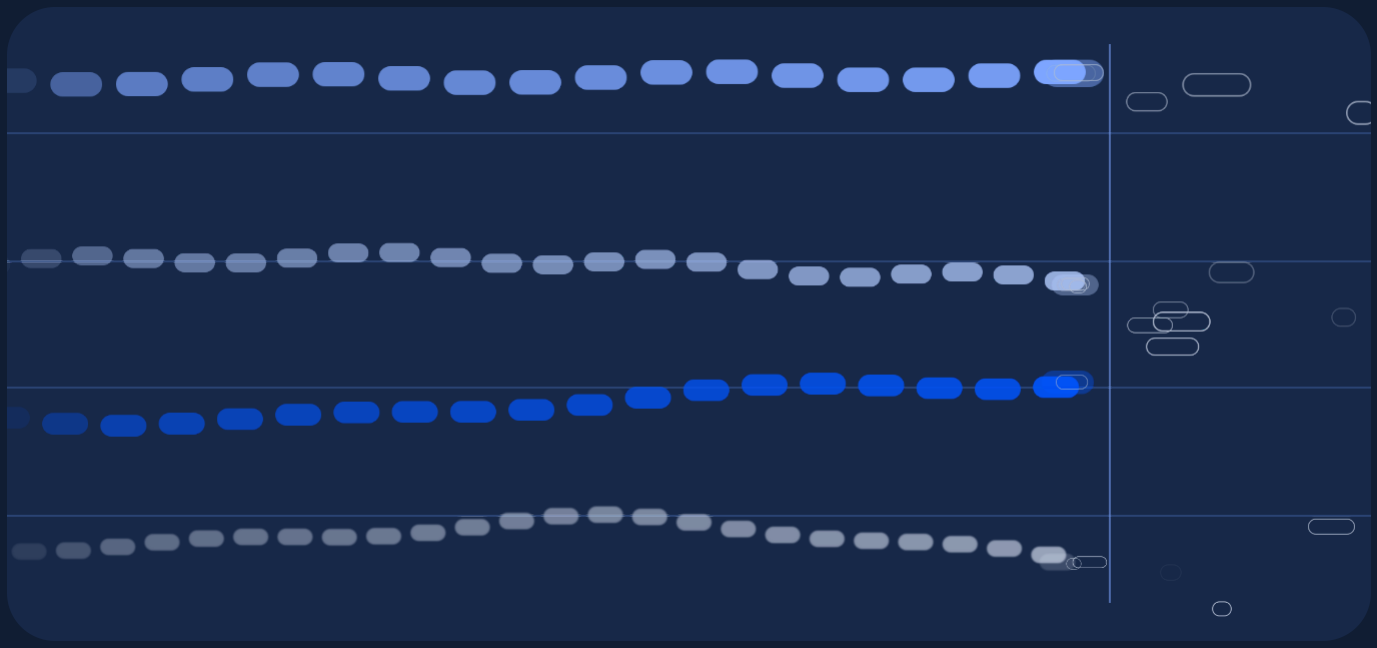
Rillor

RESEARCH REPORT

H100 rental price history, 2023 to 2026

An H100 price history from dated public sources: what list prices, rental indexes and surveyed contracts show for NVIDIA H100 GPUs from 2023 to November 2026.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Dated public sources show H100 one-year contract prices falling from 2023 to October 2025 and a daily rental index falling through 2025, both rising into 2026, while providers' list prices moved on their own schedules.

- On 10 November 2026 an NVIDIA H100 GPU read $3.06 per GPU-hour on Silicon Data's daily index of GPU-focused clouds (neoclouds) and $3.10 on Ornn's index of executed rentals. Azure's on-demand list price, viewed on 12 November, came to $12.29, its 8-GPU rate divided by eight for this report [11, 22, 34].
- SemiAnalysis's surveyed one-year H100 contract price fell from $2.70 to $3.40 per GPU-hour in the first half of 2023 to $1.45 to $1.95 in October 2025, a level it calls "a low of $1.70/hr/GPU" [31, 32].
- Prices turned up from late 2025. Silicon Data's H100 index rose from $2.00 on 9 December 2025 to $3.06 on 10 November 2026, 53.0% (computed for this report), and SemiAnalysis put one-year contracts at $2.35 by March 2026 [22, 25, 32].
- The on-demand list prices of AWS, Azure and Google Cloud, the hyperscalers, changed once in the sources reviewed: AWS cut its H100 instance (P5) rate 44% from 1 June 2025, Azure's rates carry a 1 December 2023 start date and Google's read $88.49 an hour in every capture from July 2024 [7, 11, 12].
- Neocloud list prices moved in steps. Lambda's 8-GPU H100 rate went from $2.59 per GPU-hour in September 2023 to $2.99 from August 2024 and $3.99 from April 2026, with $4.59 posted to take effect on 16 November 2026 [13, 14, 15].
- AWS raised its Capacity Block reservation price for P5 three times in 2026, from $3.933 to $5.970 per accelerator-hour, 51.8% in all (computed for this report), and its page names January 2027 for the next update [8].
- The hyperscaler premium narrowed. Silicon Data's regional data put hyperscaler H100 rates at 3 to 4.5 times other providers' at the start of 2025 and roughly 2 to 2.5 times in the third quarter of 2026 [28].
- Most GPU rental capacity is sold on long-term terms. Committed contracts made up over 98% of CoreWeave's 2025 revenue, and SemiAnalysis writes that most of the GPU rental market trades on contracts of six months or longer [17, 32].

Background

NVIDIA announced the H100 on 22 March 2022 in SXM and PCIe form factors, with the SXM version on HGX H100 server boards in four-way and eight-way configurations [1]. On 20 September 2022 it said the GPU was in full production and named Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure among the first clouds to deploy H100-based instances, starting in 2023 [2]. NVIDIA's product page, viewed on 8 November 2026, lists two variants: the H100 SXM with 80GB of memory at up to 700W, and the H100 NVL with 94GB on a dual-slot, air-cooled PCIe card at 350 to 400W [3].

AWS made its P5 instances available on 26 July 2023 in US East (N. Virginia) and US West (Oregon) [4], and Microsoft its ND H100 v5 virtual machines on 7 August 2023 [10]. The hyperscalers price H100 capacity per hour for 8-GPU instances: AWS's p5.48xlarge, Azure's ND96isr H100 v5 and Google Cloud's a3-highgpu-8g [4, 5, 10, 11, 12]. The p5.48xlarge carries 192 vCPUs, 2,048 GiB of memory, eight 3,840 GB SSDs and 3,200 gigabits of networking, and AWS's file also lists a 1-GPU p5.4xlarge at exactly one eighth of its price [5, 8]. GPU-focused clouds, here called neoclouds, post per GPU-hour, as Lambda and Nebius do [13, 19]. Dividing an 8-GPU instance price by eight gives a per GPU-hour figure that still includes the host's processors, memory and storage. CoreWeave's 2023 page went the other way: it listed a GPU component price and charged vCPUs and memory on top [16].

The same GPU rents on several terms. Nebius offered reservations "for a commitment period, ranging from 3 months to 3 years" beside its on-demand rate in 2024 [19], and Lambda offered three-year and three-month reserved clusters in 2023 [13]. AWS's Capacity Blocks reserve GPU instances for set dates, with a fee "charged up front at the time you schedule the reservation", and AWS updates their prices "regularly based on trends in supply and demand" [8].

Five kinds of series record H100 prices over time.

- **Provider list prices**, read from pricing pages, price files and price APIs, live or from Internet Archive captures [5, 8, 11, 12, 13, 14, 15, 16, 19].
- **Daily rental indexes.** Silicon Data's SDH100RT is a "daily benchmark for NVIDIA H100 hourly rental pricing, with separate neo-cloud and hyperscaler readings" [22]. Ornn's index is built "from executed trades rather than listings or surveys" [34].
- **A surveyed contract index.** SemiAnalysis's one-year H100 rental price index "is constructed from direct survey data across a pool of 100+ market participants" [32]. Its composite index blends provider prices with those surveys [31].
- **Publishers' retrospective medians**: Silicon Data's December 2025 table of H100 prices by tier [21] and its regional quarterly medians of July 2026 [28].
- **Participants' accounts**, such as Eugene Cheah's October 2024 essay [33].

Method

What was examined. How the list, indexed and surveyed rental prices of NVIDIA H100 GPUs changed from the 2023 launch to 12 November 2026, in sources that carry a date.

Sources. Provider pricing pages, price files and price APIs from AWS [5, 8], Microsoft [11], Google Cloud [12], Lambda [13, 14, 15], CoreWeave [16] and Nebius [19], read live and from Internet Archive Wayback Machine captures. Provider announcements [1, 2, 4, 6, 7, 10]. Filings and investor documents: CoreWeave's 2025 Form 10-K [17] and Q3 2025 earnings call [18], Nebius's Q2 2026 shareholder letter [20] and CME Group's 4 November 2026 letter to the Commodity Futures Trading Commission (CFTC) [36]. Index publishers' pages and dated posts from Silicon Data [21, 22, 24, 25, 26, 27, 28, 29, 30], SemiAnalysis [31, 32] and Ornn [34]. Two press reports [9, 23], one participant's essay [33] and one academic paper [35]. Each was viewed on 8 November 2026, and [5, 8, 11, 12, 16, 19, 22, 34] again on 12 November 2026.

How. Each figure is kept with its date, provider tier, term and the kind of series it belongs to. A capture shows a price on its capture date only, so a change between two captures is dated "between" them, and a price seen in two captures is not claimed for every day in between. Per GPU-hour figures for 8-GPU instances are the instance price divided by eight, and percentage changes are worked from the cited figures. Both are marked "computed". AWS's per-accelerator Capacity Block rates are AWS's own figures. Series of different kinds are never joined into one line, and levels from different constructions by one publisher stay apart. Causes of price moves are given only as the sources state them.

Disclosure. Rillor Corporation publishes this report and administers the Rillor Compute Index. Every figure in this report comes from the dated public sources listed.

Findings

1. 2023: launch-year prices far apart by provider

AWS's on-demand price file published on 23 August 2023 listed the p5.48xlarge in US East (N. Virginia) at $98.32 an hour, $12.29 per GPU-hour (computed for this report) [5]. Azure's pay-as-you-go rate for the ND96isr H100 v5 in East US is the same $98.32 an hour, with an effective start date of 1 December 2023 [11].

Neocloud rates sat far lower. Lambda's 3 July 2023 page said its on-demand cloud "now includes the new NVIDIA H100 Tensor Core GPUs (PCIe)", at $1.99 an hour for one GPU [13]. Its 9 September 2023 page priced 8x H100 SXM instances at $2.59 per GPU-hour ($20.72 an hour per instance), reserved H100 clusters at $1.89 per H100-hour for three years and a three-month "Sprint" term at $4.85 [13]. CoreWeave's 6 July 2023 page listed an HGX H100 GPU component at $4.76 an hour and an H100 PCIe at $4.25, with vCPUs at $0.01 and memory at $0.005 per GB-hour on top, under a banner of "starting at $2.23/hr" that names no term or configuration [16].

The retrospective series start in 2023 too. SemiAnalysis's pricing history puts one-year H100 contracts at $2.70 to $3.40 per GPU-hour in the first half of 2023 and $2.65 to $3.30 in the second, as 25th to 75th percentile ranges, with its composite at $6.62 for the second half [31]. Silicon Data's December 2025 analysis gives hyperscaler monthly averages of $7.62 to $7.77 from August to December 2023, a median of $7.76, and notes that in 2023 "Only Hyperscaler is present" [21]. The same post says that for much of 2023 "On-demand pricing from major clouds often exceeded $12.00 per GPU-hour", which it does not reconcile with its median [21].

Cheah, who disclosed that he co-founded an inference provider that rents GPUs, recalled in October 2024 that "the original rental rates of H100 started at approximately $4.70 an hour but were going for over $8" [33].

2. 2024: a neocloud rise and cut, and contracts easing

Lambda raised its rates on 1 January 2024. Its page showed a "2023 Price" and a "2024 Price (effective 1/1/2024)": the 8x H100 SXM went from $20.72 to $27.92 an hour, $2.59 to $3.49 per GPU-hour, up 34.7%, and the 1x H100 PCIe from $1.99 to $2.49, up 25.1% (all computed) [13]. The 8x SXM price stayed at $27.92 in captures through 28 July 2024 [13]. By 15 August 2024 the page showed a "New Price" of $2.99 per GPU-hour, 14.3% below $3.49 (computed), and reserved 8x H100 clusters as "CONTACT SALES" [13].

Cheah wrote that "At the start of 2024, the H100 prices reached approximately $2.85 across multiple providers", and that in August 2024 a renter willing to "auction for a small slice of H100 time (days to weeks)" could find H100s for $1 to $2 an hour [33]. These are one participant's observations, not a measured series.

AWS's list price held where the sources show it. An AWS post of 11 September 2024 gave the p5.48xlarge at $98.32 on demand and $43.18 on a three-year reserved rate, $12.29 and $5.40 per GPU-hour (computed) [6]. The Capacity Block page captured on 6 July 2024 listed P5 at $39.33 per instance-hour and $4.916 per

accelerator-hour, unchanged in captures through 12 January 2025 [8]. CoreWeave's GPU component prices stayed at $4.76 and $4.25 in captures from December 2023 to 6 December 2024 [16]. Nebius's page of 15 November 2024 listed one H100 SXM at $2.95 an hour on demand and $2.00 reserved [19].

SemiAnalysis's one-year range eased each quarter of 2024: $2.50 to $3.10, $2.20 to $2.50, $2.15 to $2.45 and $1.90 to $2.10 [31]. Its composite went from $5.78 in the first quarter to $3.91 in the fourth, and its on-demand column from $3.90 in the second quarter to $2.85 to $3.15 in the fourth [31].

Silicon Data's retrospective table moved the other way for hyperscalers, with medians of $7.92 from January to May 2024 and $9.34 from June to December [21]. Its marketplace readings start in June 2024 and its neocloud readings in July, with second-half medians of $2.58 and $2.99, the point at which, it writes, "A clear tiering appears" [21].

3. 2025: hyperscaler cuts, a daily index and the October contract low

AWS cut first in its Capacity Blocks. P5 fell from $4.916 to $3.933 per accelerator-hour between the captures of 12 January and 8 March 2025, by 20.0% (computed) [8]. On 5 June 2025 AWS announced price cuts for its NVIDIA GPU instances "from May 31, 2025 baseline prices", 44% on demand for P5 (H100), for On-Demand purchases from 1 June in all Regions [7]. It wrote that the growth in demand for GPU capacity "has outpaced industry-wide supply", and that it works to lower costs "so that we can pass those savings back to our customers" [7]. The price file published on 10 November 2026 lists the p5.48xlarge at $55.04 an hour, 44.0% below its August 2023 figure and $6.88 per GPU-hour (both computed) [5].

Silicon Data's retrospective medians show the step. Its hyperscaler median was $8.96 from January to May 2025, $6.94 in June, "A visible step-down month for Hyperscaler", and $6.26 from July to December [21]. Its marketplace median fell from $2.29 to $1.95 between those two periods, and its neocloud median went from $3.50 to $3.33 [21].

IEEE Spectrum's report on Silicon Data's daily index, published on 28 May 2025, gave an average hourly H100 rental price of $2.37 on 27 May [23]. It described the price going "up mildly to $2.50 per hour" after a model release in January 2025, then sliding to $2.30 "for much of February" [23]. In August 2025 the index moved between 2.22 and 2.26, with volatility of about 1.7%, and by mid-September it dipped as low as 2.13 before recovering toward 2.16 [24].

Contract prices reached what SemiAnalysis calls "a low of $1.70/hr/GPU in October 2025" [32]. Its table gives a one-year range of $1.45 to $1.95 that month, and a composite between $2.79 and $2.90 in each month from July to December [31].

Neocloud list prices held through the year. Lambda showed $2.99 per GPU-hour for the 8x H100 SXM in captures from 1 January to 10 October 2025 [13, 14]. Nebius showed $2.95 in captures of 23 May and 21 November 2025 [19]. CoreWeave's page from 1 January 2025 priced a full 8-GPU HGX H100 instance at $49.24 an hour, a different unit from its earlier GPU component price [16].

4. Late 2025 to mid-2026: the turn in contracts and indexes

On its earnings call of 10 November 2025, CoreWeave said that "in Q3, we saw our first 10,000-plus H100 contract approaching expiration", and that two quarters in advance the customer re-contracted "at a price within 5% of the original agreement" [18]. The original price is not public.

Silicon Data then reported a 10% rise in its H100 index between 9 December 2025 and 6 January 2026, from $2.00 to $2.20 an hour, while its A100 and B200 indexes were stable [25]. SemiAnalysis wrote on 2 April 2026 that "Rental pricing broke above $2/hr/GPU for a 1y contract by late January" and reached $2.35 by March [32]. In the same report it wrote that "On-Demand GPU rental capacity is sold out across all GPU types" and that "H100s are getting renewed at the exact same rate they were signed at 2-3 years ago" [32].

Period	SemiAnalysis one-year range	Midpoint, computed	On-demand column
First half 2023	$2.70 to $3.40	$3.05	Not shown
Fourth quarter 2024	$1.90 to $2.10	$2.00	$2.85 to $3.15
October 2025	$1.45 to $1.95	$1.70	$2.20 to $3.00
February 2026	$1.80 to $2.35	$2.08	Sold Out
March 2026	$2.00 to $2.70	$2.35	Sold Out
April 2026	$2.10 to $2.70	$2.40	Sold Out

Source: [31]. Ranges are 25th to 75th percentiles, and the one-year column assumes prepayment, "typically 25%" [31]. The midpoint of $1.70 matches the October low and $2.35 the March level that SemiAnalysis reports [32]. From $1.70 to $2.35 is a rise of 38.2% (computed) [31, 32].

The hyperscaler tier stayed flat while the neocloud tier rose. Silicon Data's H100 Hyperscaler On-Demand Index traded in a band of $7.40 to $7.52 from 1 March to 20 April 2026, unchanged on 39 of 50 daily transitions, for a net fall of 0.27% [26]. Over April and May its band was $7.40 to $7.55, unchanged on 43 of 60 transitions, while its neocloud index went "from a $2.47 trough on April 8 to a $2.76 peak on May 30, an 11.7% trough-to-peak swing" [27]. The neocloud rate was $2.57 on 4 May and $2.75 on 27 July, after reaching what Silicon Data called "its series high of $2.79 on June 5" [29].

5. 2026: neocloud list prices and Capacity Blocks rose

Lambda's 8x H100 SXM went from $2.99 to $3.44 per GPU-hour between the captures of 3 February and 10 March 2026, then to $3.99 by 1 April, still shown on 1 November 2026, which is 15.1% and then 33.4% above $2.99 (computed) [14]. Its instances page, viewed on 8 November 2026, posts $4.59 per GPU-hour for the 8-GPU instance with the note "These prices take effect November 16, 2026, at 00:00 UTC" [15]. That rate is 15.0% above $3.99, 53.5% above the February 2026 rate and 77.2% above September 2023's $2.59 (all computed) [13, 14, 15].

Nebius's on-demand H100 rate went from $2.95 to $3.85 between the captures of 7 May and 12 July 2026, up 30.5%, then to $4.50 in a column headed "GPU-hour (Effective October 1, 2026)", up a further 16.9% (both computed) [19]. Its live page shows $4.50 on demand and preemptible capacity "from $0.79" [19]. Nebius told shareholders on 12 August 2026 that its deals were "driven by increasing prices for new-generation GPUs and more than 30% higher pricing on older-generation GPUs versus Q1" [20]. The letter does not say which GPUs it counts as older-generation.

AWS raised its Capacity Block rates three times. P5 went from $3.933 to $4.326 per accelerator-hour between the captures of 31 March and 15 May 2026, to $5.191 by 25 August and to $5.970 on the live page [8]. The steps are 10.0%, 20.0% and 15.0%, and 51.8% from $3.933 in all (computed) [8]. Every capture from July 2024 names the next update: 1 October 2024 in the 6 July 2024 capture, April 2026 in that of 31 March 2026, July 2026 in captures of 15 May to 4 July, October 2026 in those of 25 August and 1 October, and now "January, 2027" [8]. Tradingpedia reported the July step in advance, on 26 June 2026, as "about 20% starting July 1, 2026" for P5 and other GPU families [9]. The live P5 rate is $47.757 per instance-hour in every region listed [8].

Azure's and Google Cloud's on-demand rates showed no change. Azure's pay-as-you-go and one-, three- and five-year reservation prices for the ND96isr H100 v5 in East US all carry an effective start date of 1 December 2023, at $12.29, $7.87, $5.40 and $4.92 per GPU-hour (computed) [11]. Its spot rate, $18.17 an hour or $2.27 per GPU-hour (computed), carries a start date of 1 January 2026 [11]. Google Cloud's a3-highgpu-8g, with eight H100s, was $88.49 an hour on demand in captures of 2 June 2025, 8 July 2025 and 1 January 2026 and on the live page, with one- and three-year commitments at $61.38 and $38.86 an hour, or $11.06, $7.67 and $4.86 per GPU-hour (computed) [12]. CoreWeave's live page keeps the 8-GPU HGX H100 instance at $49.24 an hour, as on 1 January 2025, with a spot price of $19.71, or $6.155 and $2.46 per GPU-hour (computed) [16].

6. The hyperscaler premium narrowed

Silicon Data's regional analysis of 31 July 2026 found that "Hyperscaler H100 rates fell 26 to 54% by region since Q1 2025 while non-hyperscaler rates recovered, compressing the hyperscaler premium from roughly 250% to about 120% across the seven regions we track" [28]. As a multiple, renting from a hyperscaler cost 3 to 4.5 times the non-hyperscaler rate at the start of 2025 and roughly 2 to 2.5 times in the third quarter of 2026 [28].

Region	Hyperscaler decline since the first quarter of 2025, as reported
US East	About 26%
US West and the Middle East	Around 30%
Canada, the Nordics and Europe	Just over 40%
Asia-Pacific	About 54%

Source: [28]. In US West, hyperscaler rates fell from roughly $8.90 to about $5.60 by late 2025, then rebounded to around $6.10 [28]. In the Nordics they fell from about $8.90 to $5.20 while non-hyperscaler rates rose to roughly $3.30, a multiple of about 1.6 times [28]. These are quarterly medians, and the third quarter of 2026 is partial [28].

List prices show a like-for-like case. AWS's P5 on-demand rate was 4.7 times Lambda's 8-GPU rate in 2023, $12.29 against $2.59, and is 1.5 times the rate Lambda posts for 16 November 2026, $6.88 against $4.59 (computed) [5, 13, 15].

7. The series side by side on 10 to 12 November 2026

Series	Kind of price	Per GPU-hour	Source
Silicon Data SDH100RT, neocloud	Daily rental index	$3.06	[22]
Ornn index, H100 SXM	Daily index of executed rentals	$3.10	[34]
Lambda, 8x H100 SXM, on demand	Neocloud list price, 1 Nov 2026 capture	$3.99	[14]
Nebius, HGX H100, on demand	Neocloud list price	$4.50	[19]
AWS Capacity Block, P5	Hyperscaler reservation price	$5.970	[8]
CoreWeave, 8-GPU HGX H100, on demand	Neocloud list price	$6.155, computed	[16]
AWS p5.48xlarge, on demand	Hyperscaler list price	$6.88, computed	[5]
Google Cloud a3-highgpu-8g, on demand	Hyperscaler list price	$11.06, computed	[12]

Series	Kind of price	Per GPU-hour	Source
Azure ND96isr H100 v5, pay as you go	Hyperscaler list price	$12.29, computed	[11]

Silicon Data's index read $2.85 on 13 October and $2.92 on 31 October before $3.06 on 10 November, rises of 7.4% and 4.8% (computed) [22]. Ornn printed its H100 SXM reading with a one-month change of 16.10% and a one-day fall of 6.06% [34]. The two indexes sit further apart on newer GPUs: Silicon Data showed H200 at $3.47 and B200 at $6.84 on 11 November [22], and Ornn H200 at $7.24 and B200 at $9.90 on 10 November [34].

Spot quotes spread wider. Silicon Data's sample of H100 spot prices from 8 to 16 October 2026 had a market median of $2.33 per GPU-hour and a daily range across sellers of $1.38 to $4.49, which it charts apart from Nebius's new dynamic spot price, at $0.79 for most of the window [30].

Silicon Data's H100 index also underlies a proposed futures contract. NYMEX submitted "Silicon Data H100 Rental Index Futures and Silicon Data B200 Rental Index Futures" on 11 August 2026, and on 4 November 2026, at the CFTC's request, consented to extend the review period "until January 8, 2027" [36].

8. Where the sources disagree

- **The size of AWS's June 2025 cut.** AWS gives 44% for P5 on demand [7]. Silicon Data writes that AWS "reduced H100 pricing by roughly 30%" [21]. AWS's price files show a fall of 44.0% (computed) [5].
- **Hyperscaler prices in 2023.** Silicon Data's text says major clouds often exceeded $12.00 per GPU-hour, while its table gives a 2023 hyperscaler median of $7.76 [21]. AWS's on-demand list rate, $12.29 per GPU-hour (computed), fits the text [5].
- **Silicon Data's hyperscaler levels.** Its daily hyperscaler index sat between $7.40 and $7.55 in spring 2026 [26, 27]. Its regional quarterly medians put hyperscaler rates in US West and the Nordics at about $5.20 to $6.10 from late 2025 [28], and its July to December 2025 median is $6.26 [21]. These are different constructions from one publisher, so their levels are not read as one series.
- **One index page, three figures.** The SDH100RT page's FAQ says "The current H100 rental price is $2.53 per GPU-hour", against its dated reading of $3.06, and its "7.4% (7D)" label matches the change since 13 October, while the change since 4 November is 3.4% (computed) [22].
- **Composite and contract.** SemiAnalysis's composite read $2.80 in March 2026, while its one-year range had a midpoint of $2.35 (computed) [31]. The composite draws on "hyperscale clouds, neoclouds, GPU marketplaces, and rental routers/aggregators" with contract surveys, "with hyperscale clouds weighted more strongly than marketplaces and neoclouds" [31].
- **Lambda's cluster table.** In March and April 2026 Lambda's 1-Click Clusters table showed H100 on demand for 2 weeks to 12 months at $2.76 under the heading "Price per hour", and from 16 May $5.54 to $6.16 by cluster size, a column the 1 November capture labels per GPU-hour [14]. The table changed form,

so the two are not a like-for-like change.

9. Most GPU rental capacity is contracted long term

CoreWeave's 2025 annual report says that "Customers generally access our platform through multi-year committed contracts, under which they purchase a specified amount of capacity on a take-or-pay basis over the contract term" [17]. Committed contracts accounted for over 98%, 96% and 88% of its revenue in 2025, 2024 and 2023 [17]. SemiAnalysis writes that "most of the GPU rental market is transacted on a long-term basis with contracts of at least 6mths and longer", and that it validates its levels "with transaction data as well as by arranging a few transactions ourselves" [32].

Bandi and Su write that "direct no-arbitrage links between futures prices and current spot prices fail due to the non-storable nature of compute" [35]. An unused GPU-hour cannot be stored for later sale, so today's rent and a later one are not tied by inventory.

Implications

For buyers comparing a quote with the past. An H100 price needs its date, tier, term and kind before it can be set against history. Silicon Data's H100 index rose 53.0% from 9 December 2025, Nebius's on-demand rate 52.5% from its 21 November 2025 capture, Lambda's rate posted for 16 November 53.5% from its 3 February 2026 capture and AWS's Capacity Block rate 51.8% from its 31 March 2026 capture, while Azure's on-demand rate kept its 2023 start date (all computed) [8, 11, 14, 15, 19, 22, 25].

For finance teams and lenders. Posted on-demand prices describe a small share of revenue at a provider whose committed contracts made up over 98% of 2025 revenue [17]. Renewal reports describe the contract side: a 10,000-plus H100 contract renewed within 5% of its original price, and renewals at "the exact same rate" [18, 32]. There the surveyed one-year index fell to $1.70 and rose to $2.35 [31, 32].

For superintelligence (SI) developers and other buyers planning capacity. Long-term reserved and committed rates sat below on-demand rates in the posted terms reviewed, from Lambda's three-year $1.89 against $2.59 in 2023 to Azure's five-year $4.92 against $12.29 (computed) [11, 13]. Lambda's three-month Sprint term, at $4.85, sat above its on-demand rate [13].

For index users and researchers. Each publisher's series is its own construction, and one publisher can carry several [21, 22, 26, 28]. Silicon Data's daily index and Ornn's executed-rental index read 4 cents apart for H100 on 10 November 2026 and dollars apart for H200 and B200 [22, 34].

Limits of this analysis

- Archived captures are irregular. A change is dated only between two captures, and AWS's $98.32 on-demand rate is shown only for August 2023 and September 2024. The fall to $55.04, 44.0% (computed), matches AWS's 44% cut from its 31 May 2025 baseline, but no source shows the price on every day in between [5, 6, 7].
- The Azure Retail Prices API reports only the current price's start date, so no Azure H100 price before 1 December 2023 is shown [11]. Google's 8 December 2023 capture does not list the A3 machine type, so no Google H100 price for 2023 is shown. Its 30 July 2024 capture lists the a3-highgpu-8g at $88.49 an hour on demand in us-central1, with one- and three-year commitments at $61.37 and $38.84 [12].
- CoreWeave's 2023 and 2024 GPU component price and its 2025 and 2026 instance price are different units, and its 2023 banner names no term [16].
- Silicon Data's hyperscaler reading is not in its index page's static text [22]. SemiAnalysis's latest values load by script, so its public monthly series here ends in April 2026 [31].
- Ornn's index counts only executed rentals, and it is read here on 10 November 2026 only [34]. CoreWeave's renewal remark gives a price relative to an undisclosed original [18]. None of the sources reviewed gives transaction prices for 2023 or 2024.
- Retrospective tables mix inputs: Silicon Data's combines "Silicon Data's internal tracking with commonly observed public-market pricing" [21]. Cheah disclosed his company's interest [33].
- H100 purchase and server prices are outside this report, and it makes no forecast of H100 rents.

Sources

1. NVIDIA Announces Hopper Architecture, the Next Generation of Accelerated Computing. NVIDIA Corporation, 22 March 2022, viewed 8 November 2026. https://nvidianews.nvidia.com/news/nvidia-announces-hopper-architecture-the-next-generation-of-accelerated-computing

2. NVIDIA Hopper in Full Production. NVIDIA Corporation, 20 September 2022, viewed 8 November 2026. https://nvidianews.nvidia.com/news/nvidia-hopper-in-full-production

3. NVIDIA H100 GPU. NVIDIA Corporation, viewed 8 November 2026. https://www.nvidia.com/en-us/data-center/h100/

4. New: Amazon EC2 P5 Instances powered by NVIDIA H100 Tensor Core GPUs for Accelerating Generative AI and HPC Applications. AWS News Blog. Amazon Web Services, 26 July 2023, viewed 8 November 2026. https://aws.amazon.com/blogs/aws/new-amazon-ec2-p5-instances-powered-by-nvidia-h100-tensor-core-gpus-for-accelerating-generative-ai-and-hpc-applications/

5. Amazon EC2 On-Demand price file, US East (N. Virginia), Linux. Amazon Web Services, published 23 August 2023 (Internet Archive Wayback Machine capture of 25 August 2023) and 10 November 2026, viewed 12 November 2026. https://web.archive.org/web/20230825130912/https://b0.p.awsstatic.com/pricing/2.0/meteredUnitMaps/ec2/USD/current/ec2-ondemand-without-sec-sel/US%20East%20(N.%20Virginia)/Linux/index.json, 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

6. Christensen. A review of purpose-built accelerators for financial services. AWS Machine Learning Blog. Amazon Web Services, 11 September 2024, viewed 8 November 2026. https://aws.amazon.com/blogs/machine-learning/a-review-of-purpose-built-accelerators-for-financial-services/

7. Yun. Announcing up to 45% price reduction for Amazon EC2 NVIDIA GPU-accelerated instances. AWS News Blog. Amazon Web Services, 5 June 2025, viewed 8 November 2026. https://aws.amazon.com/blogs/aws/announcing-up-to-45-price-reduction-for-amazon-ec2-nvidia-gpu-accelerated-instances/

8. Amazon EC2 Capacity Blocks for ML pricing. Amazon Web Services, live page and Internet Archive Wayback Machine captures of 1 November 2023, 6 July 2024, 12 January 2025, 8 March 2025, 31 March 2026, 15 May 2026, 4 July 2026, 25 August 2026 and 1 October 2026, viewed 12 November 2026. https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20231101120811/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20240706033947/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20250112155010/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20250308091133/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20260331164533/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20260515000834/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20260704092824/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20260825162005/https://aws.amazon.com/ec2/capacityblocks/pricing/, https://web.archive.org/web/20261001113914/https://aws.amazon.com/ec2/capacityblocks/pricing/

9. Fisher. AWS Raises GPU Capacity Block Prices 20% for AI Workloads. Tradingpedia, updated 26 June 2026, viewed 8 November 2026. https://www.tradingpedia.com/2026/06/26/aws-raises-gpu-capacity-block-prices-20-for-ai-workloads/

10. Scale generative AI with new Azure AI infrastructure advancements and availability. Azure Blog. Microsoft, 7 August 2023, viewed 8 November 2026. https://azure.microsoft.com/en-us/blog/scale-generative-ai-with-new-azure-ai-infrastructure-advancements-and-availability/

11. Azure Retail Prices API, Standard_ND96isr_H100_v5, East US. Microsoft, viewed 12 November 2026. https://prices.azure.com/api/retail/prices?api-version=2023-01-01-preview&$filter=armRegionName%20eq%20%27eastus%27%20and%20armSkuName%20eq%20%27Standard_ND96isr_H100_v5%27

12. Accelerator-optimized VM pricing, and VM instance pricing. Google Cloud, live page and Internet Archive Wayback Machine captures of 8 December 2023, 30 July 2024, 2 June 2025, 8 July 2025 and 1 January 2026, viewed 12 November 2026. https://cloud.google.com/products/compute/pricing/accelerator-optimized, https://web.archive.org/web/20231208085211/https://cloud.google.com/compute/vm-instance-pricing, https://web.archive.org/web/20240730111140/https://cloud.google.com/compute/vm-instance-pricing, https://web.archive.org/web/20250602102901/https://cloud.google.com/compute/vm-instance-pricing, https://web.archive.org/web/20250708184348/https://cloud.google.com/compute/vm-instance-pricing, https://web.archive.org/web/20260101105218/https://cloud.google.com/compute/vm-instance-pricing

13. GPU Cloud pricing. Lambda, Internet Archive Wayback Machine captures of 3 July 2023, 9 September 2023, 3 December 2023, 1 January 2024, 28 July 2024, 15 August 2024, 1 January 2025 and 16 March 2025, viewed 8 November 2026. https://web.archive.org/web/20230703030126/https://lambdalabs.com/service/gpu-cloud, https://web.archive.org/web/20230909194011/https://lambdalabs.com/service/gpu-cloud, https://web.archive.org/web/20231203205050/https://lambdalabs.com/service/gpu-cloud, https://web.archive.org/web/20240101073433/https://lambdalabs.com/service/gpu-cloud, https://web.archive.org/web/20240728044745/https://lambdalabs.com/service/gpu-cloud, https://web.archive.org/web/20240815220723/https://lambdalabs.com/service/gpu-cloud, https://web.archive.org/web/20250101080209/https://lambdalabs.com/service/gpu-cloud, https://web.archive.org/web/20250316063514/https://lambdalabs.com/service/gpu-cloud

14. Pricing. Lambda, Internet Archive Wayback Machine captures of 4 June 2025, 10 October 2025, 3 February 2026, 10 March 2026, 1 April 2026, 16 May 2026 and 1 November 2026, viewed 8 November 2026. https://web.archive.org/web/20250604191644/https://lambda.ai/pricing, https://web.archive.org/web/20251010223149/https://lambda.ai/pricing, https://web.archive.org/web/20260203015539/https://lambda.ai/pricing, https://web.archive.org/web/20260310143253/https://lambda.ai/pricing, https://web.archive.org/web/20260401163040/https://lambda.ai/pricing, https://web.archive.org/web/20260516025341/https://lambda.ai/pricing, https://web.archive.org/web/20261101120500/https://lambda.ai/pricing

15. Instances. Lambda, viewed 8 November 2026. https://lambda.com/instances

16. Pricing. CoreWeave, Internet Archive Wayback Machine captures of 6 July 2023, 13 December 2023, 6 December 2024 and 1 January 2025, and the live page, viewed 12 November 2026. https://web.archive.org/web/20230706175917/https://www.coreweave.com/gpu-cloud-pricing,

https://web.archive.org/web/20231213134534/https://www.coreweave.com/gpu-cloud-pricing, https://web.archive.org/web/20241206151408/https://www.coreweave.com/gpu-cloud-pricing, https://web.archive.org/web/20250101002525/https://www.coreweave.com/pricing, https://www.coreweave.com/pricing

17. CoreWeave, Inc. Annual Report on Form 10-K for the year ended 31 December 2025. CoreWeave, Inc., via SEC EDGAR, filed 2 March 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/crwv-20251231.htm

18. CoreWeave, Inc. (CRWV) Q3 2025 Earnings Call, corrected transcript. CoreWeave, Inc., 10 November 2025, viewed 8 November 2026. https://s205.q4cdn.com/133937190/files/doc_financials/2025/q3/CORRECTED-TRANSCRIPT_-CoreWeave-Inc-CRWV-US-Q3-2025-Earnings-Call-10-November-2025-5_00-PM-ET.pdf

19. Prices. Nebius, Internet Archive Wayback Machine captures of 15 November 2024, 23 May 2025, 21 November 2025, 7 May 2026, 12 July 2026 and 1 October 2026, and the live page, viewed 12 November 2026. https://web.archive.org/web/20241115043809/https://nebius.com/prices, https://web.archive.org/web/20250523205822/https://nebius.com/prices, https://web.archive.org/web/20251121191852/https://nebius.com/prices, https://web.archive.org/web/20260507111158/https://nebius.com/prices, https://web.archive.org/web/20260712235400/https://nebius.com/prices, https://web.archive.org/web/20261001131140/https://nebius.com/prices, https://nebius.com/prices

20. Letter to Shareholders, Q2 2026, Exhibit 99.2 to Form 6-K. Nebius Group N.V., via SEC EDGAR, 12 August 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1513845/000110465926094568/tm2622968d1_ex99-2.htm

21. Li. H100 Rental Price Over Time (2023-2025): A Complete Market Analysis. Silicon Data, 21 December 2025, viewed 8 November 2026. https://www.silicondata.com/blog/h100-rental-price-over-time

22. H100 Rental Price Index: SDH100RT. Silicon Data, updated 11 November 2026, viewed 12 November 2026. https://www.silicondata.com/products/silicon-index/h100

23. Moore. Price Index Could Clarify Opaque GPU Costs for AI. IEEE Spectrum, 28 May 2025, viewed 8 November 2026. https://spectrum.ieee.org/gpu-prices

24. Li. H100 Rental Market Cools in September: Price Index Slips as Volatility Rises. Silicon Data, 20 October 2025, viewed 8 November 2026. https://www.silicondata.com/blog/h100-rental-market-update-september-2025

25. Li. H100 Price Spike: Understanding the 10% Surge in GPU Rental Costs. Silicon Data, 9 January 2026, viewed 8 November 2026. https://www.silicondata.com/blog/h100-price-spike

26. H100 Hyperscaler Index, April 2026: A Benchmark in Flat Mode. Silicon Data, 27 April 2026, viewed 8 November 2026. https://www.silicondata.com/blog/h100-hyperscaler-index-april-2026

27. Li. H100 Index, May 2026: Same Chip, Two Tiers, Two Regimes. Silicon Data, 16 June 2026, viewed 8 November 2026. https://www.silicondata.com/blog/h100-index-may-2026-two-tiers-two-regimes

28. Gao. H100 Rental Prices by Region (Q3 2026): The Hyperscaler Premium Has Halved. Silicon Data, 31 July 2026, viewed 8 November 2026. https://www.silicondata.com/blog/h100-rental-prices-by-region-q3-2026

29. Li. H200 vs H100 Rental Prices, May to July 2026: The Premium That Doubled. Silicon Data, 4 August 2026, viewed 8 November 2026. https://www.silicondata.com/blog/h200-vs-h100-rental-prices-may-july-2026

30. Dynamic Pricing Brings Price Discovery to GPU Spot Rental Markets. Silicon Data, 21 October 2026, viewed 8 November 2026. https://www.silicondata.com/blog/gpu-spot-prices-dynamic-pricing-price-discovery

31. H100 Daily & Contract Pricing Index. SemiAnalysis, viewed 8 November 2026. https://gpu-index.semianalysis.com/

32. Nishball, Nanos, Cheang et al. The Great GPU Shortage: Rental Capacity, Launching our H100 1 Year Rental Price Index. SemiAnalysis, 2 April 2026, viewed 8 November 2026. https://newsletter.semianalysis.com/p/the-great-gpu-shortage-rental-capacity

33. Cheah. $2 H100s: How the GPU Rental Bubble Burst. Latent Space, 11 October 2024, viewed 8 November 2026. https://www.latent.space/p/gpu-bubble

34. GPU Market Prices: the Ornn Compute Price Index. Ornn Data, latest settlement 10 November 2026, viewed 12 November 2026. https://data.ornn.com/markets

35. Bandi and Su. (Early) AI Compute Asset Pricing. arXiv 2607.12156 v4. arXiv, 21 October 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.12156

36. Marchese. Voluntary Extension of Review Period, NYMEX Submission No. 26-370. CME Group Inc., filed with the Commodity Futures Trading Commission, 4 November 2026, viewed 8 November 2026. https://www.cftc.gov/filings/documents/2026/orgdcmnymexvolexten261105.pdf

About this report

Published by Rillor on 13 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. H100 rental price history, 2023 to 2026. 13 November 2026. https://rillor.com/insights/h100-rental-price-history-2023-to-2026

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