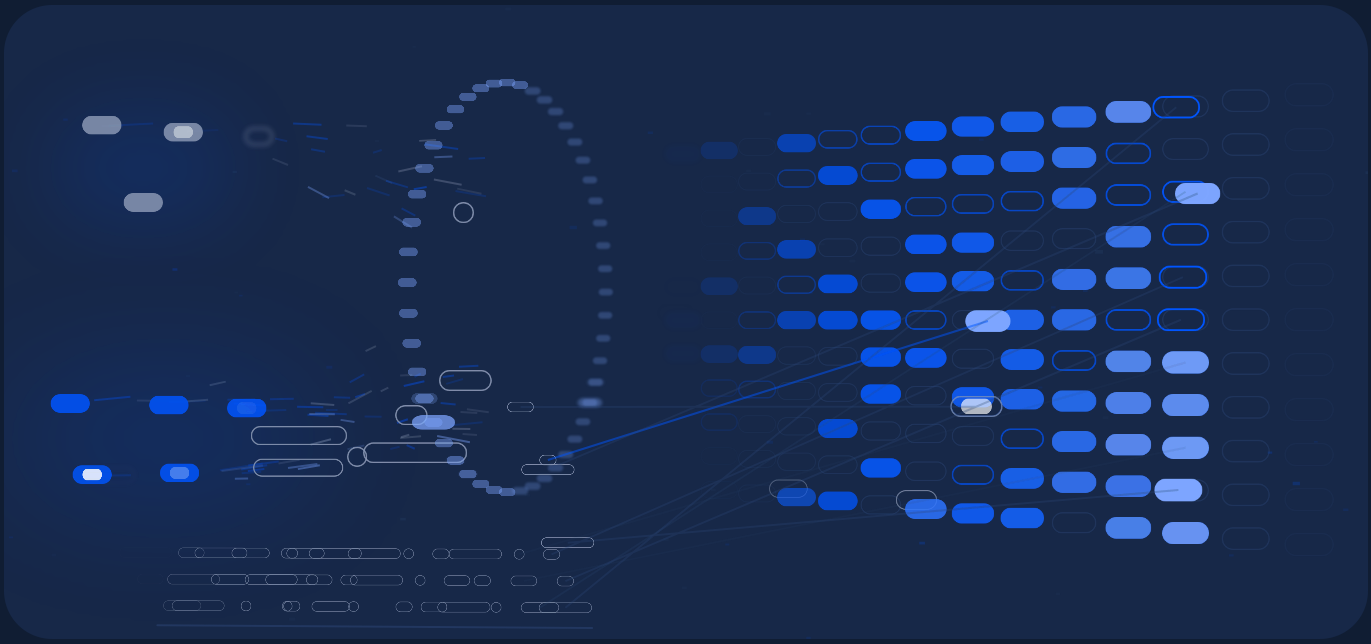
Rillor

EXPLAINER

What posted GPU cloud prices tell you, and what they do not

A posted GPU price is a listing for one configuration, term and moment. Comparing two of them takes seven adjustments, and none turns a listing into a transaction.



Rillor

Contents

01 Summary

02 Background

03 Data and method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

A posted GPU cloud price is a provider's public answer to one question: what this configuration costs, under these terms, if capacity is there. It is a real fact. It is a narrower fact than most uses of it assume.

- **One machine, many posted prices.** On 9 October 2026, AWS's on-demand, reservation, Savings Plans and Spot prices for the same 8-GPU H100 instance in one region ranged from $2.59 to $6.88 per GPU-hour, depending on pricing type, term and payment [3][4][5][6]. Spot prices for that instance differed by up to 28% across Availability Zones in the same region on the same day [5].
- **Posted prices change by decision.** AWS cut P5 on-demand prices by 44% in June 2025 [7], raised Capacity Block prices by about 15% in January 2026 and about 20% in July 2026 [8][9], and describes Capacity Block prices as updated with supply and demand [3].
- **Availability flags are thin signals.** Provider APIs report where an instance type can be launched, or how likely a Spot request is to succeed, without guarantees or quantities [10][11][12][13].
- **Posted prices are not transaction prices.** Large contracts are priced privately, sometimes per megawatt rather than per GPU-hour [20][21]. The CFTC's 2026 request for comment asks what proportion of compute transactions occur at publicly disclosed prices [28]. One index publisher describes the hyperscaler on-demand tier as a posted ceiling above reserved capacity [14].
- **Normalization is where comparisons fail.** GPU variant, interconnect, node size, region, term, bundled CPU, memory and storage, and billing rules each change what a GPU-hour buys. Two posted prices are comparable only after all of them are stated.
- **The Rillor Compute Index methodology** treats a posted rental price as a listing, records availability as reported, normalizes each observation on those dimensions, keeps pricing types and the hyperscaler and neocloud segments apart, and takes a median for each provider before the median across providers [25].

Background

GPU cloud prices are easy to find and hard to use. Hyperscalers publish price lists and pricing APIs. Neoclouds publish rate cards. Marketplaces show live listings. Index publishers turn these into daily series, and in 2026 exchanges proposed futures that settle on such series [24][26].

Each of these uses assumes that a posted price measures something stable: the market rate for a GPU-hour. This explainer sets out what a posted price measures, the ways it differs from what buyers pay, and the adjustments needed before two posted prices can be compared. It closes with how the Rillor Compute Index methodology treats posted prices.

Data and method

Primary example. We use one instance type throughout so that the hardware stays fixed: AWS's p5.48xlarge, which has 8 NVIDIA H100 GPUs [6]. Its on-demand and Savings Plans prices were taken from AWS's Price List and Savings Plans APIs on 9 October 2026, and its Spot prices from the EC2 Spot price history API as in effect at 18:00 UTC that day [5]. Its Capacity Block price was read from AWS's Capacity Blocks pricing page on the same day [3]. All prices are for Linux in U.S. dollars. Per GPU-hour figures divide the instance price by 8.

Documentation. Provider documentation for Spot, preemptible and reserved capacity at AWS, Google Cloud and Microsoft Azure, AWS documentation for Capacity Blocks, Savings Plans and Spot placement scores, and Lambda's public cloud API reference for availability fields [1][2][4][10][11][12][13].

Price history and market evidence. AWS announcements and press reports of AWS price changes [7][8][9], Silicon Data's published index commentary [14][15][16][17][18], exchange and regulatory documents on compute futures [22][23][24][26][28], and operator filings that describe contract pricing [20][21].

Hardware specifications. NVIDIA's H100 specifications and AWS's P5 instance specifications [6][19].

Findings

1. One machine, many prices

On 9 October 2026, AWS posted these prices, among others, for a p5.48xlarge in US East (N. Virginia) [3][4][5].

Pricing type	Term and payment	Price per instance-hour	Per GPU-hour
On-Demand	None	$55.04	$6.88
Capacity Block	Reserved block for a future date, paid upfront	$47.757	$5.970
Compute Savings Plan	1 year, no upfront	$43.344	$5.42
EC2 Instance Savings Plan	1 year, no upfront	$34.6752	$4.33
EC2 Instance Savings Plan	3 years, no upfront	$23.77728	$2.97
EC2 Instance Savings Plan	3 years, all upfront (effective hourly rate)	$20.69504	$2.59
Spot	None, interruptible	$28.89 to $37.10 by Availability Zone	$3.61 to $4.64

The hardware is identical in every row. The on-demand rate is about 2.7 times the three-year, all-upfront rate. The Spot range covers the price in effect in each of six Availability Zones in the region at 18:00 UTC that day [5]. In US West (Oregon) the same instance's Spot prices ranged from $22.83 to $32.04 per instance-hour, or $2.85 to $4.00 per GPU-hour [5].

Each row answers a different question. On-demand is what AWS charges for capacity with no commitment, at a static hourly rate [1]. A Capacity Block reserves instances for a future date, up to eight weeks ahead, cannot be cancelled, and is charged a reservation fee whether or not the instances are used [2][3]. A Savings Plan is a commitment to a stated amount of compute use, measured per hour, for one or three years [4]. Spot is spare capacity that AWS can reclaim with a two-minute warning [1].

Google Cloud and Azure post the same kinds of prices with their own rules. Google describes Spot VMs as excess capacity with discounts of up to 91%, preemptible at any time, excluded from its SLA, and priced at rates that "can change up to once every day" [10]. Azure evicts Spot VMs with 30 seconds' notice when it needs the capacity back or when the price exceeds a buyer's maximum, and its Spot pricing varies by region and size [11].

2. Posted prices change by provider decision

A posted price moves when the provider decides to move it, and different pricing types at the same provider can move in opposite directions.

Date	Change	Source
June 2025	AWS cut P5 (H100) on-demand prices by 44% and P5en (H200) by 25%, against 31 May 2025 prices	[7]
January 2026	AWS raised Capacity Block prices by about 15%, and p5e.48xlarge rose from $34.61 to $39.80 per hour in most regions	[8]
1 July 2026	AWS raised Capacity Block prices by about 20% for P6-B300, P6-B200, P5, P5e, P5en and P4de, and P5 in U.S. regions became $5.191 per accelerator-hour	[9]
9 October 2026	AWS's page lists P5 Capacity Blocks at $5.970 per accelerator-hour in U.S. regions and says the next update is scheduled for January 2027	[3]

In January 2026, an AWS spokesperson said that Capacity Block pricing is dynamic and that AWS's "commitment to not raise pricing on fixed pricing models like On Demand and Savings Plans remains unchanged" [8]. A series built from on-demand rates and a series built from reservation rates at the same provider can therefore diverge for reasons that have nothing to do with the hardware.

Index publishers see the same pattern in their data. Silicon Data reported that its hyperscaler H100 on-demand index was unchanged on 39 of 50 daily transitions between 1 March and 20 April 2026, and described it as a price ceiling posted above a large book of already-priced reserved capacity whose stability is "administrative as much as market-derived" [14]. Over April and May its neocloud H100 index moved with annualized volatility of 18.4%, against 3.8% for the hyperscaler tier [15].

3. Availability flags say less than they appear to

Providers expose availability in different forms.

Provider signal	What it reports	Stated limit
Lambda public cloud API, `regions_with_capacity_available`	"A list of the regions in which this instance type is available," alongside `price_cents_per_hour` [12]	A region list, with no quantity
AWS Spot placement score	How likely a Spot request is to succeed in a Region or Availability Zone, as a recommendation [13]	"does not provide any guarantees in terms of available capacity or risk of interruption" [13]
Azure Spot eviction rates	Hourly eviction probability based on the last 7 days of history, shown per size and region [11]	Historical, with no SLA for Spot VMs [11]
Google Cloud Spot VMs	Spot VMs "might not always be available" [10]	Excluded from the Compute Engine SLA [10]
AWS On-Demand	If capacity is not available at launch, the request returns an insufficient capacity error [1]	Known only at launch

A flag that says an instance type is available in a region says nothing about how many GPUs can be launched, for how long, or at what cluster size. A posted price with no availability attached may describe capacity that cannot be bought at that moment. Price and availability have to be read together, and each has to keep its capture time.

4. Posted prices are not transaction prices

Three kinds of evidence show the gap.

Large contracts are priced privately and in other units. Nebius reported in August 2026 that four large contracts carried annual contract value of $20 million to $25 million per megawatt [20]. Nscale's registration statement describes take-or-pay agreements that "provide fixed pricing and contracted revenue over the term" [21]. Neither figure appears on a rate card, and the first is not expressed per GPU-hour at all.

Regulators ask how much trading is visible. The CFTC's August 2026 request for comment on compute derivatives asks "What proportion of compute transactions occur at publicly disclosed prices, and what proportion does not?" and whether it would be appropriate to permit a derivative "settling to a price computed from data that the Commission may not be able to observe, verify, or surveil, in whole or in part" [28]. It states the Commission's preliminary understanding that non-public, bilateral agreements carry the majority of economic value but tend to be undisclosed [28]. It also notes that certain published compute

price series are built in whole or in part from posted or listed rates that providers themselves administer, and asks what would prevent a provider from manipulating a settlement index by adjusting a posted rate [28]. Sidley's summary lists opacity, concentration and limited fungibility and standardization among the obstacles the CFTC identified [23].

Index designs differ on this point. ICE and Ornn describe Ornn's Compute Price Index as built "only from printed transactions" [26]. CME Group's proposed contracts track Silicon Data indexes measuring hourly GPU rental costs, and CME and Silicon Data describe those indices as daily benchmarks for on-demand rental rates [24][27]. Both designs answer real questions. They answer different ones.

A posted price is a listing: an asking price a seller publishes. It shows what the seller asks anyone for that configuration at that time, if capacity is available. It does not show what a buyer with a multi-year, multi-megawatt contract paid.

5. Normalization pitfalls

Comparing two posted prices requires putting them on the same basis. Each dimension below changes what a GPU-hour buys.

Dimension	Example from the sources	Effect on a per GPU-hour comparison
GPU variant	H100 SXM: 80 GB, 3.35 TB/s, up to 700 W, NVLink 900 GB/s. H100 NVL: 94 GB, 3.9 TB/s, 350 to 400 W, NVLink 600 GB/s [19]	Two "H100" prices can describe different chips
Generation and memory	H200 carries 141 GB of HBM3e against 80 GB on the H100 [17]	Silicon Data's neocloud H200 premium over the H100 averaged 5.8% in May and 12.1% in July 2026 [17]
Interconnect	p5.48xlarge: 3,200 Gbps EFA and 900 GB/s NVSwitch between GPUs. p5.4xlarge: one H100, 100 Gbps, no GPUDirect RDMA [6]	AWS posts both at $6.88 per GPU-hour on demand in N. Virginia [5]. Per GPU, that is 400 Gbps of network bandwidth against 100 Gbps, and only the larger node has the NVSwitch fabric
Node size	NVIDIA lists H100 SXM systems with 4 or 8 GPUs, and H100 NVL systems with 1 to 8 GPUs [19]	Single-GPU offers and 8-GPU nodes serve different workloads
Region	p5.48xlarge on demand: $6.88 per GPU-hour in N. Virginia and Oregon, $8.944 in London, $8.60 in Tokyo [5]	A 30% difference for identical hardware at one provider

Dimension	Example from the sources	Effect on a per GPU-hour comparison
Term and payment	$6.88 on demand against $2.59 for a three-year, all-upfront commitment [4][5]	A 2.7-fold difference from commitment alone
Bundled CPU, memory and storage	p5.48xlarge includes 192 vCPUs, 2 TiB of memory and 8 x 3.84 TB of local NVMe [6]	A bare GPU listing without these is a smaller product
Billing rules	Capacity Blocks: reservation fee paid upfront for every reserved hour, no cancellation, plus an operating system fee when running [2][3]. Spot: interruption with two minutes' notice at AWS and 30 seconds at Azure [1][11]	Equal hourly prices can produce different effective costs
Index construction	Silicon Data's daily hyperscaler on-demand H100 index was near $7.46 in May 2026 [15], while its quarterly median series put hyperscaler H100 medians at $4.80 to $6.10 for the third quarter through July [16]	Two series from one publisher differ because they are built differently

Silicon Data's own description of index construction makes the general point: pricing is "inherently multidimensional," a simple average "hides underlying variability," and a short Spot rental in one region "is not economically equivalent to a multi-year Reserved contract in another" [18]. It normalizes heterogeneous units to a standard configuration with a framework it describes as proprietary [18]. Normalization is where index methods differ, and readers should ask each publisher how it is done.

6. How the Rillor Compute Index methodology treats posted prices

The Rillor Compute Index methodology v1.0, published on 9 October 2026, defines four series: RCI Rental, RCI Availability, RCI Hardware and RCI Rent-to-Value [25]. Its treatment of posted rental prices follows the findings above.

- **Evidence class.** A rental price that a provider publishes through a pricing API or pricing page is a listing, "an asking price a seller publishes." Every observation carries its evidence class, so a listing never counts as a sale [25].
- **Capture.** Each observation records one price or availability report, its source and its capture time [25].
- **Availability.** RCI Availability measures the share of tracked offers that providers report as available to rent, recorded as reported, by GPU class and segment [25].
- **Pricing types.** On-demand, spot or preemptible, and reserved prices stay separate. Reserved prices are recorded with the length of the term in months [25].

- **Segments.** Hyperscaler and neocloud sources are shown separately in every series that uses them [25].
- **Normalization.** Prices are put on a per GPU-hour basis and mapped to GPU class and variant, interconnect, node size, region, term and pricing type [25].
- **Calculation.** Each series takes a median for each provider, then the median across providers [25]. A provider with many listings counts once.
- **Conflicts.** Rillor's own quotes and agreements are excluded from every published series [25].
- **Publication.** The methodology sets data rules for the coverage a series needs before a value is published [25].

The series are for research and planning use and are not offered for settlement of any contract [25].

Implications

For buyers. Before comparing a quote with a posted price, state the variant, interconnect, node size, region, term, payment and what the instance includes. A neocloud price for a bare GPU and a hyperscaler price for a full 8-GPU node are different products. Read availability next to price, with the time each was taken.

For builders and operators. Posted rates are public signals that buyers and index publishers read. Rate changes on one pricing type can move a published series even when contract pricing is unchanged.

For lenders. A borrower's projected revenue may rest on contract prices that differ from any posted rate. Ask which evidence a revenue assumption uses: posted on-demand rates, committed rates or the borrower's signed contracts. Posted prices for one machine at one provider span a factor of about 2.7 across terms.

For researchers and index users. Ask four questions of any GPU price series: which evidence class it uses, how it normalizes configurations, how it weights providers, and whether hyperscaler and neocloud prices are mixed. The CFTC's questions on disclosed prices, verifiable settlement and posted rates [28] apply to every index, whether or not a contract settles on it.

Limits of this analysis

- The worked example uses one instance type at one provider on one day. Other providers price differently, and AWS prices change over time.
- Spot figures are single readings: the price in effect for each Availability Zone at 18:00 UTC on 9 October 2026.
- All-upfront Savings Plan figures are AWS's effective hourly rates. The buyer pays the commitment in advance.
- AWS's January and July 2026 Capacity Block changes are taken from press reports of AWS documentation.
- Contract prices discussed here come from operators' own filings and letters. We have no private transaction data.
- Silicon Data figures are reported as the publisher stated them. Its normalization method is not public.
- This explainer describes the Rillor Compute Index methodology. It reports no index values.

Sources

1. Spot Instances, Amazon EC2 User Guide. Amazon Web Services, accessed 9 October 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-instances.html

2. Capacity Blocks for ML, Amazon EC2 User Guide. Amazon Web Services, accessed 9 October 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-capacity-blocks.html

3. Amazon EC2 Capacity Blocks for ML pricing. Amazon Web Services, accessed 9 October 2026. https://aws.amazon.com/ec2/capacityblocks/pricing/

4. What are Savings Plans?, Savings Plans User Guide. Amazon Web Services, accessed 9 October 2026. https://docs.aws.amazon.com/savingsplans/latest/userguide/what-is-savings-plans.html

5. AWS Price List Query API (GetProducts, AmazonEC2), Savings Plans API (DescribeSavingsPlansOfferingRates) and Amazon EC2 API (DescribeSpotPriceHistory), queried for p5.48xlarge and related instance types. Amazon Web Services, retrieved 9 October 2026. https://docs.aws.amazon.com/aws-cost-management/latest/APIReference/API_pricing_GetProducts.html ; https://docs.aws.amazon.com/savingsplans/latest/APIReference/API_DescribeSavingsPlansOfferingRates.html ; https://docs.aws.amazon.com/AWSEC2/latest/APIReference/API_DescribeSpotPriceHistory.html

6. Amazon EC2 P5 Instances. Amazon Web Services, accessed 9 October 2026. https://aws.amazon.com/ec2/instance-types/p5/

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

8. Emma Woollacott, AWS just quietly increased EC2 Capacity Block prices. ITPro, 6 January 2026. https://www.itpro.com/cloud/cloud-computing/aws-amazon-ec2-capacity-blocks-price-increase

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

10. Spot VMs, Compute Engine documentation. Google Cloud, accessed 9 October 2026. https://docs.cloud.google.com/compute/docs/instances/spot

11. About Azure Spot Virtual Machines. Microsoft Learn, updated 24 June 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/spot-vms

12. Lambda Cloud API reference, List available instance types. Lambda, accessed 9 October 2026. https://docs.lambda.ai/api/cloud

13. Spot placement score, Amazon EC2 User Guide. Amazon Web Services, accessed 9 October 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-placement-score.html

14. Carmen Li, H100 Hyperscaler Index, April 2026: A Benchmark in Flat Mode. Silicon Data, 27 April 2026. https://www.silicondata.com/blog/h100-hyperscaler-index-april-2026

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

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

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

18. Carmen Li, Building a Robust GPU Index. Silicon Data, 12 March 2026. https://www.silicondata.com/blog/building-a-robust-gpu-index

19. NVIDIA H100 Tensor Core GPU, specifications. NVIDIA Corporation, accessed 9 October 2026. https://www.nvidia.com/en-us/data-center/h100/

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

21. Nscale Ltd, Registration Statement on Form S-1. U.S. Securities and Exchange Commission EDGAR, 18 September 2026. https://www.sec.gov/Archives/edgar/data/0002110365/000119312526395475/ck0002110365-20260918.htm

22. CFTC Seeks Comment on Listing of Compute Derivatives, Signaling Forthcoming Market Rules for Contracts Tied to AI Computing Power. Baker Botts, 31 August 2026. https://www.bakerbotts.com/thought-leadership/publications/2026/august/cftc-seeks-comment-on-listing-of-compute-derivatives-signaling-forthcoming-market-rules

23. CFTC Seeks Comment on Listing of Compute Derivatives Contracts. Sidley Austin LLP, 4 September 2026. https://www.sidley.com/en/insights/newsupdates/2026/09/cftc-seeks-comment-on-listing-of-compute-derivatives-contracts

24. CME Group and Silicon Data to Launch Compute Futures on October 5 to Unlock New Way to Hedge AI Risks. CME Group and Silicon Data (PR Newswire), 11 August 2026, text as reproduced by StockTitan. https://www.stocktitan.net/news/CME/cme-group-and-silicon-data-to-launch-compute-futures-on-october-5-to-wgqge1aqpsum.html

25. Rillor Compute Index, Methodology v1.0. Rillor Corporation, 9 October 2026. https://rillor.com/markets/compute-index

26. ICE and Ornn to Launch GPU Compute Futures Contracts. Intercontinental Exchange, Inc. (Business Wire), 19 May 2026. https://s2.q4cdn.com/154085107/files/doc_news/ICE-and-Ornn-to-Launch-GPU-Compute-Futures-Contracts-2026.pdf

27. CME Group and Silicon Data Partner to Launch First Compute Futures. CME Group and Silicon Data (PR Newswire), 12 May 2026, text as reproduced by StockTitan. https://www.stocktitan.net/news/CME/cme-group-and-silicon-data-partner-to-launch-first-compute-xsea0cl8o8z2.html

28. Request for Comment on the Listing of Compute Derivatives Contracts, 91 FR 54259. Commodity Futures Trading Commission (Federal Register), 21 August 2026. https://www.govinfo.gov/content/pkg/FR-2026-08-21/pdf/2026-17163.pdf

About this report

Published by Rillor on 10 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. What posted GPU cloud prices tell you, and what they do not. 10 November 2026. https://rillor.com/insights/posted-gpu-prices-what-they-tell-you

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 GPU classes, regions and pricing types you need, and how often you need them. Contact us at rillor.com/contact?type=index or contact@rillor.com. More on GPU and neocloud pricing: rillor.com/markets/gpu-pricing