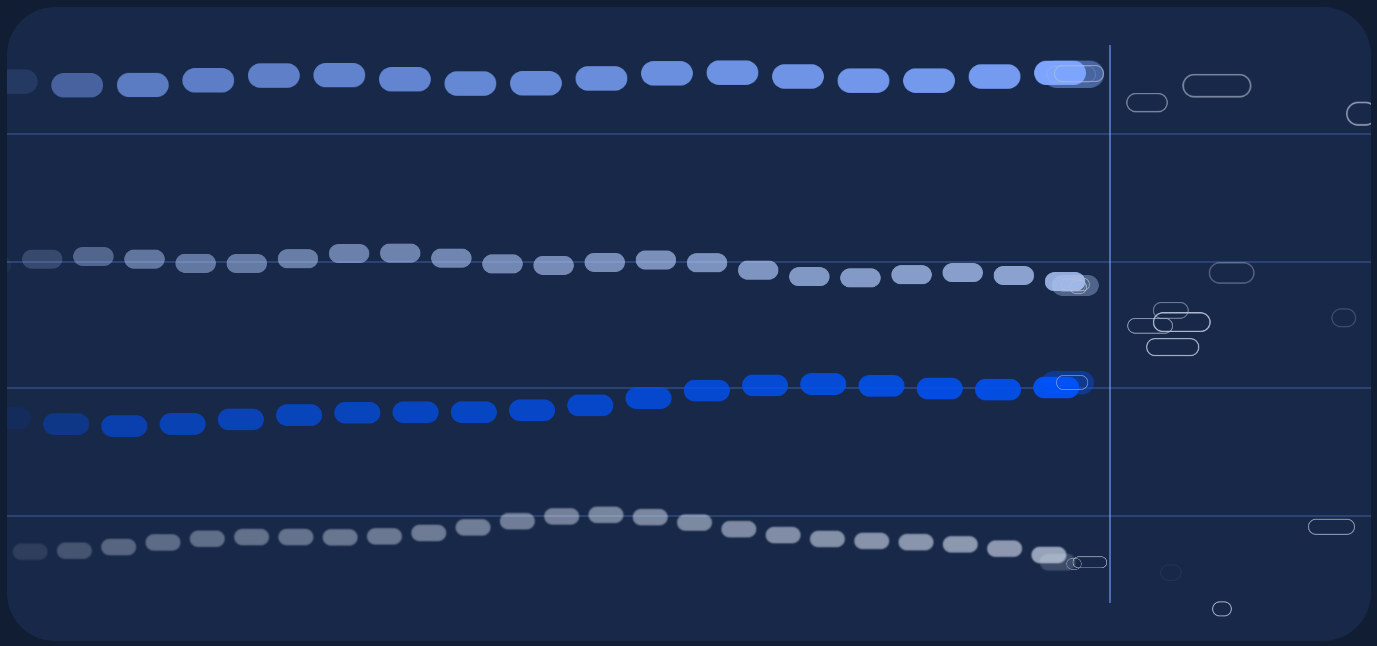
Rillor

METHODS NOTE

Rent or buy GPUs: a break-even method from public inputs

Owning GPUs costs less than renting on demand only above a break-even utilization, and against committed terms the test is cost per reserved hour, built from dated public inputs.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- On dated public inputs, one HGX B200 GPU costs about $2.43 per hour of the year to own, or $2.84 with NVIDIA's software subscription. The inputs are an October 2026 server listing with network cards, 9.0% over five years, NVIDIA's power figure, a PUE of 1.54, August 2026 U.S. industrial electricity and CBRE's colocation rate. Staff, cluster networking, storage, insurance and tax are excluded [15, 16, 18, 24, 25, 27, 29].
- Against on-demand and index B200 rates on 24 October 2026, owning costs less above 17% to 44% utilization, and above 57% to 67% against CoreWeave's spot rate [3, 8, 9, 36].
- Against committed terms both sides pay for every hour, so the test is cost per reserved hour. Committed B200 rates run from about $2.69 per GPU-hour, derived from a 24-month contract in a filing, to $14.21 for an AWS Capacity Block [1, 13].
- The rental option, by provider, term and pricing type, moves the answer more than any ownership input. One input moved across its sourced range shifts owning cost by no more than $0.47 per GPU-hour, and all of them moved together span $1.86 to $3.57. H100 rates alone span $1.40 to $12.29 [1, 3, 4, 6, 9, 12, 13].
- A longer commitment is not always cheaper per hour. AWS prices its B200 and H200 Capacity Blocks within 0.3% of on-demand, and Lambda's clusters, which add an InfiniBand fabric and management nodes, cost more per GPU-hour than its 8-GPU instances [1, 3, 9, 10].

Background

A team that needs GPU compute can rent by the hour, rent spot capacity that the provider can reclaim, commit to a term of weeks to years, or buy servers and run them in a colocation facility. The rent-or-buy question asks which option costs less for the hours the team expects to use.

The inputs move. SemiAnalysis reported that one-year H100 contract prices rose almost 40%, from a low of $1.70 per GPU-hour in October 2025 to $2.35 by March 2026 [12]. AWS reprices Capacity Blocks on a schedule, next in January 2027 [1]. U.S. industrial electricity averaged 9.64 cents per kWh in August 2026, against 8.98 cents a year earlier [24]. CBRE's average asking rate for 250 to 500 kW of colocation in primary North American markets rose 6.6% in 2025 to $196.25 per kW-month, and a further 4.3% in the first half of 2026 [27, 28]. A break-even computed once goes stale. A method with dated inputs can be rerun.

This report takes the buyer's view: what an hour of capacity costs under each option. The rent-to-value note takes an owner's view of payback, and this report does not repeat it.

Method

What was examined. The cost of owning one GPU per hour, built line by line, against rental prices for the same GPU class by provider, term and pricing type. The main case is the HGX B200, for which system listings, power figures and rental prices are all public. A second case uses the HGX H200.

Sources. Provider price pages and price files [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], rental indices [11, 12, 36], SEC filings [13, 14, 15, 21, 22, 23], NVIDIA documents [16, 17, 29], server configurators [18, 19, 20], federal statistics and guidance [24, 32, 33, 34], industry research [25, 26, 27, 28] and two reliability studies [30, 31], all viewed on 24 October 2026.

Evidence classes. A listing is an asking price a seller publishes, including rental prices on provider pages and APIs, and a reported sale is a completed sale, as reported by a party to it [35]. The classes stay separate in calculation [35]. This report labels contracts and purchases reported by a party to them, such as WhiteFiber's service orders, IREN's purchases and Exascale Labs' agreement, as reported contracts and reported purchases. Listings and reported transactions are never averaged together.

How. Prices are per GPU-hour in U.S. dollars, with 8,760 hours in a year, the convention IREN uses for annualized revenue [15]. Derived figures are labeled as our arithmetic. The analysis is pre-tax and uses no Rillor prices, quotes or data.

Definitions

- **Capital recovery factor (CRF).** "The ratio of a constant annuity to the present value of receiving that annuity for a given length of time" [34]. For an interest rate i over n years, CRF = i(1 + i)^n ÷ ((1 + i)^n - 1) [34].
- **Utilization (u).** The share of the year's 8,760 hours in which the GPU does work the buyer would otherwise rent.
- **Owning cost per used hour.** (P × CRF + F) ÷ (8,760 × u) + V, where P is the price per GPU, F the fixed yearly costs (colocation, software) and V the variable cost per used hour (electricity). This is the simple levelized cost of energy with utilization in place of the capacity factor [34]. C is its value at u = 1.
- **Break-even utilization** against a pay-as-you-go rate R is (P × CRF + F) ÷ (8,760 × (R - V)). With electricity treated as fixed, it is C ÷ R.
- **Cost per reserved hour.** A committed rental's rate, paid whether the hour is used or not, compared with C for ownership.

Like the energy model, which leaves out "future replacement or degradation costs" [34], the method assumes no resale value and constant prices over the life.

Inputs and their sources

Input	Base value	Source type	Sources
B200 server price per GPU	$57,225	Listing	[18]
Life	5 years	Filings	[13, 14]
Rate	9.0%	Reported financing	[15]
IT power per GPU	1.79 kW	Specification	[16]
PUE	1.54	Survey	[25]
Electricity	9.64 cents per kWh	Federal statistic	[24]
Colocation	$196.25 per kW-month	Survey of asking rates	[27]
Software	none in the base case	Listing	[29]

Finding 8 gives the sourced range for each input.

Findings

1. Hardware price evidence, with scope and evidence class

Evidence	Class and date	Price	Per GPU	Scope
Exxact HGX B200, Essential build	Listing, 24 Oct 2026	$443,278	$55,410	2 AMD EPYC 9554, 1,536 GB memory, no network card, 3-year warranty [18]
Same, with 8 ConnectX-7 400G cards at $1,815	Listing, 24 Oct 2026	$457,798	$57,225	Base case [18]
Thinkmate GXS6 HGX B200 ($504,905), with 8 NDR400 cards at $2,211	Listing, 24 Oct 2026	$522,593	$65,324	1 TB memory, 5-year parts replacement [19]
Thinkmate HGX H200 models, starting prices	Listing, 24 Oct 2026	$407,660 to $456,253	$50,958 to $57,032	Starting build of each model [20]
IREN, about 2,400 B200 and B300 GPUs	Reported purchase, 3 Jul 2025	about $130 million	about $54,200	Servers, storage, labor, InfiniBand, racks and cabling [21]
Exascale Labs, 96 B300 GPUs	Reported contract, Oct 2026	$8,208,000	$85,500	Network, management servers, deployment and 36 months of IT operations [23]

Per-GPU figures are our arithmetic.

The listings differ in scope. Exxact's base build selects no network card [18]. Thinkmate's B200 build has 1 TB of memory, under the 1.5 TB minimum the same page says NVIDIA recommends, and its prices exclude shipping, handling and taxes [19]. Neither configurator offers network switches, network cables, racks or installation, so neither configured price includes them [18, 19].

Reported deals bundle more. IREN's July 2025 figure covers fit-out as well as servers, for about 2,400 GPUs of two types, and still comes in below the single-server listings [21]. Exascale Labs' price includes three years of operations, with no split from the equipment [23]. IREN's November 2025 purchase from Dell, about $5.8 billion, bundles deployment services, software and licensing and gives no GPU count, so it yields no per-GPU price [22].

Thinkmate's server list shows HGX H200 systems and no HGX H100 system [20], so the second case uses the H200.

2. Capital cost: life, tax and financing

Life and rate	CRF	Yearly charge on $57,225	Per GPU-hour at full use
5 years at 9.0%	0.2571	$14,712	$1.68
5 years at 6.0%	0.2374	$13,585	$1.55
6 years at 9.0%	0.2229	$12,757	$1.46
6 years at 6.0%	0.2034	$11,637	$1.33

Sources: [13, 14, 15, 18, 34]. The arithmetic is ours.

Book life. WhiteFiber depreciates cloud service equipment over five years [13]. CoreWeave depreciates technology equipment over six years, having changed its estimate for computing equipment in data centers from five to six years from 1 January 2023 [14]. The base case uses five.

Tax life. IRS Publication 946 puts "any qualified technological equipment" in the 5-year property class and lists computers and peripheral equipment at five years under the Alternative Depreciation System [32]. It reports that P.L. 119-21 reinstated the 100% special depreciation allowance for certain qualified property acquired and placed in service after 19 January 2025 [32]. Tax timing changes the after-tax cost of owning. This report works before tax and gives no tax advice.

Financing. IREN reported investment-grade GPU financing at 6.0%, a weighted average excluding fees, and a $2.4 billion financing at a 9.0% fixed rate for non-investment-grade customer deployments that funds 90% of the related GPU capital spending [15]. These are an operator's rates against contracted revenue. A buyer's own cost of capital replaces them in the formula.

The CRF assumes nothing is left at the end of the life. A buyer who expects to resell can subtract the present value of the resale price from P. Resale values are not public in these sources.

3. Power: system draw, PUE and electricity prices

NVIDIA lists the DGX B200 at about 14.3 kW maximum for eight GPUs [16], or 1.79 kW per GPU, and the DGX H100/H200 at 10.2 kW maximum [17], or 1.28 kW per GPU. Third-party B200 builds are close: 14,147 W at Exxact [18] and an estimated 14,382 W at Thinkmate [19]. IREN plans its Blackwell capacity at 1.93 kW per GPU including ancillary draw, with a PUE of 1.1 [21].

PUE is total facility power divided by IT power [25]. Uptime Institute's 2025 survey found a weighted average of 1.54, "marking the sixth consecutive year that this headline figure has virtually stood still" [25]. Facilities commissioned within five years averaged 1.48, sites of 20 MW and above 1.44, and 15% of respondents reported 1.3 or better [25].

Tariff, August 2026, preliminary	Cents per kWh	Per GPU-hour	Per GPU-year
Texas, industrial	7.21	$0.20	$1,739
U.S., industrial	9.64	$0.27	$2,325
Virginia, industrial	9.98	$0.27	$2,407
Illinois, industrial	11.33	$0.31	$2,732
U.S., commercial	14.47	$0.40	$3,489
California, industrial	24.12	$0.66	$5,816

Sources: [16, 24, 25]. Each line is 1.79 kW × PUE 1.54 × the price, at full draw every hour.

At the base inputs electricity is about 11% of owning cost. With IREN's design figures of 1.93 kW and a PUE of 1.1, it would cost about $0.20 per GPU-hour at 9.64 cents [21, 24].

Two choices here lean high. Maximum draw every hour is an upper bound. The PUE uplift fits an owner who pays a facility's whole energy bill, while a colocation tenant may pay for metered IT power with cooling inside the per-kW rate. CBRE notes that tenants historically "paid electricity costs based on actual server-level power utilization" [28].

4. Space: colocation asking rates per kW

CBRE gives the H2 2025 average asking rate for a 250 to 500 kW requirement in primary wholesale colocation markets as $196.25 per kW-month, up 6.6% year over year, and $195.94 in the same report's market overview [27]. Rates are quoted asking rates for 250 kW and above at N+1/Tier III requirements [27, 28]. In Q1 2026 CBRE put Chicago at $200 to $230 and Northern Virginia at $190 to $235, with Northern Virginia vacancy at 0.3% [26]. The 250 to 500 kW average rose a further 4.3% in H1 2026 [28].

Rate per kW-month	IT load per GPU	Per GPU-year	Per GPU-hour
$190	1.79 kW	$4,076	$0.47
$196.25	1.79 kW	$4,210	$0.48
$235	1.79 kW	$5,041	$0.58
$196.25	1.93 kW	$4,545	$0.52

Sources: [16, 21, 26, 27]. The arithmetic is ours.

Colocation costs more per GPU-hour than electricity at the base inputs, $0.48 against $0.27. CBRE's band of 250 to 500 kW is about 130 to 260 B200 GPUs at IREN's 1.93 kW per GPU [21, 27]. Smaller deployments fall outside it. The rack power and liquid cooling explainer covers how rack density shapes these costs.

CBRE also reports that minimum power utilization floors "are becoming increasingly common in the U.S., with minimum utilization levels ranging from 60% to 85% of allocated power" [28]. Tenants "may be required to pay for a minimum share of their reserved power capacity, even if their actual usage is below that level" [28]. A floor turns part of the electricity bill into a fixed cost, which is why the base case treats power as fixed.

5. Software, support, failures, spares and staff

Software. NVIDIA AI Enterprise lists at $4,500 per GPU for one year and $18,000 for five, "five years for the price of four years", and a footnote says it is "Included with NVIDIA H100, H200 NVL and A800 PCIe GPUs" [29]. NVIDIA lists it as "Recommended" for the DGX B200 [16]. Over five years it costs $0.41 per GPU-hour, against $1 per used GPU-hour on top of the instance in cloud marketplaces [29].

Warranty and spares. Exxact's base build carries three years of parts and labor [18], and Thinkmate's five years of advanced parts replacement [19]. A five-year life on a three-year warranty leaves two years of repairs to the owner. No source here gives a spares ratio, so spares are a reader input.

Failures. A 2024 study of two research clusters spanning 24,000 A100 GPUs over 11 months found 6.50 and 2.34 job failures per thousand node-days [30]. A 2024 report on a 54-day training period on up to 16,000 H100 GPUs counted 419 unexpected interruptions. About 78% were attributed to confirmed or suspected hardware issues, and GPU issues accounted for 58.7% [31]. The team reported higher than 90% effective training time and noted that a single GPU failure may require a restart of the entire job [31]. Owned hours and productive hours diverge.

Utilization. The two research clusters averaged 83% and 85% utilization [30]. No source here reports what enterprise owners reach.

Staff. The median annual wage for network and computer systems administrators was $99,130 in May 2025 [33], or $11.32 for each hour of the year. Staff cost per GPU-hour is $11.32 divided by the GPUs one administrator runs, before benefits and overhead.

6. Rental prices by term and pricing type

B200 option	Pricing type and term	Per GPU-hour	Sources
AWS p6-b200.48xlarge, US East (N. Virginia)	On-demand	$14.24	[3]
AWS p6-b200.48xlarge Capacity Block	Reserved up to six months, prepaid	$14.21	[1, 2]
AWS Compute Savings Plan	1 year, no upfront	$11.22	[4]
Lambda 1-Click Clusters, 16 to 256+ GPUs	Reserved, 2 weeks to 1 year	$9.86 to $8.87	[9]
CoreWeave HGX B200, North America	On-demand	$8.60	[8]
Lambda 8-GPU instance	On-demand	$6.69	[9]
Silicon Data B200 index	Neocloud index	$6.45	[36]
AWS EC2 Instance Savings Plan	3 years, no upfront to all upfront	$6.15 to $5.35	[4]
CoreWeave HGX B200	Spot	$4.26	[8]
CoreWeave committed usage	Up to 60% off on-demand	from $3.44	[8]
WhiteFiber service order, 384 GPUs	Reported contract, 24 months from Jan 2026	about $2.69, derived	[13]

H100 option	Pricing type and term	Per GPU-hour	Sources
Azure ND96isr H100 v5, East US	Pay-as-you-go	$12.29	[6]
Azure	Reservation, 1 year	$7.87	[6]

H100 option	Pricing type and term	Per GPU-hour	Sources
AWS p5.48xlarge, US East (N. Virginia)	On-demand	$6.88	[3]
AWS p5.48xlarge Capacity Block	Reserved up to six months, prepaid	$5.97	[1]
Azure	Reservation, 5 years	$4.92	[6]
Lambda 8-GPU instance	On-demand	$3.99	[9]
AWS EC2 Instance Savings Plan	3 years, no upfront to all upfront	$2.97 to $2.59	[4]
SemiAnalysis 1-year contract index	Reserved, 1 year, March 2026	$2.35	[12]
WhiteFiber service order, 256 servers	Reported contract, 24 months from Mar 2026	about $1.40, derived	[13]

Viewed 24 October 2026 unless dated. Instance prices are divided by eight GPUs.

What each commitment buys. AWS Capacity Blocks reserve one to 64 instances for up to six months, booked up to eight weeks ahead and paid up front [1, 2]. AWS states that "Savings Plans doesn't provide capacity reservations" [5], and its price file lists separate No Upfront, Partial Upfront and All Upfront rates for each term [4]. Lambda's clusters hold 16 to 512 GPUs on "an NVIDIA Quantum-2 400 Gb/s InfiniBand non-blocking fabric", plus three CPU management nodes [10]. CoreWeave offers "up to 60% discounts over our On-Demand prices for committed usage" [8].

Azure reservations. The Retail Prices API returned one-, three- and five-year reservations for the ND H100 v5, though its documentation describes one- or three-year terms [6, 7]. Each record prices the whole term, as in the documentation's sample [7]. Per hour, the reservations sit 36%, 56% and 60% below pay-as-you-go [6].

Reported contracts. WhiteFiber reports service orders by GPU count, term and "aggregate revenue opportunity" [13]. Dividing gives about $2.69 per GPU-hour for 384 B200 GPUs over 24 months ($18.1 million), $1.40 for 256 H100 servers over 24 months ($50.2 million, at the filing's own equivalence of eight GPUs per server) and $1.59 for 10 H200 servers over 14 months ($1.3 million) [13]. The values may include services.

Posted prices and contract prices. Silicon Data says its index observations are standardized "for machine specs, rental terms, platform performance, and geolocation" [11, 36]. SemiAnalysis writes that "most of the GPU rental market is transacted on a long-term basis with contracts of at least 6mths and longer", negotiated bilaterally and not typically posted, and that on-demand pricing works by "fixing price at a constant level with take-up rates or utilization rates the variable" [12]. The posted GPU prices explainer covers what such listings show.

7. Break-even utilization and cost per reserved hour

B200 against pay-as-you-go rates. The base owning cost at full use is $2.43 per GPU-hour: capital $1.68, electricity $0.27 and colocation $0.48. With the software subscription it is $2.84.

B200 rate, 24 Oct 2026	Per GPU-hour	Break-even, power fixed	With software	Power variable
AWS on-demand	$14.24	17.0%	19.9%	15.5%
CoreWeave on-demand	$8.60	28.2%	33.0%	25.9%
Lambda 8-GPU instance	$6.69	36.3%	42.4%	33.6%
Silicon Data B200 index	$6.45	37.6%	44.0%	34.9%
CoreWeave spot	$4.26	56.9%	66.6%	54.0%

Sources: [3, 8, 9, 36] for rates. The arithmetic is ours.

Treating power as variable lowers each break-even by 1 to 3 percentage points. The provider and pricing type move it from 17% to 57%.

Utilization	Owning cost per used hour, power fixed	Power variable
100%	$2.43	$2.43
85%	$2.85	$2.81
70%	$3.46	$3.35
50%	$4.85	$4.59
30%	$8.08	$7.47

B200 against committed terms. CoreWeave describes its committed contracts as typically take-or-pay, "requiring payment regardless of the level of utilization" [14]. An owner also pays for every hour, so the comparison is per reserved hour.

Committed B200 option	Per GPU-hour	Above the $2.43 owning cost
WhiteFiber 24-month order, derived	$2.69	$0.26
CoreWeave committed, full 60% discount	$3.44	$1.01
AWS 3-year Instance Savings Plan	$5.35 to $6.15	$2.92 to $3.72
Lambda 1-Click Clusters	$8.87 to $9.86	$6.44 to $7.43
AWS Capacity Block	$14.21	$11.78

Sources: [1, 4, 8, 9, 13, 14]. The arithmetic is ours.

The derived WhiteFiber rate sits $0.26 above the base owning cost and $0.15 below the cost with software. A gap of $0.26 equals the median administrator wage of $11.32 an hour [33] spread over about 43 GPUs. A Capacity Block costs nearly six times the base owning cost, for a reservation of up to six months with no hardware to manage [1, 2]. A 24-month contract also covers only 40% of a five-year life. After it ends the owner keeps the hardware and its re-rental or resale value, which the rent-to-value note's full report discusses under residual value.

The H200 case. The low end of Thinkmate's HGX H200 starting prices, $407,660 [20], plus eight NDR400 cards at the $2,211 Thinkmate lists for its B200 system [19], gives $53,169 per GPU. With NVIDIA's 10.2 kW figure [17] and the other base inputs [24, 25, 27], owning costs $2.09 per GPU-hour at full use (capital $1.56, electricity $0.19, colocation $0.34), or $2.50 with software. At the high starting price it is $2.27.

H200 rate	Per GPU-hour	Break-even, power fixed	With software
AWS p5en on-demand	$7.91	26%	32%
CoreWeave on-demand	$6.31	33%	40%
Silicon Data H200 index	$3.48	60%	72%
CoreWeave spot	$2.62	80%	96%

Sources: [3, 8, 11] for rates. The arithmetic is ours.

Against committed H200 terms, $2.09 compares with about $1.59 for WhiteFiber's 14-month order, $2.97 to $3.42 for AWS's three-year Savings Plans and $7.90 for a Capacity Block [1, 4, 13]. Against each comparable rate, the H200 owner breaks even at higher utilization than the B200 owner, because H200 rents sit closer to its owning cost.

8. Sensitivity: what moves the result

Input moved, others at base	Range from sources	Owning cost at full use	Change
Server price per GPU	$54,167 to $65,324	$2.34 to $2.66	$0.33
Life	6 to 5 years	$2.20 to $2.43	$0.22
Financing rate	6.0% to 9.0%	$2.30 to $2.43	$0.13
Electricity	7.21 to 24.12 cents per kWh	$2.36 to $2.82	$0.47
PUE	1.1 to 1.54	$2.35 to $2.43	$0.08
IT power per GPU	1.79 to 1.93 kW	$2.43 to $2.48	$0.06
Colocation	$190 to $235 per kW-month	$2.41 to $2.52	$0.11
Software subscription	none to $18,000 over 5 years	$2.43 to $2.84	$0.41

Sources: [13, 14, 15, 16, 18, 19, 21, 24, 25, 26, 27, 29]. The arithmetic is ours.

With every input at the low end ($54,167, 6.0%, six years, Texas industrial power, a PUE of 1.1, $190 per kW-month, no software), owning costs $1.86 per GPU-hour. With every input at the high end ($65,324, 9.0%, five years, California industrial power, a PUE of 1.54, $235 per kW-month, software), it costs $3.57. Break-even against Lambda's $6.69 then runs from 28% to 53%, and against AWS's $14.24 from 13% to 25%.

The rental side moves further. B200 rates in this report span $2.69 to $14.24 per GPU-hour [3, 13], a spread about 25 times the $0.47 change from electricity, which moves owning cost more than any other single input. H100 rates span $1.40 to $12.29 [1, 3, 4, 6, 9, 12, 13]. Switching from CoreWeave's on-demand rate to its spot rate moves the B200 break-even from 28% to 57%, and switching from AWS on-demand to Lambda's instance moves it from 17% to 36% [3, 8, 9]. Electricity's full sourced range moves it by no more than 7 points against Lambda's rate. On these inputs, the provider, term and pricing type move the answer more than any single ownership cost.

Implications

For ML platform leads and CTOs. Two questions decide the comparison. The first is the term the team would sign. Against on-demand and index B200 rates in these sources, owning breaks even between 17% and 44% utilization, before staff and networking [3, 8, 9, 36]. Against committed rates, the reserved-hour price decides, and those prices span about $2.69 to $14.21 [1, 13]. The second is the utilization the team expects to reach. The only published levels here come from two large research clusters, at 83% and 85% [30], and failure reports show owned hours exceeding productive hours on large jobs [31].

For procurement and finance teams. A committed rate carries terms beyond its price: reserved capacity and prepayment for a Capacity Block [1, 2], a price with no capacity reservation for a Savings Plan [5], and an InfiniBand fabric with management nodes for a Lambda cluster [10]. A hardware price needs its scope stated: network cards, memory, warranty years, and whether switches, installation, shipping and tax are included [18, 19]. The financing rate and the life together move the base owning cost by up to $0.35 per GPU-hour.

For researchers. A rental price without its term, pricing type, node size and interconnect cannot be compared with an owning cost. The same GPU class spans almost nine times from low to high in these sources [1, 3, 4, 6, 9, 12, 13].

The Rillor Compute Index lists weighing renting against owning among the uses of its series, with RCI Rental broken out by GPU class, region, term and pricing type and RCI Hardware by comparable system grade [35]. Those are the two price inputs this method compares.

Limits of this analysis

- Every worked number illustrates the method on public inputs. None is a Rillor Compute Index value, a quote or an offer.
- The owning cost excludes staff, cluster switches and cables, storage, insurance, taxes, shipping, installation, deployment time and resale value. It is a floor.
- Hardware prices are listings and reported transactions of different dates, scopes and buyer sizes. No current listing for a new HGX H100 system was found among the sellers opened.
- The H200 case uses a starting price, NVIDIA's DGX H100/H200 power figure as a proxy for third-party servers, and the network card price from Thinkmate's B200 configurator.
- Electricity is charged at maximum draw every hour with the PUE uplift, an upper bound for many colocation contracts. August 2026 EIA prices are preliminary.
- CBRE figures are asking rates for 250 kW and above at N+1/Tier III requirements, and Azure per-hour reservation rates divide whole-term prices by the term's hours.
- WhiteFiber rates divide reported contract values that may include services, and the H100 count rests on eight GPUs per server.
- Silicon Data's index pages showed one reading in their headline and a different one in their questions section on 24 October 2026: $2.88 and $2.53 for the H100 [11], and $6.45 and $5.45 for the B200 [36]. This report uses the dated headline readings. At the B200 questions-section reading of $5.45, break-even would be 44.5%, or 52.0% with software.
- Google Cloud prices were not included, because its pricing pages did not render for automated reading.
- The analysis is pre-tax and assumes constant prices over the life. It is not tax, accounting or investment advice.

Sources

1. Amazon EC2 Capacity Blocks for ML pricing. Amazon Web Services, viewed 24 October 2026. https://aws.amazon.com/ec2/capacityblocks/pricing/

2. Amazon EC2 Capacity Blocks for ML. Amazon Web Services, viewed 24 October 2026. https://aws.amazon.com/ec2/capacityblocks/

3. Amazon EC2 On-Demand price data, US East (N. Virginia), Linux. Amazon Web Services, published 23 October 2026. https://b0.p.awsstatic.com/pricing/2.0/meteredUnitMaps/ec2/USD/current/ec2-ondemand-without-sec-sel/US%20East%20%28N.%20Virginia%29/Linux/index.json

4. AWS Price List, Savings Plans for Compute, US East (N. Virginia), version 20261023221239. Amazon Web Services, published 23 October 2026. https://pricing.us-east-1.amazonaws.com/savingsPlan/v1.0/aws/AWSComputeSavingsPlan/20261023221239/us-east-1/index.json

5. Compute Savings Plans and Reserved Instances. Savings Plans User Guide, Amazon Web Services, viewed 24 October 2026. https://docs.aws.amazon.com/savingsplans/latest/userguide/sp-ris.html

6. Azure Retail Prices API, query for Standard_ND96isr_H100_v5 in East US. Microsoft, viewed 24 October 2026. https://prices.azure.com/api/retail/prices?$filter=armSkuName%20eq%20%27Standard_ND96isr_H100_v5%27%20and%20armRegionName%20eq%20%27eastus%27

7. Azure Retail Prices REST API overview. Microsoft Learn, viewed 24 October 2026. https://learn.microsoft.com/en-us/rest/api/cost-management/retail-prices/azure-retail-prices

8. Pricing. CoreWeave, viewed 24 October 2026. https://www.coreweave.com/pricing

9. GPU cloud pricing, archived capture of 25 October 2026 showing the rates viewed on 24 October 2026. Lambda, via the Internet Archive Wayback Machine, viewed 7 November 2026. https://web.archive.org/web/20261025040216/https://lambda.ai/pricing

10. 1-Click Clusters: Introduction. Lambda Docs, viewed 24 October 2026. https://docs.lambda.ai/public-cloud/1-click-clusters/

11. H100 Rental Price Index (SDH100RT), with related index readings. Silicon Data, viewed 24 October 2026. https://www.silicondata.com/products/silicon-index/h100

12. The Great GPU Shortage: Rental Capacity: Launching our H100 1 Year Rental Price Index. SemiAnalysis, 2 April 2026. https://newsletter.semianalysis.com/p/the-great-gpu-shortage-rental-capacity

13. WhiteFiber, Inc. Quarterly Report on Form 10-Q for the quarter ended 30 June 2026. U.S. Securities and Exchange Commission, filed 12 August 2026. https://www.sec.gov/Archives/edgar/data/2042022/000121390026088026/ea0300877-10q_whitefiber.htm

14. CoreWeave, Inc. Annual Report on Form 10-K for the year ended 31 December 2025. U.S. Securities and Exchange Commission, filed 2 March 2026. https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/crwv-20251231.htm

15. IREN Reports FY26 Results (Exhibit 99.1 to Form 8-K). IREN, 27 August 2026. https://www.sec.gov/Archives/edgar/data/0001878848/000187884826000051/irenreportsfy26results.htm

16. NVIDIA DGX B200. NVIDIA, viewed 24 October 2026. https://www.nvidia.com/en-us/data-center/dgx-b200/

17. Introduction to NVIDIA DGX H100/H200 Systems. NVIDIA DGX H100/H200 User Guide, viewed 24 October 2026. https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html

18. Exxact TensorEX 8U HGX B200 Server TS4-142487900, configurator. Exxact Corporation, viewed 24 October 2026. https://configurator.exxactcorp.com/configure/TS4-142487900

19. GPX GXS6-2808-8HGX-B2, configurator. Thinkmate, viewed 24 October 2026. https://www.thinkmate.com/system/gpx-gxs6-2808-8hgx-b2

20. GPX GPU-Optimized Servers. Thinkmate, viewed 24 October 2026. https://www.thinkmate.com/systems/servers/gpx

21. IREN Expands AI Cloud with 2.4k NVIDIA Blackwell GPUs (Exhibit 99.1 to Form 8-K). IREN, 3 July 2025. https://www.sec.gov/Archives/edgar/data/1878848/000114036125024747/ef20051566_ex99-1.htm

22. IREN Secures $9.7bn AI Cloud Contract with Microsoft (Exhibit 99.1 to Form 8-K). IREN, 3 November 2025. https://www.sec.gov/Archives/edgar/data/1878848/000114036125040072/ef20058139_ex99-1.htm

23. Exascale Labs Holdings Inc. Current Report on Form 8-K (purchase and sale agreement). U.S. Securities and Exchange Commission, dated 15 October 2026. https://www.sec.gov/Archives/edgar/data/2109869/000182912626011139/exascalelabs_8k.htm

24. Electric Power Monthly, Table 5.6.A, Average Price of Electricity to Ultimate Customers by End-Use Sector, by State, August 2026 and 2025. U.S. Energy Information Administration, released 23 October 2026. https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a

25. Uptime Institute Global Data Center Survey 2025 (UII Keynote Report 180). Uptime Institute, July 2025. https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2025.Annual.Survey.Report.pdf

26. Global Data Center Trends 2026. CBRE, 17 June 2026. https://www.cbre.com/insights/reports/global-data-center-trends-2026

27. North America Data Center Trends H2 2025. CBRE, 25 February 2026. https://www.cbre.com/insights/books/north-america-data-center-trends-h2-2025

28. North America Data Center Trends H1 2026. CBRE, 27 August 2026. https://www.cbre.com/insights/books/north-america-data-center-trends-h1-2026

29. NVIDIA AI Enterprise pricing. NVIDIA Enterprise Licensing Guide, viewed 24 October 2026. https://docs.nvidia.com/ai-enterprise/planning-resource/licensing-guide/latest/pricing.html

30. Revisiting Reliability in Large-Scale Machine Learning Research Clusters (Apostolos Kokolis, Michael Kuchnik, John Hoffman and others). arXiv, 29 October 2024. https://arxiv.org/html/2410.21680v1

31. The Llama 3 Herd of Models (Aaron Grattafiori and others), version 3. arXiv, 23 November 2024. https://arxiv.org/html/2407.21783v3

32. Publication 946 (2025), How To Depreciate Property. Internal Revenue Service, viewed 24 October 2026. https://www.irs.gov/publications/p946

33. Network and Computer Systems Administrators, Occupational Outlook Handbook. U.S. Bureau of Labor Statistics, viewed 24 October 2026. https://www.bls.gov/ooh/computer-and-information-technology/network-and-computer-systems-administrators.htm

34. Simple Levelized Cost of Energy (LCOE) Calculator Documentation. National Laboratory of the Rockies, Energy Systems Analysis, last updated 6 December 2025. https://www.nlr.gov/analysis/tech-lcoe-documentation

35. Rillor Compute Index. Rillor Corporation, viewed 24 October 2026. https://rillor.com/markets/compute-index

36. B200 Rental Price Index (SDB200RT). Silicon Data, viewed 24 October 2026. https://www.silicondata.com/products/silicon-index/b200

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Rent or buy GPUs: a break-even method from public inputs. 7 November 2026. https://rillor.com/insights/rent-or-buy-gpus-break-even

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Rillor Compute Index series are for research and planning use. They are not offered for settlement of any contract.

NVIDIA, AMD, Ceph, Kubernetes and related product names are trademarks of their owners and are used only to identify products. No affiliation or endorsement is implied.

Work with Rillor

Send the series, GPU classes and history you need, and how you plan to use them. Contact us at rillor.com/contact?type=index or contact@rillor.com. More on the Rillor Compute Index: rillor.com/markets/compute-index