

On 17 September 2026 CoreWeave, one of the largest companies renting out AI accelerators, told investors that since the end of June it had signed three-to-six-month contracts at about \$40m a year for each megawatt of power its customers' clusters need. Five days later it closed a \$4.2bn issue of convertible notes. Figures like these sit at the centre of this year's argument over whether the hundreds of billions of dollars going into AI hardware will be earned back. The per-megawatt figures are prices; the notes are borrowing. What the hardware costs depends on three quantities that no price reveals: how much capital was spent, over how many years it is assumed to earn, and how much of the time it is actually at work. A Chicago electricity executive laid out that arithmetic in 1898, and it still sorts the current dispute into what is known and what is merely asserted.

1. A price per megawatt

CoreWeave's prospectus supplement of 17 September 2026, filed for a programme of sales of its Class A shares, contains under "New Contracts and Customer Commitments" a sentence that defines its own number:

Since June 30, 2026, we have signed short-dated customer contracts at pricing of approximately \$40.0 million per megawatt, calculated as annualized revenue divided by power required to service the related clusters, with terms of approximately three to six months.

The figure is revenue, not profit, and it applies only to short contracts signed since the end of June; the same document gives no per-megawatt price for the company's longer contracts, nor the chip generation involved. Five days later, on 22 September, CoreWeave completed an upsized private offering of \$4.2bn of 2.875% convertible senior notes due 2033. CoreWeave's earlier senior notes, listed in the same filings, carry coupons of 8.5% to 9.75%.

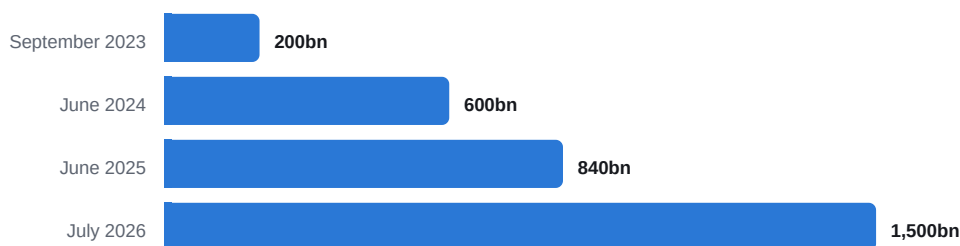
Its peers publish similar numbers. Nebius's letter to shareholders of 12 August 2026 reports that second-quarter deals "saw an average yield of more than \$20 million per megawatt", that four large deals averaged "a yield of \$20-25 million per megawatt", and that the company sees "a price opportunity in the \$40-50 million per MW range". It also gives an expected payback, "for the associated capex and related operating costs", of one year and ten months — down from two to three years — which a footnote describes as based on "forecast costs and contracted future capacity, including capacity not yet built". IREN, in its results of 27 August 2026, cites "Recent 3-year contracts >\$20m revenue per MW (IT), representing a ~2 year payback", where the payback is estimated capital spending on GPUs and ancillaries divided by contracted revenue less estimated direct costs.

Operator	Figure	What it measures, in the company's words	Date
CoreWeave	~\$40.0m per MW	"annualized revenue divided by power required to service the related clusters"; terms of three to six months	17 Sep 2026
Nebius	>\$20m per MW; \$20–25m for four large deals	"yield", shown in the letter as annual contract value per MW, on a revenue-recognition basis excluding prepayments	12 Aug 2026
Nebius	payback of 1 year 10 months	capex and related operating costs of Q2 deals; a forecast including capacity not yet built	12 Aug 2026
IREN	>\$20m revenue per MW (IT); ~2-year payback	three-year contracts; GPU and ancillary capex ÷ (contracted revenue – direct costs)	27 Aug 2026

All four rows are the companies' own figures; none is audited as a statement of cost, and each uses a different definition. The per-megawatt rows are nonetheless the nearest thing the industry publishes to a unit price for AI capacity; the paybacks are forecasts of how quickly that price recovers the investment.

The sums behind the argument are larger still. The International Energy Agency's "Key Questions on Energy and AI" (16 April 2026) says that the capital expenditure of the largest technology companies "exceeded USD 400 billion in 2025 – and is expected to jump by another 75% in 2026". That is total capital spending by a group of companies, not spending on AI alone. The agency adds that "Data centre investments have grown too large to be funded from company balance sheets alone", so that the pace of building "will be sensitive to market sentiment, including expectations for returns on investment in data centres and AI deployment". The best-known sceptical arithmetic comes from David Cahn of Sequoia Capital, who in June 2024 asked "AI's \$600B question" and on 8 July 2026 raised it to "AI's \$1.5T question". His figure is a construct rather than a measured shortfall:

Take Nvidia's projected Q4 run-rate data center revenue x 2 (to reflect total data center CapEx, including non-chip expenses) x 2 (to reflect a 50% margin across the hyperscaler and the AI product company). This analysis arrives at the lifetime end-customer revenue requirement for a single year of CapEx.

Figure 1. David Cahn's 'question': the revenue he calculates AI spending requires, by date of his estimate

Source: David Cahn, 'AI's \$1.5T Question', 8 July 2026, which lists the four estimates; 'AI's \$600B Question', Sequoia Capital, 20 June 2024. Each figure is a required-revenue construct built from NVIDIA's projected data-centre revenue and two doubling assumptions, not a measured loss or revenue gap. The dates are unevenly spaced.

▼ Table view

Figure 1. David Cahn's 'question': the revenue he calculates AI spending requires, by date of his estimate

Date of estimate	Required end-customer revenue (\$bn)
September 2023	200bn
June 2024	600bn
June 2025	840bn
July 2026	1,500bn

Cahn's July update concedes that "With the rise of AI coding, there is a more clear path to monetizing data center CapEx than there was when I first started publishing these analyses." The question has grown; so, in his reading, has the evidence that it can be answered.

2. Two readings that mislead

2.1 "The price shows the cost"

It does not, and the clearest evidence is a price that rose while the thing being priced aged. Silicon Data publishes a daily index of the hourly rental price of NVIDIA's H100 accelerator, drawn from "observations across neo-cloud providers, hyperscalers, colocation markets, and private rental platforms", with separate readings for neoclouds and for the hyperscalers, whose on-demand rates are far higher (\$7.19 on 28 September 2026). By the firm's account its H100 index rose from \$2.00 on 9 December 2025 to \$2.20 on 6 January 2026, a jump it called "startling"; the neocloud reading stood at \$2.72 on 28 September 2026. The chips being priced are the same model, older each month.

Figure 2. Hourly rental price of an NVIDIA H100, as reported by Silicon Data

Sources: Silicon Data, 'H100 Price Spike', blog, January 2026 (readings of 9 December 2025 and 6 January 2026; the blog does not name the segment, and its level is consistent with the neocloud series rather than the hyperscaler one); Silicon Data H100 Rental Price Index page, neocloud reading as of 28 September 2026, read 29 September 2026. The page's dated daily series reads \$2.60 on 22 September and \$2.72 on 28 September; an undated answer in the page's FAQ gives \$2.53. The first two readings are the firm's own description of its H100 index and are not identified as the same series as the third.

▼ Table view

Figure 2. Hourly rental price of an NVIDIA H100, as reported by Silicon Data

Date and series	USD per GPU-hour
9 Dec 2025 (blog; segment not stated)	2\$
6 Jan 2026 (blog; segment not stated)	2.2\$
28 Sep 2026 (neocloud index, SDH100RT)	2.7\$

The sellers explain the rise as demand outrunning supply. Michael Intrator, CoreWeave's chief executive, told investors on 11 August 2026 that "Pricing and margins for our Blackwell and Vera Rubin SKUs are setting new highs while pricing for prior generation SKUs is at or above where it was years ago. Our near-term capacity remains effectively sold out." Nebius attributed its second-quarter pricing to "increasing prices for new-generation GPUs and more than 30% higher pricing on older-generation GPUs versus Q1". The sellers read higher rental prices as a sign of scarce capacity. The prices themselves do not reveal what the chips cost when they were bought, nor whether renting them is profitable, which depends on that cost and on how many hours are sold.

The same distinction runs through the falling price of AI answers. Epoch AI's "The plunging price of thought" (22 September 2026) finds that the price of reaching a given benchmark score has fallen about 47% a quarter since 2023, and offers "One possible explanation" for why the fall is steepest at first: "when a performance level is first achieved, AI companies can briefly charge a premium for it, before competition and technological improvement quickly drive down the price." That explanation concerns pricing power; it does not measure what the answers cost to produce. Epoch's own method makes the same point from the other side: for open models not sold through an API it used "the cost of their rented hardware", which is a cost to the renter but a price to the hardware's owner. And the falling price of a fixed level of performance coexists with rising bills at the frontier: Gundlach and colleagues at MIT FutureTech (arXiv 2511.23455) estimate that "the price of running frontier models is rising between 3× to 18× per year due to bigger models and larger reasoning demands".

2.2 "The dispute is about when chips break"

On 10 November 2025 the investor Michael Burry wrote on X that "Understating depreciation by extending useful life of assets artificially boosts earnings -one of the more common frauds of the modern era", that buying NVIDIA hardware "on a 2-3 yr product cycle should not result in the extension of useful lives of compute equipment", and that "By my estimates they will understate depreciation by \$176 billion 2026-2028." The obvious rejoinder — that old chips keep working and keep renting — arrived quickly. Nine days later, on NVIDIA's results call, its chief financial officer, Colette Kress, said without naming him that "the A100 GPUs we shipped six years ago are still running at full utilization today". In August 2026 CoreWeave's finance chief said it had "recently signed an A100 contract that extends into 2029".

Burry's reply, on 9 July 2026, reframed the dispute:

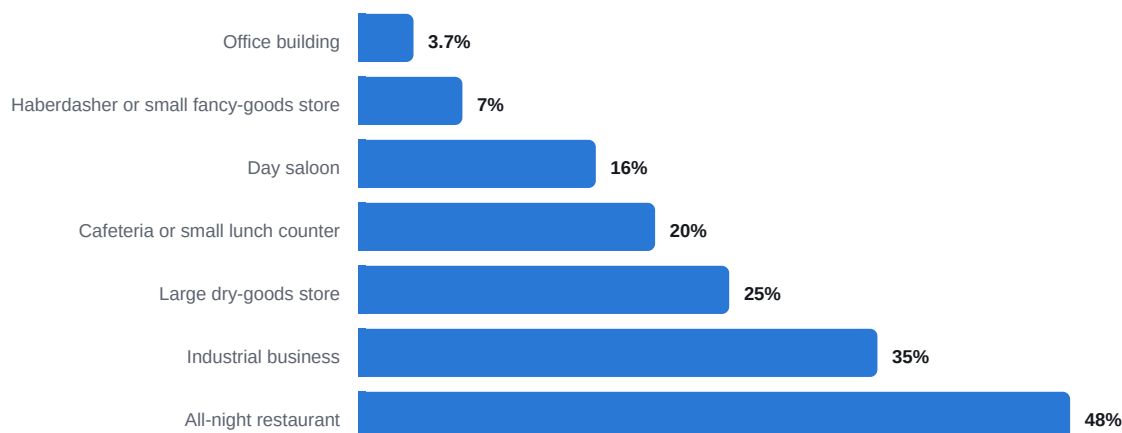
Depreciation is not a bet on when a chip stops working.

"A chip can rent and still depreciate very fast economically," he went on. "I never said the A100s would stop functioning." The \$176bn is his own estimate. On his account, the disagreement is not about hardware reliability. It is about how quickly a chip's earning power falls once better chips exist, and that is a question about cost.

3. The idea: capital, life and hours of use

3.1 Insull and the load factor

The arithmetic was worked out for electricity. On 17 May 1898 Samuel Insull, president of the Chicago Edison Company and formerly Thomas Edison's private secretary, lectured at Purdue University on "The Development of the Central Station". A power station must be built for its maximum load, so the capital sits in the plant whether or not anyone is using it; as Insull put it, "If your maximum is very high and your average consumption very low, heavy interest charges will necessarily follow." Insull's measure was the load factor — average load as a share of maximum demand — and he set it out for seven classes of customer, from an office building at about 3.7% (the investment "in use the equivalent of a little over 323 hours a year") to an all-night restaurant at 48%.

Figure 3. Insull's illustrative load factors by class of customer, 1898

Source: Samuel Insull, 'The Development of the Central Station', lecture at Purdue University, 17 May 1898, printed in *Central-Station Electric Service* (Chicago, privately printed, 1915), pages 26-27. Insull's own illustrative figures for 'seven different classes of business commonly taken by electric-light-and-power companies in any large city'.

▼ Table view

Figure 3. Insull's illustrative load factors by class of customer, 1898

Class of customer	Load factor (average load as a share of maximum demand)
Office building	3.7%
Haberdasher or small fancy-goods store	7%
Day saloon	16%
Cafeteria or small lunch counter	20%
Large dry-goods store	25%
Industrial business	35%
All-night restaurant	48%

This question of load factor is by all means the most important one in central-station economy.

Selling "at cost", Insull said, meant charging "an amount sufficient to cover his operating, repairs, and renewals, general expense, and interest and depreciation" — a definition any modern analyst of AI infrastructure could use unchanged. On that basis the customer with the poorest load factor would have to pay "over four times as much per unit" as the customer with the best, and Insull thought it unjust to charge them the same:

It does not seem to be just that a man who only uses your investment, say, 100 hours a year should be able to buy your product at precisely the same price as the man who uses your investment, say, 3,000 hours a year, when the amount of money invested to take care of either customer is precisely the same.

His reason was that "interest is the largest factor in cost". The remedies he listed are recognisable: some companies offered discounts to heavy users, and some used "the two-rate scheme, charging one rate for electricity used during certain hours of the day and a lower rate for electricity used during the remainder of the day". Seventeen years earlier, writing from Edison's laboratory at Menlo Park in 1881, Insull had put the same economics in one line about Edison's plan to sell power by day as well as light by night: "his plant is never idle, his capital is never running to waste".

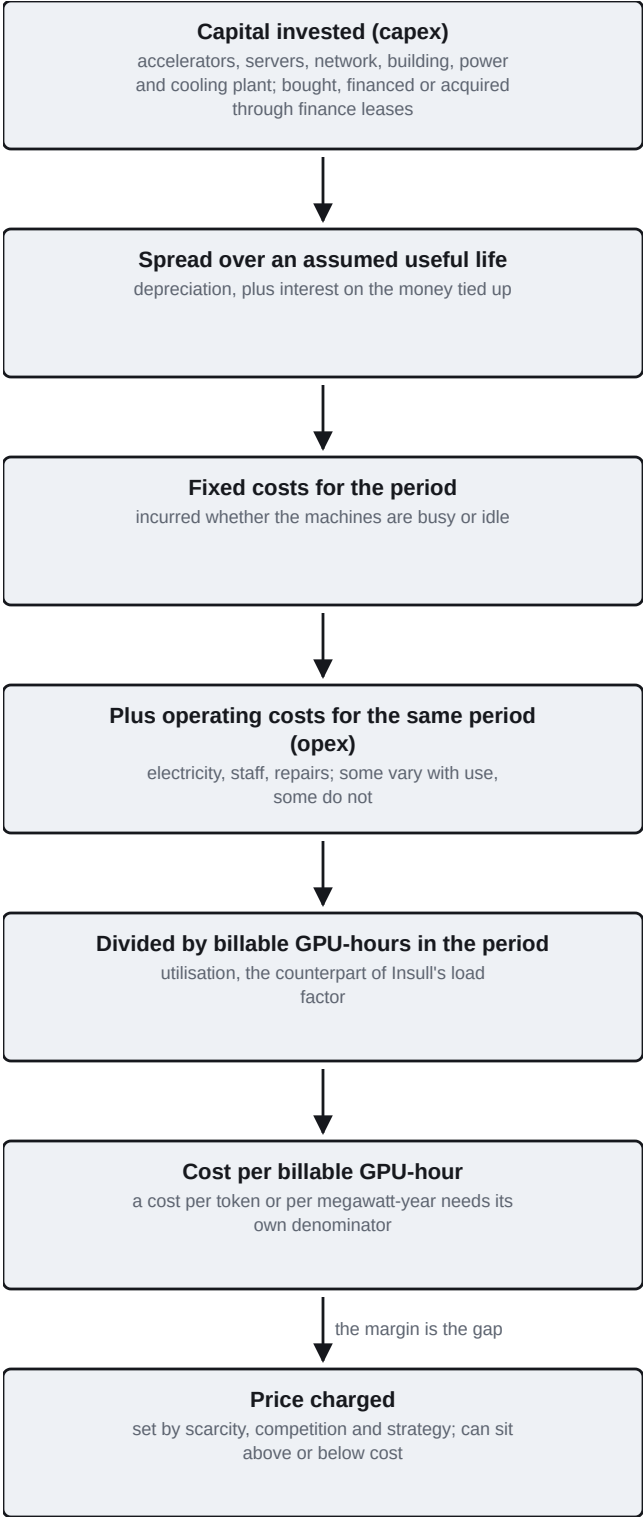
3.2 From a power station to a room of accelerators

The translation to AI hardware is direct. Capital expenditure, or capex, is investment in long-lived assets — accelerators, servers, networking, buildings, and the electrical and cooling plant — which may be paid for in cash, financed, or acquired through finance leases. Operating expenditure, or opex, covers what is spent to run them: electricity, staff, repairs, rent. Depreciation converts capex into an annual charge by spreading the purchase over the years the equipment is expected to earn, its useful life. Interest, or the return that shareholders forgo, is the cost of the money tied up in the meantime. Utilisation — the share of a fleet's capacity doing paid work over a period — plays the part that load factor played for Insull. The two are not identical: load factor compares average demand with peak demand, and a machine can run all day at part load. Both measure how hard an investment sized for the peak is actually worked.

Much of the confusion in the current dispute comes from treating different clocks as one. They run at different speeds and answer different questions.

Clock	What it measures	Example from the filings
Physical life	How long the hardware keeps working	NVIDIA's A100s "shipped six years ago", still in service (company's claim, November 2025)
Economic life	How long it earns more than it costs to run	The subject of Burry's objection; no filing measures it directly
Accounting life	The period over which the purchase is depreciated	Five to six years for servers at the large buyers
Contract term	How long a customer has agreed to pay	Three to six months (CoreWeave's short contracts) to one to six years (its committed contracts)
Debt maturity	When the money borrowed must be repaid	CoreWeave's convertible notes, due 2033
Payback	When a defined stream of receipts covers a defined outlay	About two years, as forecast by Nebius and IREN on their own definitions

Figure 4. What an hour of AI computing costs, and why the price is a separate number



Schematic, for one period of account. Every box above the price involves estimates the operator makes; only the price is observed by the customer. A change of assumed life moves the second box; idle hours move the fifth.

▼ Table view

Figure 4. What an hour of AI computing costs, and why the price is a separate number — stages

#	Stage	Note
---	-------	------

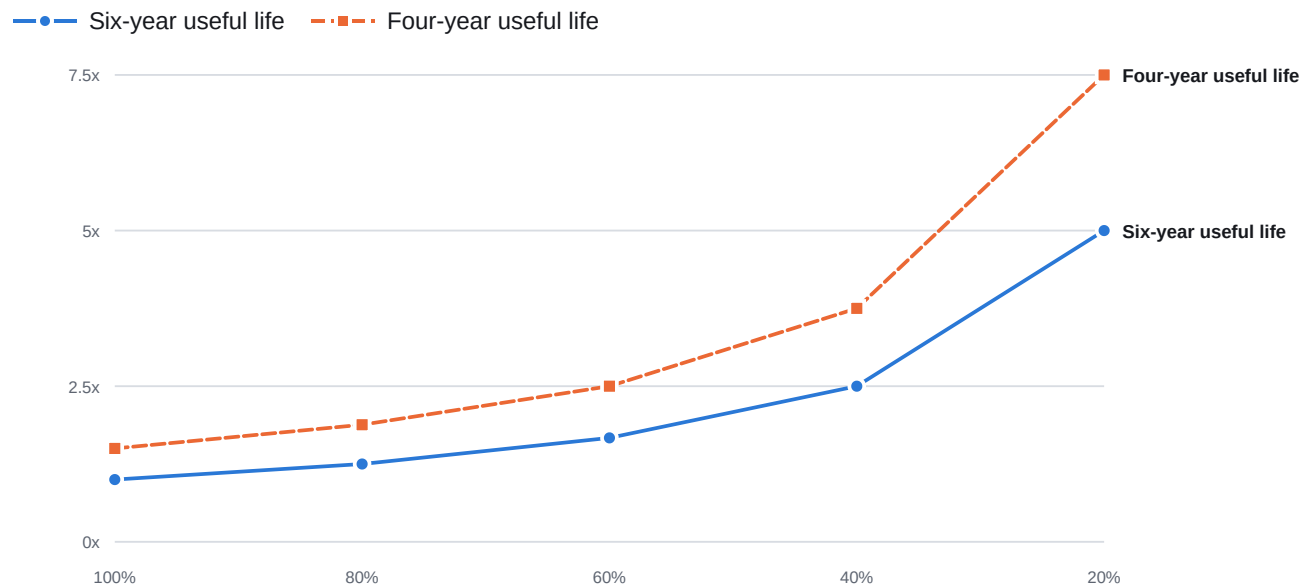
#	Stage	Note
1	Capital invested (capex)	accelerators, servers, network, building, power and cooling plant; bought, financed or acquired through finance leases
2	Spread over an assumed useful life	depreciation, plus interest on the money tied up
3	Fixed costs for the period	incurred whether the machines are busy or idle
4	Plus operating costs for the same period (opex)	electricity, staff, repairs; some vary with use, some do not
5	Divided by billable GPU-hours in the period	utilisation, the counterpart of Insull's load factor
6	Cost per billable GPU-hour	a cost per token or per megawatt-year needs its own denominator
7	Price charged	set by scarcity, competition and strategy; can sit above or below cost

Figure 4. What an hour of AI computing costs, and why the price is a separate number — connections

From	To	Label
Capital invested (capex)	Spread over an assumed useful life	
Spread over an assumed useful life	Fixed costs for the period	
Fixed costs for the period	Plus operating costs for the same period (opex)	
Plus operating costs for the same period (opex)	Divided by billable GPU-hours in the period	
Divided by billable GPU-hours in the period	Cost per billable GPU-hour	
Cost per billable GPU-hour	Price charged	the margin is the gap

Two of those estimates move the answer a great deal. Straight-line depreciation divides the purchase evenly across the useful life, so assuming six years rather than four cuts each year's charge by a third; assuming four rather than six raises it by half. And because the fixed charge accrues by the hour whether or not the machines are busy, halving utilisation doubles the fixed cost carried by each busy hour. The combined effect is shown below in index form, which requires no assumption about what any particular accelerator cost.

Figure 5. Fixed charge carried by each busy hour, by utilisation and assumed useful life (index: six-year life at full use = 1.0)



Illustrative arithmetic, not any company's figures: straight-line depreciation with no residual value; interest, electricity and other running costs excluded. The index is $(6 \div \text{useful life in years}) \div \text{utilisation}$. A four-year life at 40% use carries 3.75 times the fixed charge per busy hour of a six-year life at full use.

▼ Table view

Figure 5. Fixed charge carried by each busy hour, by utilisation and assumed useful life (index: six-year life at full use = 1.0)

Utilisation (share of capacity-hours doing paid work)	Six-year useful life	Four-year useful life
100%	1x	1.5x
80%	1.2x	1.9x
60%	1.7x	2.5x
40%	2.5x	3.8x
20%	5x	7.5x

None of this sets the price. Insull's customers paid what his tariffs said; an AI customer pays what the market will bear, and scarcity moves that independently of cost.

3.3 Where scarcity comes from

The inputs to AI hardware are themselves scarce. The IEA's April update says that "a shortage of high-bandwidth memory – integral to AI chip production – has developed over the past six months and is anticipated to persist through at least the end of 2027". NVIDIA's quarterly report for the period to 26 July 2026 records that it raised its supply commitments "from \$ 119 billion last quarter to \$ 279 billion", commitments that are "primarily memory and manufacturing facilities". Microsoft's chief financial officer, Amy Hood, told investors on 29 July 2026 that its \$41bn of quarterly capital expenditure included "the impact from higher component pricing". The electrical equipment is scarce too: Wood Mackenzie reported in October 2025 that lead times for power transformers had eased to an average of 128 weeks.

Scarcity therefore works on both sides of the ledger at once, but not symmetrically. Dear memory raises what a new data centre costs to build; scarce capacity raises what an existing one can charge. A company whose chips were bought before the squeeze can price at today's rate on yesterday's cost; one buying now pays today's cost and depends on tomorrow's price. That asymmetry, rather than any single number, is what makes a price per megawatt so hard to read.

4. What happened: three numbers, 2023–2026

4.1 The useful life

The large buyers' assumed lives have moved in both directions, and each change is disclosed with its effect. The lengthenings Burry objected to were mostly made in 2022 and 2023, when Microsoft and Alphabet each moved servers from four years to six.

Company	Servers and network equipment	Change, and its stated effect	Source
Microsoft (earlier change)	6 years, from 4	servers and network equipment lengthened, fiscal 2023 (from July 2022): \$3.7bn more operating income, \$3.0bn more net income that year	10-K for fiscal 2023, 27 Jul 2023
Alphabet (earlier change)	6 years, from 4	servers lengthened (certain network equipment from 5), January 2023: \$3.9bn less depreciation, \$3.0bn more net income that year	10-K for 2023, 31 Jan 2024
CoreWeave	6 years ("Technology equipment")	lengthened from 5 to 6 years, 1 January 2023	10-K for 2025, 2 Mar 2026
Amazon	5 to 6 years	5 → 6 years, January 2024; a subset 6 → 5 years, 1 January 2025: \$1.4bn more depreciation, \$1.0bn less net income in 2025	10-K for 2025, 6 Feb 2026
Meta	5 to 5.5 years	most lengthened to 5.5 years, 1 January 2025: \$2.92bn less depreciation, \$2.59bn more net income in 2025	10-K for 2025, 29 Jan 2026
Alphabet	"generally" 6 years	none found in 2025–26	10-K for 2025, 5 Feb 2026
Oracle	6 years	none found	10-K for fiscal 2026, 22 Jun 2026
Microsoft	2 to 6 years	buildings lengthened from 15 to 25 years from fiscal 2027	10-K and earnings call, 29 Jul 2026
Nebius	5 years	lengthened from 4 to 5 years from the first quarter of 2026	Form 6-K, 12 Aug 2026

The most instructive pair took effect on the same day. On 1 January 2025 Meta lengthened the life of most of its servers and network assets to 5.5 years, while Amazon shortened a subset of its servers and networking equipment from six years to five and gave its reason:

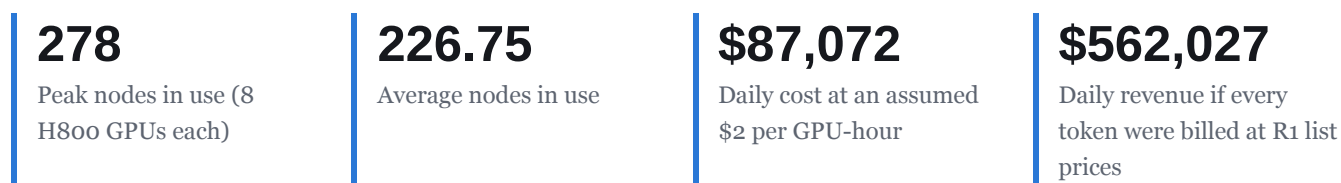
The shorter useful lives are due to the increased pace of technology development, particularly in the area of artificial intelligence and machine learning.

Burry cited the pair in July as "Same hardware but opposite conclusions by two of the biggest." The two companies' fleets and workloads differ, so the filings do not show that either estimate is wrong; they show that the estimate is a judgement, and a large one. Alphabet describes its six-year life as one it regularly evaluates "for factors such as technological obsolescence and our planned use and utilization" — utilisation entering the accounting policy itself. Nebius attributed its lengthening to "usage patterns and current utilization commitments". Microsoft's change was to buildings rather than servers, and Hood said it "affects only the timing of future depreciation and is expected to have a minimal benefit to FY27 operating income"; its larger effect was on the reported capital spending, because more future data-centre leases will be classed as operating rather than finance leases, lowering Microsoft's expectation for calendar 2026 to "approximately \$175 billion".

4.2 Utilisation, and a rare look at cost against price

Providers rarely publish their serving costs; DeepSeek is an exception. On 1 March 2025 it released an overview of its inference system with statistics for the 24 hours to midday on 28 February (Beijing time). Its load was high by day and low at night, so it ran inference on all its nodes at the daytime peak and, "During low-load nighttime periods, we reduce inference nodes and allocate resources to research and training." Occupancy peaked at 278 nodes of eight H800 GPUs and averaged 226.75.

Figure 6. DeepSeek's own account of one day of serving, 27-28 February 2025



Source: DeepSeek, 'Day 6: One More Thing, DeepSeek-V3/R1 Inference System Overview', GitHub (deepseek-ai/open-infra-index), 1 March 2025, section 'Statistics of DeepSeek's Online Service'. The cost is GPU-hours at an assumed leasing price and nothing else; the revenue is hypothetical. DeepSeek's own figures, unaudited.

▼ Table view

Figure 6. DeepSeek's own account of one day of serving, 27-28 February 2025

Measure	Value
Peak nodes in use (8 H800 GPUs each)	278
Average nodes in use	226.75
Daily cost at an assumed \$2 per GPU-hour	\$87,072
Daily revenue if every token were billed at R1 list prices	\$562,027

Valuing each GPU at a leasing price of \$2 an hour, DeepSeek put the day's cost at \$87,072. Had every token been billed at the list prices of its R1 reasoning model, revenue would have been \$562,027, "with a cost profit margin of 545%" — profit expressed as a multiple of cost, not of revenue. The company then qualified the number itself:

However, our actual revenue is substantially lower for the following reasons:

Its V3 model was priced well below R1; "web and APP access remain free"; and "Nighttime discounts are automatically applied during off-peak hours". The cost side was equally partial: GPU-hours at an assumed rental price, with nothing for staff, research or the training of the models. The \$2 rate is itself a rental price, which includes whatever capital recovery and margin a hardware owner builds into it, and the overview does not say whether DeepSeek owns or rents its chips. The disclosure shows the gap between list price and the hourly cost of a busy fleet; it does not show a profit.

4.3 A fixed-cost business, in one income statement

CoreWeave's accounts show what the arithmetic looks like at scale. In 2025 the company reported revenue of \$5,131m, depreciation and amortisation of \$2,454m, net interest expense of \$1,229m, an operating loss of \$46m and a net loss of \$1,167m. Depreciation and interest together came to about 72% of revenue: Insull's "interest and depreciation", at almost three-quarters of what an AI infrastructure company takes in.

Figure 7. CoreWeave, year to 31 December 2025



Source: CoreWeave, Form 10-K for 2025, filed 2 March 2026 (consolidated statements of operations; depreciation and amortisation from the statement of cash flows). The company's committed contracts ran for a weighted average of about five years at the end of 2025.

▼ Table view

Figure 7. CoreWeave, year to 31 December 2025

Measure	Value
Revenue	\$5,131m
Depreciation and amortisation	\$2,454m
Net interest expense	\$1,229m
Net loss	\$1,167m

The same period brought the buyers' capital plans to a new scale. Meta's second-quarter release of 29 July 2026 guides to 2026 capital expenditure, including principal payments on finance leases, of \$130bn–145bn, narrowed from \$125bn–145bn. Microsoft reported \$41bn for its fourth fiscal quarter alone, of which "Roughly two thirds" went on "short-lived assets, primarily CPUs and GPUs" serving "both AI and non-AI infrastructure" — a reminder that capital expenditure is not the same thing as spending on AI.

5. Two ways to carry the idle hours

Every operator of expensive hardware has to decide who bears the cost of the hours it is idle. Two postures are visible in the documents.

The first is to sell the hours in advance. CoreWeave's annual report describes its committed contracts, typically one to six years long, as

requiring payment regardless of the level of utilization.

For the life of such a contract the customer carries the risk of idle hours, and customers "often make a prepayment". The operator keeps the cost of the money — CoreWeave's net interest expense was \$1.2bn in 2025 — and faces the utilisation question again at renewal, when a chip three or four years old must find a new tenant at whatever the market then pays. The two-year paybacks that Nebius and IREN forecast rest on the first contracts recovering the capital before that moment. The evidence on second contracts is so far anecdotal and comes from the sellers: CoreWeave's A100 contract into 2029, and NVIDIA's claim of full utilisation for six-year-old chips.

A variant moves the risk further up the supply chain. NVIDIA's quarterly report for the period to 26 July 2026 discloses guarantees, "capped at a total of \$ 105 billion", to provide credit support on leases for about 4.25 gigawatts of IT load at a campus in Pike County, Ohio, on behalf of an affiliate of OpenAI. The obligation is triggered only by "certain tenant defaults"; each guarantee generally becomes effective as the applicable lease begins, with the first phase expected in fiscal 2029; the amount is expected to decrease over each phase's 20-year lease term; and the guarantees are "limited to defined portions of lease and power payments and not the full cost of the site". The supplier of the chips has, in a bounded way, taken on part of its customer's rent risk.

The second posture is Insull's: keep the idle hours and try to fill them. DeepSeek's March 2025 disclosure described moving night-time capacity to research and training. Its current price list, read on 29 September 2026, defines peak hours as "01:00 - 04:00 and 06:00 - 10:00 UTC, Monday through Friday, excluding Chinese public holidays" — nine to noon and two to six in Beijing — and states that "Off-peak rates are half of the peak rates." It is the two-rate scheme Insull described in 1898, applied to tokens. A lower price in quiet hours is not in itself a loss: if it covers the running cost of machines that would otherwise stand idle, it reduces the fixed charge every other hour has to carry.

	Sell the hours in advance	Fill the idle hours
Example	CoreWeave's take-or-pay contracts; Nebius's and IREN's multi-year deals	DeepSeek's off-peak half price and night-time reallocation
Who carries idle hours	The customer, for the contract term	The operator
What the operator still carries	Financing cost; renewal price for ageing chips	Demand risk every hour
What it depends on	Contracts recovering capital before the chips age	Price-sensitive demand for off-peak capacity
Historical parallel	Insull's discounts to long-hours customers	Insull's two-rate scheme

6. What to watch

- **Late October 2026, if last year's timing holds: the hyperscalers' third-quarter results.** In 2025 Microsoft, Alphabet, Amazon and Meta all reported their third-quarter results on 29 or 30 October. Meta's guidance of \$130bn–145bn of 2026 capital expenditure is the one range published in a company's own release; any change to it, and any change to an assumed server life, will be in these reports. Microsoft's first quarterly report of fiscal 2027 will also be its first filing since the longer building lives took effect.
- **November 2026, if last year's timing holds: NVIDIA's third-quarter results.** NVIDIA reported the equivalent quarter of 2025 on 19 November 2025. Its supply commitments stood at \$279bn on 26 July 2026, up from \$119bn a quarter earlier; the next figure will show whether they are still growing.
- **Continuously: the H100 rental index.** At \$2.72 on 28 September 2026, Silicon Data's reading is the most direct public test of the depreciation dispute. If older accelerators keep their rental prices as Blackwell and Rubin systems arrive in volume, the longer lives look defensible; if the index falls steeply, Burry's reading of "very fast" economic depreciation gains ground.

7. The idea to keep

For an expensive machine, cost is the purchase price spread over its useful life, plus interest and running costs, divided by the hours it actually works. Price is what a buyer will pay today, and scarcity can push it well above that cost or competition well below it. A price can cover the cost of running a machine without ever repaying the cost of buying it. Any price for AI — per token, per GPU-hour or per megawatt — therefore leaves three questions open: over how many years the hardware is assumed to earn, how much of the time it is busy, and who carries the idle hours. Samuel Insull put the second of those at the centre of the electricity business in 1898; the best short modern illustration of all three together is DeepSeek's "Day 6" inference-system overview of 1 March 2025.

Sources

Source	Date	What it supports
CoreWeave, prospectus supplement (Form 424B5)	17 Sep 2026	~\$40.0m per MW on three-to-six-month contracts
CoreWeave, Form 8-K (convertible notes)	22 Sep 2026	\$4.2bn of 2.875% notes due 2033; earlier notes' coupons
CoreWeave, Form 10-K for 2025	2 Mar 2026	six-year life; take-or-pay; 2025 revenue, depreciation, interest, loss
CoreWeave, second-quarter 2026 earnings call (corrected transcript)	11 Aug 2026	pricing of prior-generation chips; A100 contract into 2029

Source	Date	What it supports
Nebius Group, Q2 2026 letter to shareholders; operating and financial review (Form 6-K)	12 Aug 2026	yield and price per MW; payback; life lengthened to five years
IREN, FY26 results release	27 Aug 2026	>\$20m revenue per MW (IT); ~2-year payback
International Energy Agency, <i>Key Questions on Energy and AI</i> , executive summary	16 Apr 2026	capex >\$400bn and +75%; memory shortage; funding and returns
David Cahn, "AI's \$600B Question" (Sequoia Capital); "AI's \$1.5T Question"	20 Jun 2024; 8 Jul 2026	required-revenue construct and its history
Silicon Data, H100 Rental Price Index; "H100 Price Spike"	28 Sep 2026; Jan 2026	\$2.00, \$2.20, \$2.72 per GPU-hour
Epoch AI, "The plunging price of thought"	22 Sep 2026	price of a fixed performance level; premium explanation
Gundlach, Lynch, Mertens, Thompson, "The Price of Progress", arXiv 2511.23455 v2	23 Mar 2026	frontier running prices rising 3–18× a year
Michael Burry, post on X; "Short Thoughts July 8, 2026" (Cassandra Unchained)	10 Nov 2025; 9 Jul 2026	depreciation critique and its reframing
NVIDIA, Q3 fiscal 2026 earnings call; Form 10-Q for the quarter to 26 Jul 2026	19 Nov 2025; 26 Aug 2026	A100 utilisation claim; supply commitments; lease guarantees
Samuel Insull, "The Development of the Central Station" (Purdue University), in <i>Central-Station Electric Service</i>	17 May 1898 (printed 1915)	load factor; selling at cost; two-rate scheme; 1881 letter
Amazon, Form 10-K for 2025	6 Feb 2026	server lives and the effect of shortening them
Meta Platforms, Form 10-K for 2025; Q2 2026 results release	29 Jan 2026; 29 Jul 2026	5.5-year life and its effect; 2026 capex guidance
Alphabet, Form 10-K for 2025	5 Feb 2026	six-year life, evaluated for utilisation
Oracle, Form 10-K for fiscal 2026	22 Jun 2026	six-year life
Microsoft, Form 10-K for fiscal 2026; fourth-quarter earnings call	29 Jul 2026	server lives; building lives; capex composition and outlook
DeepSeek, "Day 6: One More Thing, DeepSeek-V3/R1 Inference System Overview"; API pricing page	1 Mar 2025; read 29 Sep 2026	one day of cost, theoretical revenue and utilisation; off-peak pricing
Wood Mackenzie, "Mind the gap"	15 Oct 2025	power-transformer lead times