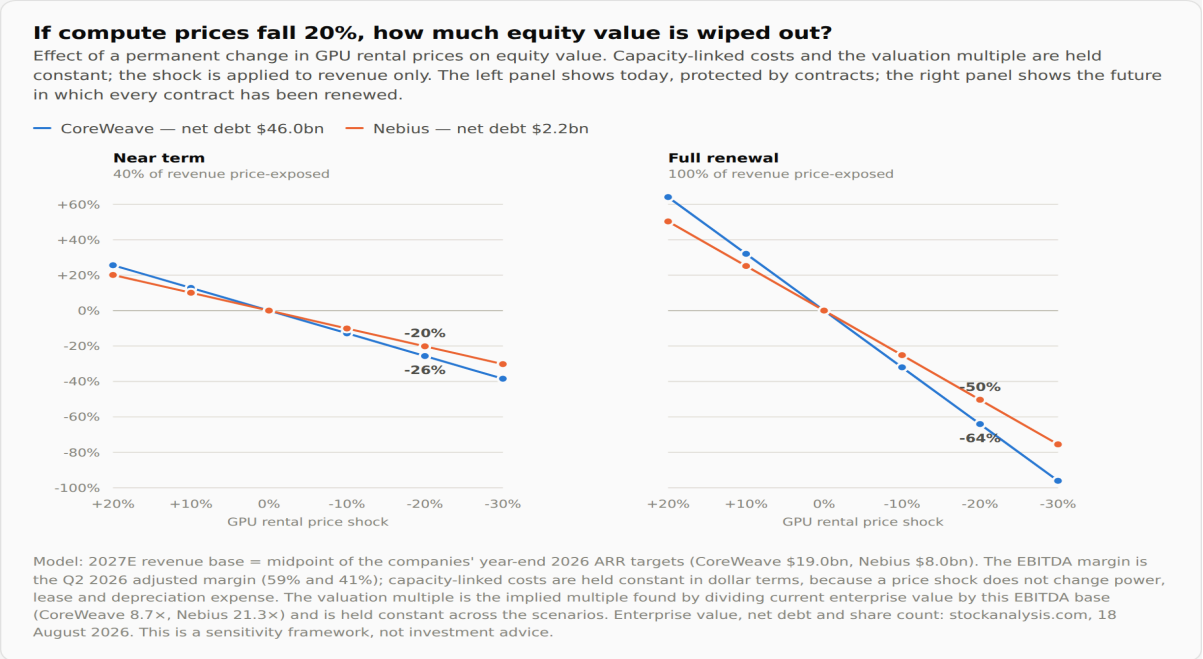

Chart 9 — Impact of a compute price shock on equity value

Near-term scenario (40% of revenue exposed):

Price shock	CoreWeave equity value	Nebius equity value
+20%	+26%	+20%
+10%	+13%	+10%
Base	**$51.7bn · $94**	**$67.7bn · $247**
−10%	−13%	−10%
−20%	**−26%**	**−20%**
−30%	−38%	−30%

Full renewal scenario (100% of revenue exposed):

Price shock	CoreWeave equity value	Nebius equity value
+20%	+64%	+50%
+10%	+32%	+25%
Base	**$51.7bn · $94**	**$67.7bn · $247**
−10%	−32%	−25%
−20%	**−64% ($34)**	**−50% ($123)**
−30%	**−96% ($4)**	−76% ($61)

The question posed was: *"if H100/B200 rental rates fall 20%, which company is more vulnerable?"*

The answer is clear: **CoreWeave.** In the full renewal scenario a 20% price decline erases 64% of equity value; at Nebius the figure is 50%. On a 30% decline CoreWeave's equity is effectively wiped out (−96%); on the model's arithmetic the threshold that zeroes out its equity is **−31.2%.** At Nebius the same threshold is **−39.7%.**

But why? Two different kinds of leverage

What is interesting is that the answer is finer than a simple "CoreWeave is riskier." Two forms of leverage work against each other.

Operating leverage — here Nebius is more fragile. Nebius's EBITDA margin is 41%, CoreWeave's 59%. The lower the margin, the more a change in revenue is amplified in EBITDA. On a 20% price decline Nebius's EBITDA erodes 49% against CoreWeave's 34%.

Financial leverage — here CoreWeave is far more fragile. CoreWeave's net debt is $46.0bn, or **47%** of enterprise value. At Nebius net debt is $2.2bn, just 3% of enterprise value. Debt dumps every fall in enterprise value straight onto the equity; the equity is the residual left above the debt.

Put the two together and financial leverage wins. Even though CoreWeave's EBITDA falls by less, the entire decline lands on a far thinner layer of equity.

> **Two kinds of leverage, in one example**
>
> Two people each open the same shop. The shop is worth $10 million and earns $1 million a year.
>
> **Anna** bought hers entirely with her own money. **Ben** borrowed $8 million and put in $2 million of his own.
>
> Now suppose the sector deteriorates and the shop's value falls 20%: from $10 million to $8 million.

> Anna's wealth went from $10 million to $8 million — **a 20% loss.** Ben's: the shop is worth $8 million, the debt is $8 million, what is left is **zero.** The same 20% decline means a one-fifth loss for one and complete wipeout for the other.
>
> CoreWeave is Ben. Net debt is 47% of enterprise value. Nebius is closer to Anna: net debt is 3%.

This also explains the divergence in share performance from Section 3: over the past twelve months CoreWeave returned −1% and Nebius +260%. The market punished the debt-financed one of two companies tied to the same commodity.

What does the forward curve say?

Now put Section 6's output into this model. The 36-month change the market is pricing for the H100 was −12.5% on a term rate basis and −25% on a derived forward basis. In the full renewal scenario:

Scenario priced by the market	CoreWeave	Nebius
Term rate basis (−12.5%)	−40% → $56	−31% → $169
Derived forward (−25%)	−80% → $19	−63% → $92

This table should be read carefully. **It is not a share price forecast.** The model assumes capacity stays constant — whereas both companies are multiplying their capacity, and growth could more than offset a price decline. The multiple is also held constant.

What the table says is narrower but sharper: **today's share prices assume the price of compute stays where it is. The forward curve says it will not.** The gap between them closes through capacity growth, multiple expansion, or the share price.

The limits of the model

To be honest about it, there are four weak points:

1. **The multiple is held constant.** In reality multiples compress in bad scenarios too — meaning the downside could be worse than shown here.
2. **Costs are assumed entirely fixed.** In reality management responds and some costs are variable — meaning the downside could be better than shown here. (1 and 2 partly offset each other.)
3. **Capacity is fixed.** Both companies are growing aggressively; this model does not account for growth.
4. **The exposure assumptions are estimates.** Both the 40% near-term figure and the business model exposure coefficients in Section 10 are the author's estimates. The overwhelming majority of CoreWeave's revenue is contracted with a weighted-average term of around five years; but new capacity is sold at current prices and the company is multiplying its capacity.

For that reason the tables should be read not as a price target but as a **sensitivity map.**

And the real point

Underneath all of this analysis sits a single sentence:

> The equity value of these two companies is tied to a variable nobody has been able to price — the price of a GPU-hour in 2029 — **with a sensitivity of between 1 and 3 times.** That is, every 1% change in the GPU-hour price moves shareholder wealth by 1–1.3% in the near term, and by 2.5–3.2% once all contracts have renewed.

From October 2026 that variable will have a market price. And that changes three things at once:

For the investor, a measurement tool: buying a neocloud stock is really taking a long position in the compute forward curve. Until now the price of that position was invisible.

For the company, a defensive tool: as we saw in Section 5, hedging renewal risk lowers the cost of capital — and at a company with $31.4bn of recourse debt, 100 basis points can amount to as much as a third of expected annual operating income.

For the market, a decomposition tool: today CoreWeave trades at 8.7× and Nebius at 21.3×. How much of that gap is growth, how much is leverage, and how much is expectations for the price of compute is currently a matter of interpretation. Once the forward curve trades, the third component becomes separable.

8. Second-round effects: the rest of the chain

So far we have looked at what the price of compute does to the neoclouds. But the neocloud is a link in the middle of the chain. Once the price signal settles there, what propagates upstream and downstream?

The chain runs like this:

> **Compute forward curve ⟶ Neocloud / hyperscaler investment decision ⟶ GPU order ⟶ Semiconductors ⟶ HBM and networking ⟶ Data centre ⟶ Power and cooling**

Every link has its own lead time, its own bottleneck and its own price dynamic. And there is an important asymmetry: **they do not all respond at the same speed.**

Where is the scarcity most severe?

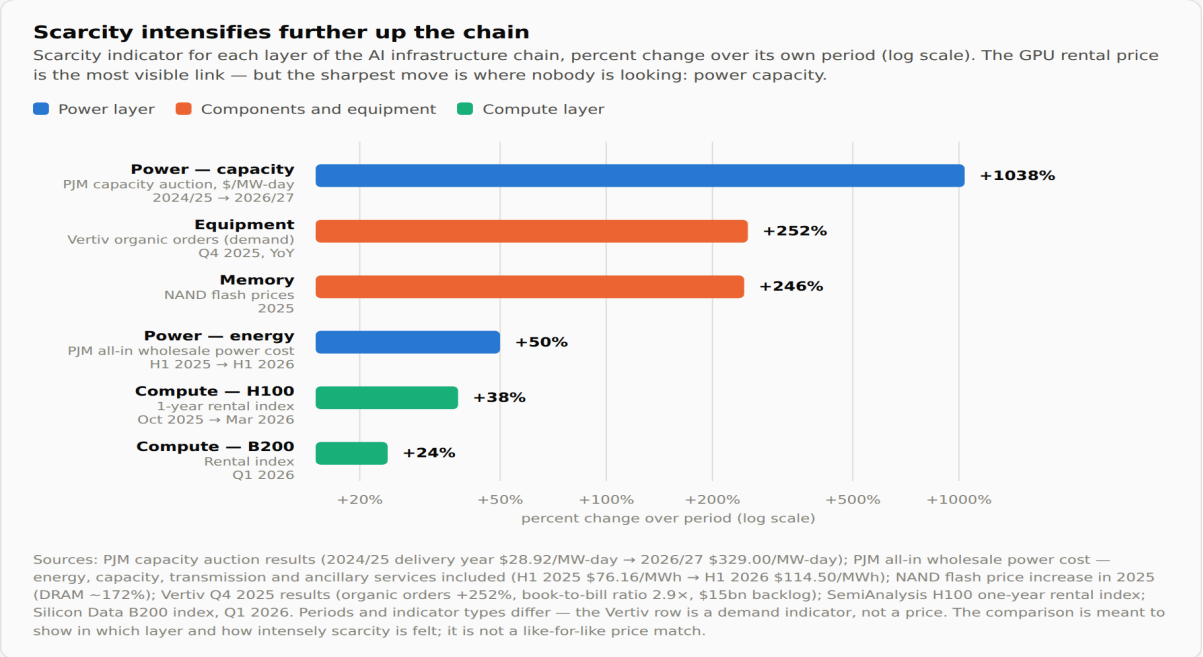

Chart 10 — Scarcity signals along the chain

Put the moves in each layer of the chain side by side and the picture that emerges is counterintuitive. The GPU rental price everyone talks about is in fact one of the **calmest** links in the chain:

Layer	Indicator	Change	Period
Power — capacity	PJM capacity auction	**+1,038%** (11.4x)	2024/25 → 2026/27
Equipment	Vertiv organic orders	+252%	Q4 2025, year on year
Memory	NAND flash prices (DRAM ~172%)	+246%	2025
Power — energy	PJM total wholesale electricity cost	+50%	H1 2025 → H1 2026
Compute	H100 one-year rental	+38%	Oct 2025 → Mar 2026
Compute	B200 index	+24%	Q1 2026

PJM's capacity auction went from $28.92 per MW-day to $329 — eleven times higher. This is the macro-level confirmation of Section 4's finding: **the bottleneck is not in silicon, it is in electricity.**

The physical reality behind the numbers is starker still:

- In the US, **more than 2,060 GW** of active projects were waiting in the interconnection queue as of the end of 2025 (down from the peak of 2,600 GW at the end of 2023, but still twice installed capacity); the median time to commercial operation **exceeds five years**, whereas a data centre is built in under three.
- PJM has a capacity shortfall of **6.6 GW** in the 2027/28 delivery year and **6.8 GW** in 2028/29.
- In 2025, **1,891 electricity projects (266 GW) were cancelled** in the US. The reasons, in order: local opposition, interconnection costs, the deteriorating economics of battery storage, a federal policy shift on offshore wind, and queue reforms.
- On the HBM side, SK Hynix and Micron have sold out their capacity through the end of 2026; Samsung is wrestling with qualification problems. SK Hynix holds 62% of the market and roughly 90% of NVIDIA's HBM supply; the remainder is split between Samsung (~30%) and Micron (~8%).
- Vertiv's book-to-bill ratio is 2.9x, its backlog is $15 billion and deliveries are scheduled **12–18 months out**. In the CEO's words, "it has never been harder to connect to the grid."

Which link in the chain has a price?

Now to the real structural question. If an investor wanted to price this chain, for which links could they find a forward market price?

Layer	Is there a forward price?	How far forward?
Power — energy	**Yes** (electricity and natural gas futures)	Years
Power — capacity	**Yes** (PJM capacity auctions)	3 years
Compute	**Yes, from 5 October 2026**	**36 months**
Data centre capacity	Partly (lease contracts, illiquid)	—
HBM / memory	No (bilateral annual contracts)	—
Networking / GPU	No (order books, disclosed with a lag)	—

The table shows the real structural contribution of compute futures: **there were prices at both ends of the chain, but not in the middle.** The price of electricity was known three years forward; the price of the compute that consumes that electricity was not. From October, the gap in between is closing.

So could this be a leading indicator?

The answer is "yes, but not in the direction people assume."

For it to be a leading indicator, the compute forward curve would have to move **before** events in the other links of the chain. The causal chain works like this:

> A neocloud looks at the expected rental price in 2029 and commits to capacity today → that commitment turns into a GPU order → the GPU order turns into HBM and networking orders → those turn into data centre, power and cooling orders.

Lead times genuinely make this chain a leading one: 12–18 months for cooling and power equipment, 5–7 years for grid interconnection. In other words, **today's compute forward curve is the decision input that determines what equipment gets ordered between 2028 and 2031.**

Three signals can be tracked concretely:

1. The curve flipping from backwardation to contango. If the market starts pricing in a deepening compute scarcity, that means capacity investment is about to accelerate. It is a bullish signal for power, cooling and HBM. Today's curve points the other way.

2. The B200–H100 forward spread. The gap between the 12-month forwards of the two generations is the market's collective expectation of how fast Blackwell supply will normalise. A narrowing spread means new-generation supply is easing; a widening spread means the bottleneck persists. This is an indicator that can be read before NVIDIA discloses shipment data.

3. The gap between the long end of the curve and depreciation assumptions. We calculated it in Section 6: the market is pricing 4.4–9.1% annual erosion for the H100; CoreWeave's accounting assumes 11.7%, Nebius's 14.0%. If that gap closes or reverses, an industry-wide change in depreciation policy — and therefore an earnings revision — follows.

But the reverse is also true: the price of compute is an *outcome*

Here it is worth being honest. The compute forward curve is not a universal leading indicator for the chain as a whole, because causality does not run in one direction.

The price of electricity is an input to the cost of compute; the price of compute is an output of electricity demand. The PJM capacity auction already produces a price three years forward, and it moved earlier than compute did (an elevenfold increase, while GPU prices were still at their lows). In other words, **the power curve leads the compute curve, not the other way round.**

The unique contribution of the compute forward curve lies elsewhere: **it prices the demand side.** Every other forward price in the chain (electricity, gas, capacity) is a supply and cost signal. The compute forward curve, by contrast, is the first market price for the future value of AI workload demand.

And why this is not a "spark spread"

In energy markets the margin of a gas-fired plant is measured by the *spark spread*: the price of electricity minus (the price of gas × the heat rate). Because both have futures markets, the plant's future margin can be traded today.

Constructing the same thing in compute looks tempting: the GPU-hour price minus the cost of electricity. But the numbers break the analogy. In an H100-hour, energy and facility cost is only **8.3%** of the price; electricity alone is **1.4%**. In a gas plant, fuel is 60–70% of cost.

Think of the difference this way: for a baker, flour is the largest cost item — when flour gets more expensive, bread gets more expensive, and the link is direct. For a neocloud, electricity is more like the salt in the price of the bread. If electricity doubled, the GPU-hour price would rise by all of 1.4%.

So a neocloud is **not a conversion business.** It does not operate like a power plant turning a cheap input into an expensive output; it operates like a business renting out scarce capacity — much more like shipping freight, data centre rent or hotel occupancy. Because more than half the price is scarcity rent, the forward curve is not an indicator of **cost margin** but an indicator of **scarcity expectations**.

That distinction is critical for any investor who intends to interpret the curve: a falling compute forward curve does not mean costs are falling. It means scarcity is easing — that is, that the neocloud's reason for existing is weakening.

Which companies in the chain are sensitive to what?

Layer	Companies	Sensitivity to the compute forward curve
Neocloud	CoreWeave, Nebius	**Direct and very high** — a 1% change in price moves equity value by 1–3x
GPU	NVIDIA, AMD	High, but lagged — the curve determines the order decision, and the order determines revenue
HBM	SK Hynix, Micron, Samsung	High; because capacity is sold out through 2026, the impact lands in 2027+
Networking	Broadcom, Arista	Moderate; depends on cluster scale
Power and cooling	Vertiv, generators, infrastructure	Moderate — but their own curves already move earlier
Hyperscaler	Microsoft, Amazon, Google	Low; compute is a cost line for them, not a revenue line

The last row matters. For hyperscalers, a falling compute price is **good news** — their inputs get cheaper. For neoclouds it is a disaster — their output gets cheaper. The same futures contract puts these two groups on opposite sides of the market.

And that is exactly what a commodity futures market needs in order to survive: two large parties with opposing interests in the same price.

9. Compute and oil: where the analogy breaks

Let us return to the question in the title. Is compute really "the oil of the AI age"?

CME itself made the comparison institutionally: it placed the contracts under the energy desk, and had the executive responsible for energy products make the announcement. The structural parallel is real too: a scarce input to production, a fragmented spot market, a reference index, a futures curve, and producers and consumers who hedge.

But when you take the analogy seriously and test it, five points emerge where it breaks. And the direction of the breaks is interesting: the further compute moves from oil, the closer it gets to **other commodities**.

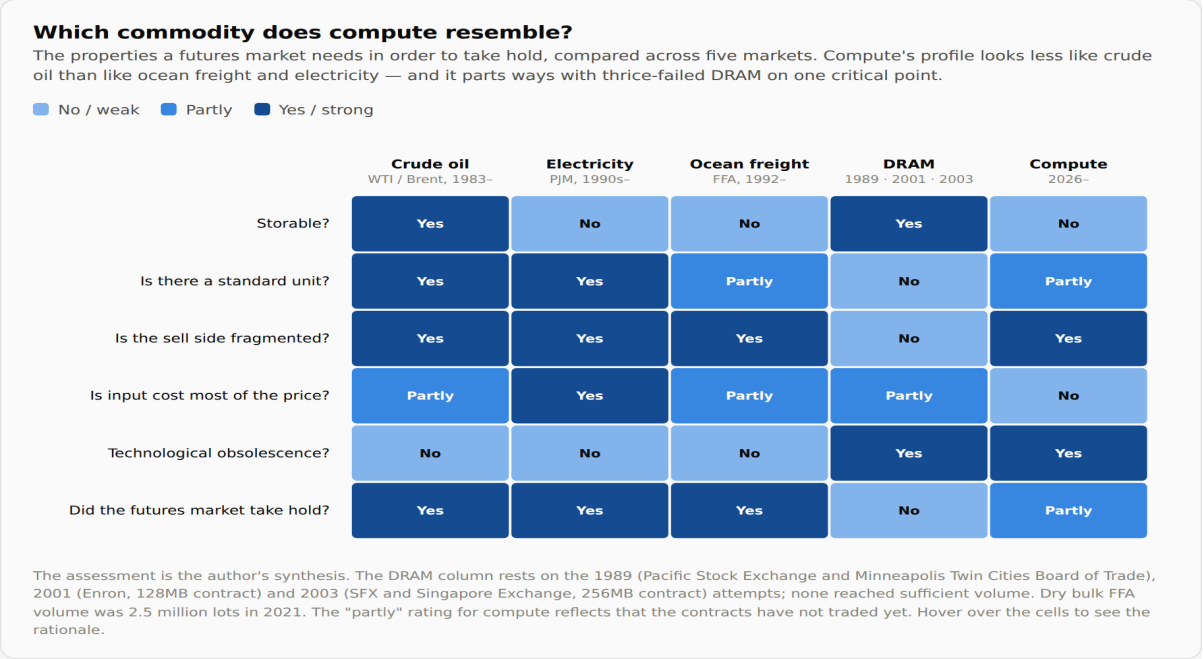

Chart 11 — Which commodity does compute resemble?

1. Compute cannot be stored — and that sets the curve free

This is the most fundamental difference, and the one whose consequences are least understood.

Oil can be stored. That single property mathematically constrains the shape of the forward curve: if the future price rises far above the spot price, buying today, storing it and selling later becomes profitable. That arbitrage caps contango within a band **the width of the cost of storage**. The oil curve is not a forecast; it is partly an accounting identity.

There is no such anchor in compute. An unused GPU-hour vanishes; you cannot "store" it and sell it next year. The compute forward curve is therefore not tethered to the cost of carry — it is **pure expectation**. That means two things:

- The curve may be more volatile and less predictable than the oil curve.
- But it is also more *informative*: it contains no cost-of-storage noise, only supply-and-demand expectations.

This property brings compute much closer to **electricity** than to oil. Electricity cannot be stored either, its curve is also pure expectation, and its markets (PJM, ERCOT) have deepened on a regional basis.

> **What does non-storability change?**
>
> Think of an airline seat. The moment the plane takes off, the empty seat ceases to exist — you cannot "inventory" it and sell it next week. A hotel room, a concert ticket, a cinema seat: all the same.
>
> Wheat is not like that. If the price is low you store it, and when the price rises you sell. That option stops the future price from straying too far from today's price — because the moment the gap widens, someone buys and stores at a profit and the difference closes.
>
> A GPU-hour is on the airline-seat side. An unused hour vanishes. That is why there is no mathematical anchor tying compute's forward curve to anything. The curve is entirely expectation — which makes it both more volatile and more informative.

2. More than half the price is not cost, it is scarcity rent

We saw it in Sections 4 and 8: in an H100-hour, energy and facility cost is **8.3%** of the price, and electricity alone roughly **1.4%**. In a gas-fired plant, fuel is 60–70% of marginal cost; in crude oil, the cost of production is the defining component of the price.

So compute is **not a conversion business.** It does not operate like a facility turning a cheap input into an expensive output. It operates like a business renting out scarce capacity.

That tells us what the right analogy is: **ocean freight.** On a dry bulk carrier too, what determines the price is not fuel but how many ships happen to be available on that route that week. A voyage that runs empty cannot be recovered, just like an idle GPU-hour. And freight derivatives (FFAs) work exactly like compute futures: no physical delivery, cash settlement against the **arithmetic average of an index over the month**. CME's preference for Asian-style settlement is no coincidence; it is the same solution to the same problem.

3. Compute is not one commodity but a family

In oil, a barrel is a barrel; there are fixed discounts for API gravity and sulphur that have not changed in decades. In compute there is the A100, H100, H200, B200, B300 and soon Rubin — each an imperfect substitute for the others, each with its own supply curve and its own obsolescence schedule.

Section 6's data showed this in measured terms: the 36-month discount is 8% for the B200, 12.5% for the H100 and 15.2% for the A100. The same service has three different forward curves, and they do not move together.

Layer a horizontal dispersion on top of that. The chart in Section 4 showed that on the same day, for the same H100, the price range across providers was $2.19–$11.06 (5.1x). Region, interconnect quality, uptime guarantees and cluster scale create economically distinct goods.

The upshot: basis risk in compute is structurally larger than in classic commodities — both vertically (across generations) and horizontally (across providers).

4. Technological obsolescence — but more measured than assumed

The opening argument of this section was: *"Oil does not become technologically obsolete in 18 months; a GPU does."*

True — and it is the one property that separates compute from every classic commodity. In Chart 11, compute shares that row's cell with **DRAM** alone.

But Section 6's data makes the size of that difference measured. The market is pricing 4.4–9.1% annual erosion for the H100; the 36-month discount is smaller than crude oil's (14.4%). Obsolescence is real, but its expression

in the price is far gentler than the "scrap in 18 months" narrative — because, as we saw in Section 4, the binding constraint is not silicon but electricity.

5. And the real difference: this has been tried three times before

The most instructive break in the analogy is historical. Because the attempt to financialise the chip market is not new — and until now it has always failed.

- **1989:** The Pacific Stock Exchange and the Minneapolis Twin Cities Board of Trade proposed futures on 256KB and 1MB DRAM. CFTC approval never came.
- **2001:** Enron announced it would launch a 128MB DRAM contract. The project went down with the company.
- **2003:** SFX in the US and the Singapore Exchange listed 256MB DRAM futures. Neither reached sufficient volume.
- **2004:** Patents were granted for a DRAM futures exchange and for TSMC fab capacity futures. None came to life.

Two things killed these attempts, and both are directly live questions for compute:

Killer one: unit standardisation. In the words of one 2003 analyst, "you can define a standard density, a standard type and a standard configuration — but there can be infinite combinations of these." A wheat contract can stay the same for decades; a chip model becomes obsolete in three to five years. That mismatch between the life of the contract and the life of the product was never bridged.

How does compute escape this trap? Here is the critical design difference: **CME's contract references not a chip but the hourly price of a service.** There is no physical delivery; what is standardised is not the hardware but the *rental rate*. A GPU-hour is far easier to index than the chip itself — because providers like Silicon Data and Ornn can normalise different configurations down to a single rate. Compute futures solve the problem DRAM futures could not, by changing the good.

Killer two: seller concentration. A futures market requires fragmented participants on both sides. The number of DRAM manufacturers was around 17 in 1989; today it is three. The producers were reluctant too: one Micron executive said "we do not adjust production to what the market needs, we run at full capacity"; Texas Instruments preferred a direct customer relationship to anonymous commodity trading. The reason oil futures took hold in 1983 was precisely the opposite: independent producers, wildcatters, a fragmented supply side looking for buyers.

Compute is differently positioned here too. GPU *manufacturing* is extremely concentrated — NVIDIA dominates, and SK Hynix holds 62% share in HBM. But the side that sells GPU-**hours** is fragmented: dozens of neoclouds, marketplaces and hyperscalers. In other words, the layer being financialised is not the layer where the concentration sits. That is an advantage DRAM never had.

But this fragmentation did not arise on its own.

The story we told in Section 3 connects precisely here. That the rental layer is fragmented while GPU manufacturing is a monopoly is no accident; it is NVIDIA's deliberate choice. The company could have sold every chip it made to three or four hyperscalers — in which case the sellers of GPU-hours would, exactly as in DRAM, have consisted of a handful of giants. Instead it created dozens of sellers from scratch, using allocation priority, capital and offtake backstops.

The result gives rise to this paradox:

> **The structural feature that allows compute futures to escape DRAM's fate — a fragmented seller side — was produced not by the market itself but by the monopoly in the supply chain.**

And what has been produced can be taken back.

This is a fragility nobody talks about. If NVIDIA changes its allocation policy, narrows its cloud partner programme or fails to renew its capacity guarantees, the neocloud layer could consolidate quickly. When that day comes, compute futures will meet the same problem that killed DRAM futures: a few giants to sell, a few giants to buy, and nobody in between to trade.

So the long-term health of this market depends on a decision nobody voted on: how many customers NVIDIA wants to have.

The real problem Enron solved

This history also has a constructive lesson, and it connects directly to Section 5's argument.

The reason Enron succeeded in natural gas was not that it managed price risk. Producers initially would not go near forward contracts. Through the "Gas Bank" structure Enron offered them prepayment and interest rate swaps — that is, **it solved not the price risk but the credit problem.** What drew the producer into the market was not hedging but financing. No such mechanism was ever built in chips.

In compute the counterpart of that mechanism already exists, in two distinct forms:

- **The Nebius model:** 70% of contracts are prepaid and the prepayments cover 50–60% of the related capital expenditure. That is a one-for-one Gas Bank structure — the customer is providing half of the producer's financing.
- **The lender's hedging requirement:** as we argued in Section 5, if hedging uncontracted capacity becomes a credit condition, it creates a natural and continuous sell side.

But the most exact counterpart is the third. Enron's Gas Bank gave the producer two things: cash up front and a price guarantee. What NVIDIA gives the neoclouds is exactly that — capital, priority allocation and an offtake backstop for capacity that cannot be sold. The only difference is that Enron was an intermediary trader, while NVIDIA is the sole producer of the good.

In other words, the compute market's Gas Bank has already been built, and it is called NVIDIA. That was the mechanism missing in chips in 1989, in 2001 and in 2003; this time it is there from the start.

So the path by which compute futures gain liquidity will probably run not through the speculator but through the **credit market**. That is exactly what DRAM did not have.

Conclusion: where does the analogy stand?

Pulling it together, compute's profile looks like this:

Property	Closest analogue
Non-storable, curve is pure expectation	**Electricity**
Capacity rental, input cost immaterial	**Ocean freight**
Asian-style cash settlement against an index	**Ocean freight (FFA)**
High basis risk, fragmented spot market	**Electricity / freight**
Technological obsolescence	**DRAM** (the only analogue)
Deep, single-reference global market	**Oil** (not yet)

The honest answer to the question in the title: **compute is not the oil of the AI age. It is the freight and the electricity of the AI age.**

That does not diminish the comparison. Freight and electricity markets are also deep, liquid and economically decisive. But knowing the difference matters, because the right analogy sets the right expectation:

- If you expect the oil model, you expect a single global reference price and a contract that stays fixed for decades. It is not coming.
- If you expect the freight-and-electricity model, you expect **fragmented curves by region and by generation**, high basis risk, and a market that deepens over time but never fully standardises. That is probably what will happen.

And there is a third possibility: the DRAM model. That is, it never catches on at all. That is what happened in 1989, 2001 and 2003. Compute has two structural advantages over those attempts — it indexes the service rather than the good, and its seller side is fragmented. Whether those advantages are enough will become clear in the first twelve months after 5 October.

10. Investment implications

We have built the chain from end to end. Now let us tie it to three scenarios — and quantify what each scenario means for the revenue of six companies.

Three scenarios

Bull case — scarcity deepens. AI demand continues to outrun GPU supply, and the electricity constraint delays new capacity from coming online. Prices rise **20% over three years**, and capacity runs completely full.

Base case — supply catches up with demand. Capacity comes online gradually and prices soften in the way the forward curve indicates today: a **12.5% decline** over three years. Utilization holds. Scale and the cost of borrowing determine the winner.

Bear case — supply overshoots demand. Electricity capacity comes online en masse, efficiency gains on the software side get the same work done with less hardware, and one or two large customers cut back their investment. Prices fall **30%** and, on top of that, **10% of capacity sits idle.** Combine the two and revenue erodes by a third.

The same shock, six different outcomes

The critical variable is how much of each company's revenue is directly tied to the GPU-hour price. Call that share the "price-exposed share of revenue."

An important reading note: these coefficients measure **business model exposure** — that is, how much of a company's business is determined by the GPU-hour price. *When* that exposure materialises is a separate question; as we saw in Sections 5 and 7, long-dated take-or-pay contracts defer the impact for years. The table below is therefore not a point estimate for 2028 but a comparison of the revenue levels that would emerge **once the shock has fully passed through.**

Company	2026 revenue run-rate	Price-exposed share of revenue	Rationale
CoreWeave	$19.0bn	**100%**	GPU-hours are the only product
Nebius	$8.0bn	**95%**	AI cloud is ~99% of revenue
Oracle OCI	$23.2bn	**70%**	Growth is almost entirely AI infrastructure
Google Cloud	$99.2bn	**35%**	High AI share but a broad classic cloud base
Azure	$100.0bn	**30%**	~90% of Microsoft Cloud revenue comes from non-frontier-lab customers
AWS	$169.0bn	**15%**	AI revenue run-rate $25bn out of $169bn total

Combine these shares with capacity-driven growth expectations and the 2028 picture is this:

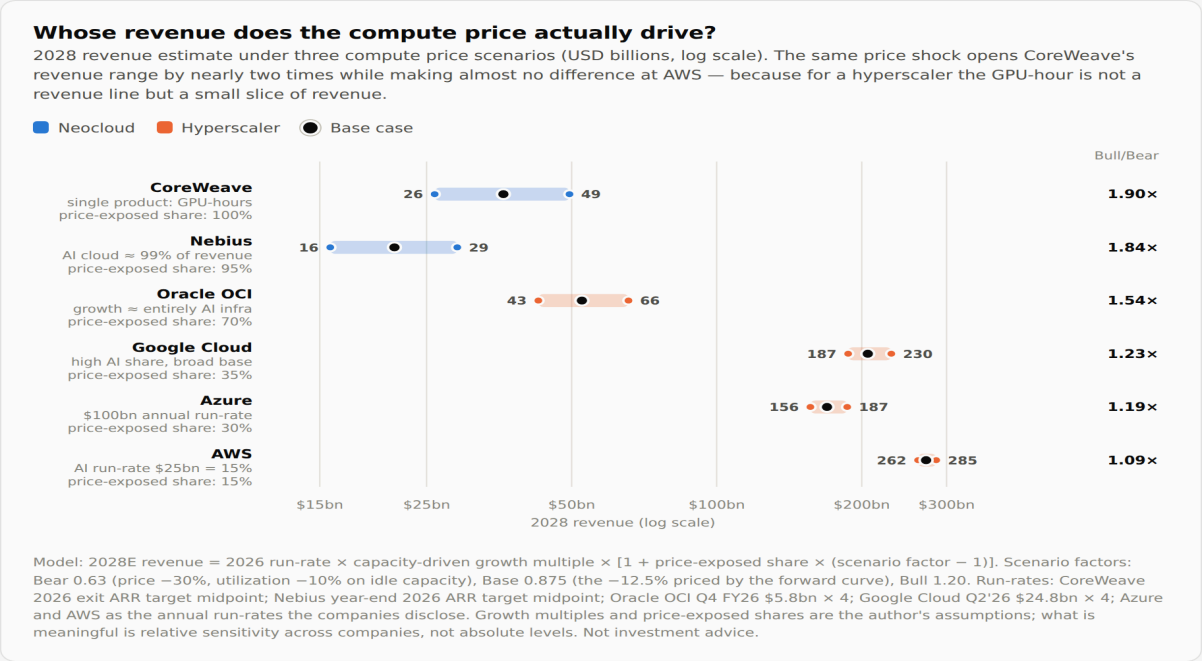

Chart 12 — 2028E revenue under three compute price scenarios

Company	Bear	Base	Bull	Spread
CoreWeave	$26.0bn	**$36.1bn**	$49.5bn	**1.90x**
Nebius	$15.8bn	**$21.4bn**	$28.9bn	**1.84x**
Oracle OCI	$42.6bn	**$52.5bn**	$65.6bn	1.54x
Google Cloud	$187.4bn	**$205.8bn**	$230.3bn	1.23x
Azure	$156.5bn	**$169.4bn**	$186.6bn	1.19x
AWS	$261.8bn	**$272.0bn**	$285.5bn	1.09x

Deviation from the base case:

Company	In the bear case	In the bull case
CoreWeave	**−28.0%**	+37.1%
Nebius	−26.4%	+35.0%
Oracle OCI	−18.8%	+24.9%
Google Cloud	−9.0%	+11.9%
Azure	−7.6%	+10.1%
AWS	**−3.7%**	+5.0%

The table reduces to a single sentence: **the price of compute is revenue for the neoclouds and a small slice of revenue for the hyperscalers.** Relative to the base case, the bear case pulls CoreWeave's revenue down 28% while making only a 3.7% difference at AWS. (These percentages are relative to the base case; the base case already embeds a 12.5% price decline, and the bear case additionally carries a 10% loss of utilization.)

And more importantly: **the sign is not even the same for a hyperscaler.** AWS and Azure both sell GPU capacity and run their own AI products on that capacity. When compute gets cheaper they lose a little on the sales side, but they gain on the margin of their own Copilot/Bedrock/Gemini. The bear case is existential for a neocloud and could be, on net, **good news** for a hyperscaler.

Protected today, exposed tomorrow

There is an observable metric for how well companies are protected in the short term: the ratio of backlog to annual revenue.

Company	Backlog	Annual revenue	Coverage
CoreWeave	$129bn	$12.8bn (2026 revenue guidance)	**10.1x**
Oracle	$638bn	$76.8bn (Q4 annualised)	**8.3x**
Nebius	~$17bn (disclosed only)	$3.0bn (ARR)	5.7x
Google Cloud	$514bn	$99.2bn	5.2x
Microsoft Cloud	$678bn	$214.0bn	3.2x
AWS	$496bn	$169.0bn	2.9x

The ratios in the table rest on different revenue definitions (guidance, annualised quarter, ARR), so they should not be compared to decimal precision. Even so the order of magnitude is clear, and it contains a paradox: **the company most exposed to a price shock is at the same time the best protected in the short term.** CoreWeave's backlog covers roughly ten years of revenue, Oracle's eight; AWS's three. So the bear case does not arrive at these two companies in 2027.

But that protection is not free. To deliver those order books:

- CoreWeave will spend $35–39 billion of capital expenditure in 2026, its net debt is $46 billion, and quarterly interest expense is $640 million.
- Oracle raised $43 billion of debt plus $5 billion of equity in FY2026, its free cash flow is **minus $23.7 billion**, and it plans roughly $40 billion more of financing in FY2027.

So the backlog is not a protection but a **commitment.** And the financing of that commitment is a bet on the price of compute on the date the contracts run out. Section 5's maturity mismatch chart described exactly this.

Forces pushing the price up

The factors feeding the bull case, most of them measurable:

1. Electricity arrives far more slowly than silicon. More than 2,060 GW of active projects sit in the US interconnection queue, and the median wait to commercial operation exceeds five years; a data centre, by contrast, is built in under three. PJM has a capacity shortfall of 6.6 GW in 2027/28 and 6.8 GW in 2028/29. In 2025, 1,891 electricity projects (266 GW) were cancelled. PJM's capacity price rose elevenfold in the 2026/27 delivery year and stayed at the same levels in the two subsequent auctions — the gap has not closed.

2. The new generation does not fit the existing data centre stock. Older facilities were designed for ~20 kW per rack; Blackwell wants 120–140 kW. This both slows new capacity coming online and preserves the value of older GPUs — CoreWeave has contracted its A100s out to 2029 and renewed its H100s at 95% of the original price.

3. The HBM bottleneck caps GPU supply. SK Hynix and Micron have sold out their capacity through the end of 2026; Samsung is having qualification problems. SK Hynix supplies ~90% of NVIDIA's HBM. NAND flash prices rose 246% in 2025 and DRAM prices ~172%, and Amazon raised its capital expenditure target from $200 billion to $220 billion citing "higher memory cost."

4. The demand side is still accelerating. AWS accelerated its growth for a fifth consecutive quarter (+36.7%), Azure +43%, Google Cloud +82%, Oracle OCI +93%. Capital expenditure guidance is being revised up: Google from $180–190 billion to $195–205 billion, Amazon to ~$220 billion. One Nebius capacity auction closed 15% above the highest price it had ever seen for Blackwell.

5. Inference load is continuous. Training comes in projects and ends; inference grows linearly with user numbers and does not stop. The floor under demand is hardening.

Forces pushing the price down

1. The high price itself calls forth supply. PJM's capacity price rising to $329 per MW-day is a signal designed precisely to trigger new generation investment. Vertiv's book-to-bill ratio of 2.9x shows that capacity is on the way. The scarcity premium — 52–64% of the price, as we saw in Section 4 — is the line item that melts when supply arrives.

2. Model efficiency could break hardware demand. Every architectural improvement that does the same work with fewer FLOPs increases effective supply. Historically this is the fastest-moving variable.

3. The forward curve is already pointing down. The market today prices backwardation in all three of the A100, H100 and B200. That means the bear case is partly priced in already — but it also shows the direction of collective expectations.

4. Financing fragility is an accelerant. CoreWeave's CDS premiums have risen to roughly 850 basis points; on a standard recovery assumption, that implies a five-year default probability of around 50%. Oracle's free cash flow is minus $23.7 billion. A price decline could create a self-reinforcing loop on these balance sheets: falling price → falling collateral value → widening spread → constrained growth.

5. A single customer stepping back. The order books are concentrated — Meta at CoreWeave ($21bn + $14bn), Microsoft at Nebius (~$17bn), a handful of giant AI labs at Oracle. A customer of that scale scaling back its plans would on its own create a supply glut.

6. A change in depreciation policy. The industry rests on a six-year useful life assumption. If an auditor or a regulator forces that shorter, earnings are revised across the board with no change in revenue.

7. NVIDIA withdrawing its support. As we saw in Section 3, the neocloud layer was built by NVIDIA: allocation priority, direct capital and an offtake backstop for capacity that cannot be sold. A narrowing of that support — a shrinking of the programme's scope, guarantees not being renewed, or new-generation allocation shifting to the hyperscalers — would hit both the neoclouds' growth and their ability to raise credit at the same time. This is the one risk on the list that depends on a single company's decision.

Company by company

CoreWeave — the highest beta. Its price-exposed share is 100%, its net debt is $46 billion, and interest expense is five times operating income. In the bull case revenue rises to $49.5 billion; in the bear case it falls to $26 billion, and on Section 7's model a price shock of that magnitude effectively wipes out the equity. The backlog protects 2027–28; the risk sits in the 2029 renewals. **The purest exposure to the compute forward curve.**

Nebius — high exposure, low leverage. Its price-exposed share is 95%, but its net debt is only $2.2 billion and 70% of its contracts are prepaid; prepayments cover 50–60% of capital expenditure and the payback period has come down to 22 months. Its operating leverage is higher than CoreWeave's (margin 41% vs 59%) but its financial leverage is far lower. **Exposure to the same commodity with a more defensible balance sheet — and at 21.3x, that has already been paid for.**

Oracle — a neocloud in hyperscaler's clothing. OCI is growing 93%, RPO is $638 billion, but free cash flow is minus $23.7 billion and another $40 billion of financing is needed in FY2027. Its price-exposed share, at 70%, is markedly higher than the hyperscalers'. **The only large-scale classic software company carrying compute price risk.**

Google Cloud, Azure, AWS — for these, the price of compute is second-order. 15–35% of their revenue is exposed, but at the same time the input cost of their own AI products is falling. The bear case is probably **net positive** for these three. Their real risk is not the price of compute but the return on capital expenditure.

What to watch

After October 2026 the evidence for this debate will sit in a chart. Concretely:

1. **The shape of the forward curve.** If backwardation is deepening, bear; if it is flattening or turning to contango, bull.
2. **The B200–H100 forward spread.** The earliest indicator of how fast Blackwell supply is normalising.
3. **The response of borrowing costs.** If Section 5's thesis is right, the existence of a hedgeable forward curve should narrow the interest rate differential applied to uncontracted capacity. If it does not narrow, the credit market does not believe in this insurance.
4. **The contract's volume.** If no meaningful open interest builds in the first twelve months, the 1989–2003 DRAM history is repeating itself and this entire analysis remains theoretical.
5. **Depreciation policies.** Will the gap between the long end of the curve and companies' assumptions close?

The last link: what this does to the AI ecosystem

Up to this point the chain ended in a financial question: what is the stock worth? But the price of compute is not a revenue line for six companies. Every answer a chatbot gives, every suggestion a coding assistant makes, every protein simulation a drug company runs — all of it consumes GPU-hours. So the fact that this price is about to have a market matters not only to CoreWeave's shareholders but to everyone building on AI.

Through four mechanisms.

First: things become plannable. Today an AI startup cannot tell its investors what its unit cost will be in 2028. Because it cannot, it cannot commit to long-dated pricing either; when an enterprise customer asks for a three-year fixed price, the startup either declines or takes the risk and gambles its margin.

Airlines are an industry that solved this problem. Before jet fuel futures became liquid, no airline could price next year's ticket today; the low-cost carriers were born once the fuel market deepened enough to let them do exactly that. A meaningful share of Southwest's cost advantage in the 2000s came not from flight operations but from the hedging desk.

In compute, the same door opens in October. The real effect on the software layer will be a change in who is able to *promise what*.

Second: cost of capital — and how it flows downstream. Section 5's finding was that the main benefit of futures to a neocloud is not revenue visibility but borrowing cost. Uncontracted capacity is charged a 375 basis point penalty today; if a hedgeable forward curve narrows that penalty, the same data center gets financed more cheaply.

That gain does not stay in one line item. If capacity becomes cheaper to build, more of it gets built; and if more of it gets built — given that, per Section 4's decomposition, more than half of a GPU-hour price is scarcity rent — the price falls. The benefit reaches an AI startup that has never seen a futures contract in its life.

The scale is not trivial. At one company: a 100 basis point narrowing on CoreWeave's $31.4bn of recourse debt is $314mn a year, roughly one third of its full-year adjusted operating income guidance. Across the industry there is more than $20bn of GPU-backed debt outstanding, and the stack is growing fast.

Third: the information gap narrows. Chart 5 shows today's reality: the same chip, on the same day, at up to a 7.7× price difference depending on the provider. What that gap means is that the large buyer knows the real price of compute and the small buyer does not. A forty-person startup has no reference against which to check whether the hourly rate it pays is above or below the market.

A public, tradable index changes that. Commodity history shows the effect repeatedly: once Brent became the reference price, small refiners could negotiate on the same information set as the large trading houses. In compute the winner will be the small buyer, and the loser the seller who benefits from today's price opacity.

Fourth: capital discipline. This is the most contested effect for the ecosystem — and it is not unambiguously good.

A forward curve does not merely show the future price; it makes the **collective belief about the future price public**. Today's curve prices roughly a 12.5% discount 36 months out. On the day that number is published, every company booking profit on a six-year depreciation assumption faces a market figure that tests the assumption.

The likely result is a slower and more selective investment cycle. In the short run that is not good news for the ecosystem: fewer data centers, slower capacity growth, prices that stay high for longer. In the long run it reduces the risk of the kind of overbuild-then-collapse seen in fiber optics in 2000 or in mortgage credit in 2007. The history of commodity markets suggests a public price does not prevent bubbles, but it shrinks them.

And the limits of the effect. Three of them matter.

Futures do not *create supply*. Section 8's finding was that the real bottleneck in the chain is electricity, not silicon: the PJM capacity price rose elevenfold in two years, and more than 2,060 GW sits in the grid interconnection queue. A compute curve prices that constraint; it does not remove it.

Basis risk hits the smallest player hardest. The index is an average; the further a small buyer's own price strays from the average, the less the hedge protects. Which means the benefit — "the information gap narrows" — and the cost — "the hedge is incomplete" — land on the same group.

And the contract is not guaranteed to work. DRAM was tried three times — 1989, 2001, 2003 — and none of the attempts built sufficient volume. Without volume, none of the effects listed here happen at all.

In sum. Compute futures do not change how many GPUs exist in the world. What they change is the quality of the *decisions* around those GPUs: who can promise what, at what rate capacity gets financed, which information is public, and how many quarters it takes for a mistake to be noticed.

For an ecosystem whose primary failure mode is hundreds of billions of dollars allocated to the wrong thing, the last item on that list is probably the most valuable.

Final word

This piece began with a news item: on 5 October, CME is launching futures on the GPU rental price.

Where it arrives is this: the hundreds of billions of dollars poured into AI infrastructure have until now been betting on a variable **that had no price** — the value of a GPU-hour three years out. That variable determines 100% of CoreWeave's equity value, 95% of Nebius's and 70% of Oracle OCI's; it determines the repayment of more than $20 billion of GPU-collateralised debt; and it determines the answer to the question of whether it is six years or five.

From October, that variable will have a market price.

This rescues nobody. A futures contract does not lower the price, does not fill an empty GPU, and does not fix a badly built capital structure. The only thing it does is fill, with a common number, a gap that everyone has until now filled with their own assumption.

The history of commodity markets shows that this is not a small thing.

This piece is not investment advice. All scenario and sensitivity calculations presented here are frameworks resting on explicitly stated assumptions; they should not be read as price or revenue forecasts.

Methodology and data note

All price, balance sheet and market data in this study were collected as of 18 August 2026. Three types of figure should be treated separately:

Disclosed data. Companies' quarterly results, credit facility announcements, exchange announcements, and index levels published by market data providers. These are given together with their sources.

Derived calculations. The cost of service and depreciation items in the margin bridge, forward rates, revenue per MW, and the sensitivity and scenario tables. All of these were calculated by the author from disclosed data; the assumptions used are stated in the relevant section and in the chart footnotes. The source code for the models in Sections 7 and 10 has been shared.

Unverifiable data. Silicon Data's 36-month term rate levels come from a subscription portal and could not be publicly verified. The base case in Sections 6, 7 and 10 rests on this data; the reader should treat it as an assumption. That the CME contracts are cash-settled is a conclusion inferred from the structure of the product and from the disclosed characteristics of ICE's equivalent contract; the definitive contract terms in CME's product notice SER-9785 could not be accessed. The default probability implied by CoreWeave's credit default swap premiums is a reading derived on a standard recovery assumption.

Sources

Exchanges and index providers

- CME Group, "CME Group and Silicon Data to Launch Compute Futures on October 5" (11 August 2026) and the initial announcement dated 12 May 2026; product notice SER-9785
- ICE and Ornn, "ICE and Ornn to Launch GPU Compute Futures Contracts" (19 May 2026)
- Silicon Data: GPU Rental Price Indices; "Building a Robust GPU Index"; "H100 Rental Price Over Time (2023–2025)"; "B200 Rental Price: March 2026 Update"; "The GPU Forward Curve"; "GPU Futures: How Compute Is Becoming a Tradable Commodity"
- SemiAnalysis, H100 one-year rental price index

Company disclosures

- CoreWeave: Q2 2026 results and investor presentation; 2025 annual report (10-K); DDTL 4.0 and DDTL 5.5 credit facility announcements; Mike Intrator's quarterly call and CNBC remarks
- Nebius Group: Q2 2026 results and shareholder letter; quarterly call transcript
- Oracle: fiscal year 2026 Q4 results
- Amazon, Microsoft, Alphabet: Q2 2026 results
- Vertiv: Q4 2025 results and 2026 guidance
- NVIDIA: cloud partner programme announcement (2 July 2026); CoreWeave and Nebius share positions (13F filings); $6.3bn capacity purchase commitment with CoreWeave (expiring April 2032)

Market data

- Interactive Brokers: CRWV and NBIS weekly closes and volatility data; NYMEX WTI crude oil futures closes (18 August 2026)
- stockanalysis.com: enterprise value, net debt and share count (18 August 2026)

Analysis and research

- Silicon Analysts, "GPU Rental Premium Decomposition 2026" (10 July 2026)
- Dave Friedman, notes on compute derivatives market structure and basis risk
- Felix Stocker, the history of semiconductor futures initiatives
- Thunder Compute, AI GPU rental market trends (August 2026)
- LBNL "Queued Up", US interconnection queue data
- PJM capacity auction results and Independent Market Monitor reports
- Cleanview, report on electricity projects cancelled in 2025
- TrendForce; Kingston / NAND price statements
- GPUSmith, IntuitionLabs, getdeploying: provider-level GPU rental price comparisons

---

This piece is not investment advice. All scenario and sensitivity calculations presented here are frameworks resting on explicitly stated assumptions; they should not be read as price or revenue forecasts.