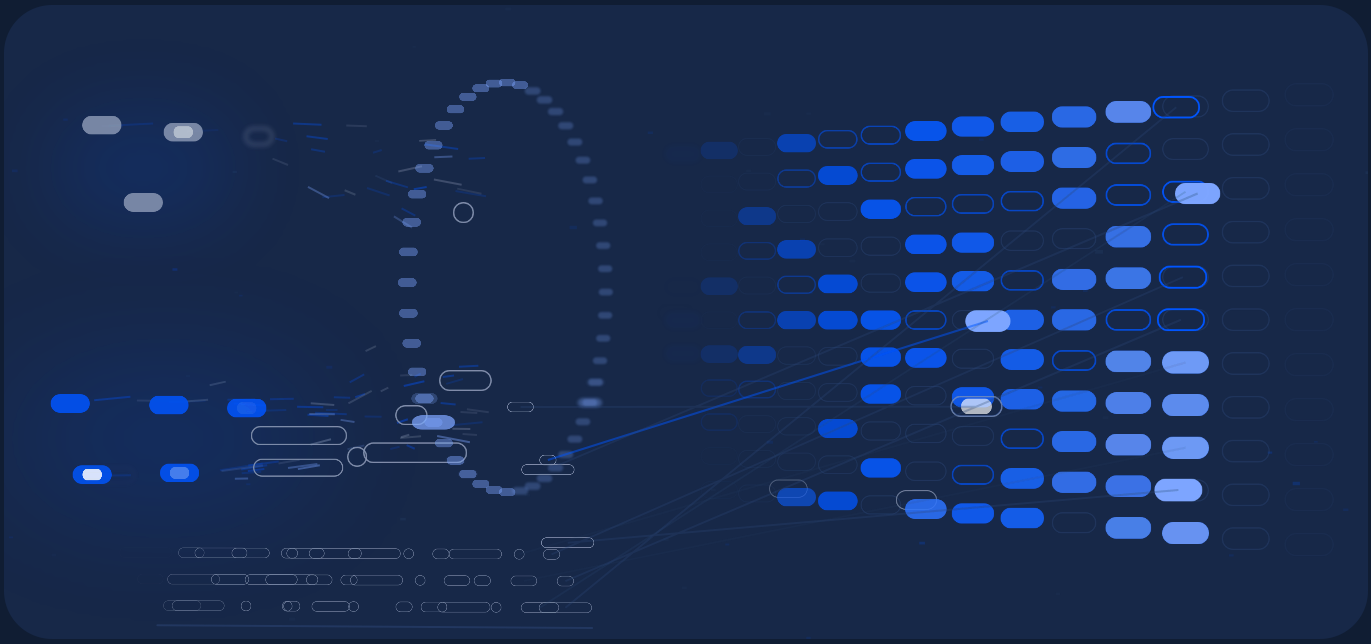
Rillor

EXPLAINER

Neocloud and hyperscaler GPU prices, compared like for like

Neoclouds and hyperscalers sell the same H100 in different packages. Posted prices for 8-GPU nodes on 7 November 2026, put on one basis, and what each price includes.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Neoclouds, the clouds built around GPU capacity, and hyperscalers, the large general-purpose clouds, posted different prices for an 8-GPU NVIDIA H100 node on 7 November 2026, and each price described a different package of hardware, network, region and term.

- On-demand prices for an 8-GPU NVIDIA H100 node, per GPU-hour, ran from $6.88 to $12.29 at four hyperscalers (AWS, Microsoft Azure, Google Cloud and Oracle) and from $3.90 to $6.16 at five neoclouds (CoreWeave, Lambda, Nebius, Crusoe and Together AI) on 7 November 2026, with Lambda at its 25 October price (computed for this report from [9, 15, 18, 22, 25, 29, 30, 33, 35]). The two ranges did not overlap in this sample, and AWS sat 12% above CoreWeave.
- Hyperscaler commitments posted on 7 November 2026 fell inside the $3.90 to $6.16 per GPU-hour range of neocloud on-demand H100 prices: Azure's three- and five-year reservations at $5.40 and $4.92, Google's three-year committed use rate at $4.86 and its Flex-start rate for runs of up to seven days at $4.79, and AWS's Capacity Block, paid up front, at $5.970 (computed for this report from [11, 15, 18, 20]).
- None of the five neoclouds studied posts an H100 price for a term longer than one year. CoreWeave offers up to 60% off on-demand for committed usage, Lambda's "1 year+" cluster term and Together AI's terms over 180 days carry no price, and Crusoe asks customers to contact it for guaranteed resources [25, 28, 33, 35].
- A GPU-hour buys different hardware at each provider that states its specifications: 12 to 28 vCPUs, 120 GB to 256 GiB of host memory and from no local storage to 7.68 TB of NVMe per GPU, with internet egress from free to $0.12 per GiB [9, 10, 17, 21, 23, 24, 26, 27, 30, 34].
- Cluster networking is part of the price. Lambda's InfiniBand-connected H100 clusters posted $5.54 to $6.16 per GPU-hour, 21% to 54% above its single 8-GPU instance and close to CoreWeave's InfiniBand node at $6.155 and AWS's 8-GPU node at $6.88 (computed for this report from [9, 25, 26, 27, 28, 29]).
- Three neoclouds posted H100 on-demand rises of 15% to 17%, effective on 1 October, 16 November and 1 December 2026: Nebius to $4.50, Lambda to $4.59 and Crusoe to $4.50 per GPU-hour (computed for this report from [27, 29, 30, 31, 33]).
- One index publisher reports that the hyperscaler premium in its H100 data narrowed from about 250% to about 120% between the first quarter of 2025 and the third quarter of 2026, and says its data cannot establish why [7].

Background

Two kinds of GPU cloud

Cloud services, in the words of the UK Competition and Markets Authority (CMA), give customers access to shared computing resources on demand, so that they do not need to own the underlying hardware and software [3]. In its final decision of July 2025, it found that cloud infrastructure built on accelerated compute "is being used by a small number of customers who are primarily AI model developers" [3]. It found Microsoft, AWS and Google to be strong vertically integrated providers of accelerated compute and AI-related cloud services, and other cloud providers significantly weaker in this respect [3].

The final report describes two groups of supplier. Microsoft, AWS, Google, Oracle and IBM all supply infrastructure based on accelerated compute alongside their infrastructure based on standard compute [4]. Beside them sit several smaller cloud providers that specialise in accelerated compute. The CMA names CoreWeave, Lambda Labs, Crusoe Cloud, Nebius and Nscale, and records that CoreWeave describes itself as a cloud provider specialising in GPU-accelerated workloads [4]. Their offers tend to target a narrower group of customers, with a more limited range of cluster management options, from basic management software to none at all, known as bare metal [4]. These providers generally do not serve the wider cloud services market [4]. The CMA also found that large public cloud providers in some instances buy bare metal services from other cloud providers [4].

Neither "hyperscaler" nor "neocloud" has a regulator's definition in these documents. The CMA uses the first term only when reporting others, such as a survey submitted by AWS that gave AWS, Azure, GCP, IBM and Alibaba as examples of hyperscale public clouds, and customer and provider submissions that speak of "hyperscalers" [4]. The definitions in use come from the companies and from research. CoreWeave's annual report for 2025 says it competes primarily with hyperscalers that offer general-purpose cloud computing as part of a broader product portfolio, several of which are also its customers and partners [1]. It argues that general-purpose cloud infrastructure was not designed for the high-density compute, advanced networking, optimized storage and distributed-systems software that large GPU workloads need [1]. NVIDIA's annual report for fiscal 2026 lists cloud service providers and neocloud builders as separate types of indirect customer [2]. A January 2025 systems paper contrasts traditional CPU-centric clouds, built on homogeneous hardware, with neoclouds such as CoreWeave, Lambda and Crusoe, built on heterogeneous platforms where tenants choose a concrete hardware platform at a fixed price up front [5].

The neocloud group is large in number. SemiAnalysis's ClusterMAX 3.0, posted on 23 September 2026, reviews 77 neocloud providers and says its wider market view covers 323 providers [6].

Why the posted price is only one view

A posted on-demand price is what anyone can pay for an hour with no commitment, and it describes only part of how GPU capacity is sold. CoreWeave offers on-demand access through a pay-as-you-go model, and says its customers generally access its platform through multi-year committed contracts, buying a specified amount of capacity on a take-or-pay basis [1]. The CMA found the use of committed spend

agreements by providers widespread [3]. How those contracts work, and what they mean for neocloud finances, is covered in Neocloud economics: contract first, hardware second (https://rillor.com/insights/neocloud-economics-contract-first). This report stays with posted prices, the prices both segments publish for any reader.

Method

Basis. One GPU on one basis: an 8-GPU NVIDIA H100 node or instance with 80 GB of memory per GPU and NVLink between the GPUs, on-demand, Linux, in a U.S. region where the provider names one, as posted on 7 November 2026. NVIDIA sells the H100 SXM with 80 GB and 900 GB/s NVLink in partner systems of 4 or 8 GPUs, and the H100 NVL with 94 GB and 600 GB/s NVLink in a PCIe form factor [8]. Every row compared except Together AI's is an 8-GPU H100 with 80 GB per GPU as its provider describes it. Together AI's page names the HGX H100 without a node size [35]. Each price is divided by eight to give U.S. dollars per GPU-hour.

Segments. Hyperscalers here are the large general-purpose clouds that sell GPU compute alongside a broad portfolio: AWS, Microsoft Azure, Google Cloud and Oracle Cloud Infrastructure, from the CMA's list of providers that supply accelerated compute alongside standard compute [4]. Neoclouds are the clouds built around GPU capacity: CoreWeave, Lambda, Nebius and Crusoe, four of the specialists the CMA names [4], and Together AI, placed here by this report's rule. Its placement, like Oracle's, is a classification choice, and no source opened classifies it. The CMA groups Oracle with Microsoft, AWS, Google and IBM [4], but its summary puts the shares of Oracle and IBM below Google's and says they do not supply as wide a range of cloud services as Microsoft, AWS and Google [3]. The AWS survey's list of hyperscale public clouds leaves Oracle out [4]. IBM and Nscale were not priced.

Sources and capture. Every source was opened on 7 November 2026, between about 08:59 and 09:40 UTC. AWS prices come from the public price files its pricing pages load [9, 12], Azure's from Microsoft's unauthenticated Retail Prices API [15, 16], Oracle's from its public list price endpoint [22] and the others from their pricing pages [18, 25, 27, 28, 30, 33, 35]. Hardware details come from product pages and documentation [10, 17, 19, 23, 26, 32, 34], earlier prices from Internet Archive captures [29, 31], and context from the CMA [3, 4], two annual reports [1, 2], a systems paper [5] and two industry publishers [6, 7]. No account, login or quote request was used.

Computation. Per GPU-hour prices are the posted instance price divided by eight. Azure's reservation prices are term totals, divided here by 8,760 hours a year and by eight GPUs. Microsoft states no hour count, so the 8,760 hours are an assumption. The unit conversions follow Reading GPU prices from cloud pricing APIs (https://rillor.com/insights/gpu-prices-from-cloud-pricing-apis). Every derived figure is marked "computed for this report".

What was left out. Together AI's Dedicated Inference price is for an inference endpoint, not raw compute, and is not compared [35]. Single-GPU instances, H100 NVL offers, transaction prices, performance and service terms sit outside the comparison.

Findings

1. The two segments sell different ranges of products

The hyperscalers sell GPU nodes as one product among many, and they post several ways to buy each one. For Google's A3 High with 8 GPUs, the documentation lists eight consumption options, from on-demand, Spot and Flex-start to three kinds of future reservation and resource-based committed use discounts [19]. The same family is not sold the same way at every size. A3 High with 1, 2 or 4 GPUs comes only as Spot or Flex-start, and Google's documentation lists no on-demand option for A3 Ultra [19], although its price table posts one [18]. Google's price table shows "N/A" in the on-demand column for its A4 High (B200) machine [18]. AWS posts on-demand, Capacity Block and Spot prices for its P5 instances, and sells Savings Plans at up to 72% off on-demand [9, 11, 12, 13]. Microsoft's API returns consumption, one-, three- and five-year reservation, Spot and Low Priority prices for its H100 VM [15].

The neoclouds post fewer pricing types, and several route longer terms to sales. CoreWeave posts on-demand and spot prices and offers "up to 60%" off on-demand for committed usage [25]. Lambda sells on-demand instances on a self-serve, first-come basis, and InfiniBand-connected clusters for two weeks to a year, with its "1 year+" term listed without a price [27, 28]. Nebius posts preemptible and on-demand prices and offers commitment discounts for large clusters reserved for multiple months [30]. Crusoe offers spot, on-demand and reserved pricing options and asks customers to contact it to lock in guaranteed resources [33]. Together AI posts preemptible, on-demand and reserved prices for terms of 3 to 180 days, with longer terms marked "Contact us" [35].

Cluster software is priced differently too. Nebius lists its managed Kubernetes service as free, and Crusoe charges $0.10 per cluster hour for managed Kubernetes [30, 33].

2. Posted on-demand prices on one basis

Provider (segment)	Product, region	Posted price	Per GPU-hour
AWS (hyperscaler)	p5.48xlarge, US East (N. Virginia)	$55.04 per instance-hour [9]	$6.88
Microsoft Azure (hyperscaler)	ND96isr H100 v5, East US	$98.32 per hour [15]	$12.29
Google Cloud (hyperscaler)	a3-highgpu-8g, Iowa (us-central1)	$88.490000119 per hour [18]	$11.06
Oracle (hyperscaler, by this report's rule)	BM.GPU.H100.8, part B98415, no region	$10 per GPU-hour [22, 23]	$10.00

Provider (segment)	Product, region	Posted price	Per GPU-hour
CoreWeave (neocloud)	HGX H100, North America	$49.24 per hour [25]	$6.155
Lambda (neocloud)	8x H100, region not stated	$3.99 on 25 October, $4.59 from 16 November [27, 29]	$3.99 or $4.59
Nebius (neocloud)	HGX H100, eu-north1	$4.50 per GPU-hour since 1 October [30, 31, 32]	$4.50
Crusoe (neocloud)	H100 HGX, U.S. and Iceland zones	$3.90, $4.50 from 1 December [33, 34]	$3.90 or $4.50
Together AI (neocloud)	HGX H100 GPU Clusters, region not stated	$3.99 per GPU-hour [35]	$3.99

Table 1. 8-GPU H100 on-demand prices posted on 7 November 2026, except where dated. Per GPU-hour figures computed for this report.

The ranges. On 7 November 2026 the hyperscalers posted $6.88 to $12.29 per GPU-hour and the neoclouds $3.90 to $6.16 (computed for this report). The two ranges did not overlap in this sample. The medians, on samples of four and five, were $10.53 and $3.99, a ratio of about 2.6 (computed for this report). With Lambda's 16 November price in place of its 25 October one, the neocloud median is $4.50. The medians describe this sample only.

The spread within each segment. Azure's on-demand price was 1.79 times AWS's, and CoreWeave's was 1.58 times Crusoe's (computed for this report). Across segments, AWS sat 12% above CoreWeave, while Azure was 3.2 times Crusoe.

Region. The rows do not all share a region. Nebius documents its H100 only in eu-north1 [32], Lambda and Together AI state no region [27, 35] and Oracle's price record carries none [22]. CoreWeave posts the same $49.24 in its North America and Europe tables, with spot at $19.71 and $19.51 [25]. Silicon Data reports that the multiple between hyperscaler and other H100 rates in its own data runs from about 1.6 times in the Nordics to about 3 times in US West and the Middle East [7]. Region alone can move a cross-segment comparison.

Dates. On 7 November, Lambda's page showed prices that take effect on 16 November 2026 and asked visitors to log in to see current pricing [27]. Its 25 October capture showed $3.99 [29].

Units. CoreWeave's table carries $6.16 in a column labelled for single-GPU inference prices, and its footnote says GPU-based pricing "is exclusively applicable to the CoreWeave inference platform customers" [25]. This report uses the $49.24 node price, which is $6.155 per GPU-hour (computed for this report).

3. What the hourly price includes

Provider	vCPUs per GPU	Host memory per GPU	Local storage per GPU	Scale-out network	Internet egress
AWS p5.48xlarge	24	256 GiB	3.84 TB NVMe	3,200 Gbps EFA per instance, GPUDirect RDMA	100 GB a month free, then $0.09 per GB for the next 10 TB [9]
Azure ND96isr H100 v5	12	237.5 GiB	3.5 TiB NVMe (28 TiB per VM, read as a total), plus 128 GiB temporary disk	400 Gbps InfiniBand per GPU	First 100 GB free, then $0.087 per GB [15]
Google a3-highgpu-8g	26	234 GB	750 GiB Local SSD	Up to 1,000 Gbps per VM, GPUDirect-TCPX	$0.12 per GiB to 1 TiB, then $0.11 and $0.08 [21]
Oracle BM.GPU.H100.8	28	256 GB	7.68 TB NVMe	8 x 2 x 200 Gbps RDMA	First 10 TB a month free [24]
CoreWeave HGX H100	16	256 GB	7.68 TB raw, 3.84 TB usable	400G NDR InfiniBand fabric	Storage has no egress fees [25]
Lambda 8x H100	26	225 GiB	2.75 TiB	Not stated for single instances	No egress fees [27]
Nebius HGX H100	16	200 GiB	None bundled	400 Gbps InfiniBand (8× ConnectX-7)	Free [30]
Crusoe H100 HGX	22	120 GB	960 GB NVMe, ephemeral	3,200 Gbps InfiniBand	No ingress or egress fees "at this time" [33]
Together AI H100 cluster	Not stated	Not stated	Not stated	Not stated	No price found

Table 2. What each on-demand node carries. vCPUs, host memory and local storage per GPU, as instance totals divided by eight. Network and egress as each provider states them (computed for this report from [9, 10, 15, 17, 19, 21, 23, 24, 25, 26, 27, 30, 32, 33, 34, 35]).

Host resources. Host memory per GPU ran from 120 GB at Crusoe to 256 GiB at AWS and 256 GB at Oracle and CoreWeave, and vCPUs from 12 at Azure to 28 at Oracle [10, 17, 23, 26, 34]. Oracle counts one x86 OCPU as two vCPUs [23]. Google bills A3 machines for their attached GPUs, predefined vCPUs, memory and bundled Local SSD, and bills Persistent Disk, Hyperdisk and network usage separately [18, 19].

Storage. CoreWeave's pricing page lists 61.44 TB of local storage per node [25]. Its documentation configures eight 7.68 TB drives as RAID 10 and gives 30.72 TB as usable, noting that usable storage is less than raw capacity under RAID [26]. Microsoft's size page lists eight NVMe disks and an NVMe size of 28 TiB without saying whether that is per disk or in total [17], and this report reads it as the total. Nebius bundles no local storage with its GPU price and lists local SSD block volumes at $0.065 per GiB a month [30]. Adding the 3.84 TB per GPU that AWS bundles, at that Nebius rate, would cost about $232 a month per GPU, or about $0.32 per GPU-hour over 730 hours, about 7% of Nebius's on-demand price (computed for this report from [9, 30]).

Egress. Take one node running all 730 hours of a month and sending 10 TB, counted as 10,000 GB, to the internet from one account in a U.S. region. At AWS, after the free 100 GB, about 9,900 GB at $0.09 costs about $891, or about $0.15 per GPU-hour over 5,840 GPU-hours, 2.2% of the on-demand price [9]. Under Google's Premium Tier, the default for all its data transfer, from a North American region to North America, the same volume is about 9,313 GiB and costs about $1,035, about $0.18 per GPU-hour and 1.6% of its on-demand price [21]. Oracle charges nothing for the first 10 TB a month [24], and Lambda, Nebius and Crusoe post no egress fees [27, 30, 33]. Both egress costs are computed for this report. CoreWeave's pricing page says its storage carries no ingress, egress or transfer fees [25], and its 10-K describes a Zero Egress Migration program that can reduce or eliminate data transfer or egress fees [1]. The CMA found egress fees "a key commercial barrier", highest relative to total cloud spending for smaller customers and those with large proportions of stored data [3].

4. Pricing types each segment publishes

Provider	Rate	Per GPU-hour	Against own on-demand
AWS	Capacity Block, US East (N. Virginia), paid up front [11]	$5.970	13% below
AWS	Spot, us-east-1, about 09:07 UTC [12]	$2.66	61% below
Azure	Reservation, 1 year [15]	$7.87	36% below
Azure	Reservation, 3 years [15]	$5.40	56% below
Azure	Reservation, 5 years [15]	$4.92	60% below

Provider	Rate	Per GPU-hour	Against own on-demand
Azure	Spot [15]	$2.27	82% below
Google	DWS Flex-start, up to seven days [18, 20]	$4.79	57% below
Google	DWS Calendar mode [18]	$5.20	53% below
Google	Spot [18]	$6.59	40% below
Google	Resource CUD, 1 year [18]	$7.67	31% below
Google	Resource CUD, 3 years [18]	$4.86	56% below
CoreWeave	Spot, North America [25]	$2.46	60% below
CoreWeave	Committed [25]	"Up to 60%" off, priced by sales	
Lambda	1-Click Clusters, two weeks to one year, 16, 64 and 256 GPUs [28]	$6.16, $5.85, $5.54	Above
Nebius	Preemptible, dynamic [30]	"From $0.79"	
Together AI	Preemptible, reserved 3 to 30, 31 to 90 and 91 to 180 days [35]	$1.99, $3.69, $3.45, $3.19	50%, 8%, 14% and 20% below

Table 3. Committed, reserved and interruptible rates on the basis of Table 1. Per GPU-hour figures and percentages computed for this report, Azure's at 8,760 hours a year.

Hyperscaler commitments sit inside the neocloud on-demand range. Azure's three- and five-year reservations, Google's three-year committed use rate, Flex-start and calendar-mode rates, and AWS's Capacity Block all fall between $3.90 and $6.16 per GPU-hour, the range of neocloud on-demand prices on the same day (computed for this report). Azure's one-year reservation ($7.87) and Google's one-year committed use rate ($7.67) sit above it. A fair pair is term for term.

Neocloud commitments are mostly unpriced in public. Of the five neoclouds, only Lambda and Together AI post prices for fixed terms, Lambda for clusters of two weeks to one year and Together AI for 3 to 180 days [28, 35]. CoreWeave states only a ceiling for committed discounts [25], Nebius describes commitment

discounts without a price [30], and Crusoe asks customers to contact it for guaranteed resources [33].

Stated ceilings are not prices. AWS states savings of up to 72% against on-demand for Savings Plans [13] and Spot discounts of up to 90% [12]. Azure's Retail Prices API returns Microsoft retail prices without discount [16]. The Savings Plans rows for the same AWS instance are set out in What posted GPU cloud prices tell you, and what they do not.

Interruptible prices move on their own clocks. AWS sets Spot prices that adjust gradually with long-term supply and demand, updates its table every 5 minutes, and shows "the lowest prices per instance type in the region" [12]. Nebius says its spot prices are dynamic and may change as often as every 15 minutes [30]. Google says its DWS prices are variable but updated infrequently [18]. On 7 November, Google's Spot rate for the 8-GPU A3 sat above its own Flex-start and three-year rates [18]. Spot rates at AWS, Azure and CoreWeave, from $2.27 to $2.66, were close to one another across segments (computed for this report from [12, 15, 25]). What interruptions cost a training run is covered in What research shows about training on spot GPUs (https://rillor.com/insights/spot-gpu-training-research).

5. Cluster networking is part of the price

Lambda posts two products with the same GPU. Its single 8-GPU H100 instance costs $3.99 per GPU-hour in the 25 October capture and $4.59 from 16 November [27, 29], and its page states no scale-out network for it [27]. Its 1-Click Clusters are dedicated and InfiniBand-connected, for two weeks to one year, at $6.16 per GPU-hour for 16 GPUs, $5.85 for 64 and $5.54 for 256 [28]. The cluster prices are 21% to 54% above the instance prices, and close to CoreWeave's node at $6.155, whose nodes are linked with a 400G NDR InfiniBand fabric [25, 26], and AWS's p5.48xlarge at $6.88, with 3,200 Gbps of EFA and GPUDirect RDMA [9, 10] (computed for this report).

The other nodes describe their networks in their own terms. Azure gives each GPU a dedicated 400 Gbps Quantum-2 CX7 InfiniBand connection, 3.2 Tbps per VM [17]. Oracle lists 8 x 2 x 200 Gbps of RDMA [23]. Crusoe lists 3,200 Gbps of InfiniBand, with 175 Gbps of VPC networking [34]. Nebius lists InfiniBand at "400 Gbps (8× ConnectX-7)" [32]. Google lists up to 1,000 Gbps per VM and GPUDirect-TCPX, which moves packet payloads directly from GPU memory to the network interface [19].

These are specifications. No benchmark reviewed shows that the nodes compared deliver the same training or inference throughput.

6. Posted prices moved, and not in step

Three neoclouds posted on-demand H100 rises within two months. Nebius showed $3.85 on 15 September, showed $3.85 beside $4.50 "Effective October 1, 2026" on 1 October, and showed $4.50 on 7 November, a rise of 16.9% [30, 31]. Lambda's 8-GPU price rises from $3.99 in the 25 October capture to $4.59 from 16 November, 15.0%, and its single H100 SXM is posted at $4.89 from that date [27, 29]. Crusoe's table shows its

H100 at $3.90 and $4.50 from 1 December, 15.4% [33]. The same Crusoe table moves the H200 from $4.29 to $5.40 and the B300 from $8.96 to $9.90, posts the B200 at $8.50 where it said "Contact sales", and leaves the A100 SXM unchanged [33]. All percentages are computed for this report.

Nebius's preemptible price changed form in the same period. Its 15 September capture showed $2.15, and on 7 November the page showed "from $0.79", with a note that spot prices are dynamic [30, 31].

On the hyperscaler side, AWS announced a 44% reduction in P5 on-demand prices against 31 May 2025 baseline prices, applying to on-demand purchases from 1 June 2025 [14]. Its p5.48xlarge price on 7 November 2026 was $55.04 an hour [9]. AWS says Capacity Block prices are updated regularly with supply and demand trends, and that the current prices are scheduled to be updated next in January 2027 [11].

Silicon Data, an index publisher, reported on 31 July 2026 that hyperscaler H100 rates in its data fell 26% to 54% by region since the first quarter of 2025, while other providers' rates recovered, compressing the hyperscaler premium from roughly 250% to about 120% across the seven regions it tracks [7]. It puts the multiple at roughly 2 to 2.5 times as of the third quarter of 2026 [7]. It also writes that its quarterly medians cannot distinguish list-price normalization from shifts in the mix of contracts and configurations behind each median [7]. All three neocloud rises in this report took effect, or are dated to take effect, after that article.

7. Where the posted gap comes from

The posted pages show five parts of the distance between the segments' on-demand prices.

- **Term.** Multi-year hyperscaler commitments fall inside the neocloud on-demand range (Finding 4).
- **Bundle.** A GPU-hour buys different host memory, vCPUs, storage and egress terms (Finding 3). In the examples, storage and egress move the comparison by about 1.6% to 7% (computed for this report).
- **Network.** At the one neocloud that posts both, a cluster costs more than an instance (Finding 5).
- **Region.** One neocloud row is in a European region, and three rows carry no region (Finding 2). In Silicon Data's data, the cross-segment multiple varies by region from about 1.6 to about 3 [7].
- **Date.** Nebius's price rose on 1 October, and Lambda and Crusoe post rises from 16 November and 1 December 2026 (Finding 6), so the same comparison made a month apart gives a different answer.

Other parts are not priced on the pages. Silicon Data writes that hyperscaler rates bundle value that other providers typically do not carry: managed orchestration and the surrounding platform services, enterprise support agreements and service level agreements, and certified compliance regimes [7]. This report did not test that view and opened no service level, support or compliance page. The CMA's account of the specialists, with their narrower customer group and more limited cluster management [4], and its finding that large public clouds in some instances buy bare metal from other cloud providers [4], show that the segments also serve different buyers and sometimes each other.

Both segments sell under negotiated commitments. The CMA found committed spend agreements widespread [3]. CoreWeave says customers generally buy multi-year take-or-pay contracts [1] and offers up to 60% off on-demand for committed usage [25]. SemiAnalysis describes, in one provider's negotiations, prepayments that reach 100% on one-year commitments, and writes that smaller labs get stuck with large prepayments and worse prices than profitable ones [6]. No source opened for this report shows the contract prices themselves, so the gap that matters to a large buyer cannot be measured from posted prices.

Implications

For engineers and buyers comparing offers. Compare on-demand with on-demand and a three-year term with a three-year term (Finding 4). Ask what a node includes per GPU: host memory, vCPUs, local storage as raw and usable capacity, and the scale-out network (Findings 3 and 5). Price egress at the expected monthly volume, since it ranges from free to $0.12 per GiB [9, 21, 24, 27, 30, 33]. For multi-node training, set cluster prices beside cluster prices (Finding 5). Date every price and check its region, because three neoclouds posted H100 rises effective from 1 October to 1 December 2026 and one row sat in Europe [27, 30, 32, 33].

For finance and procurement teams. Posted prices are listings. Microsoft's retail API returns prices "without discount" [16], CoreWeave states committed discounts only as a ceiling [25], and committed spend agreements are widespread [3]. A segment median built from posted prices says little about what one contract costs. Prepayment and take-or-pay terms also change the effective price, as Neocloud economics: contract first, hardware second (https://rillor.com/insights/neocloud-economics-contract-first) sets out.

For analysts building price series. Keep the segments apart, and keep pricing types apart inside each one (Findings 2 and 4). Record each price with its GPU variant, interconnect, GPUs per node, region, term and pricing type, and record the effective date where a provider posts a future price (Finding 6). Check units before dividing: one neocloud's per-GPU figure sat in a column labelled for single-GPU inference (Finding 2). Rillor's GPU and neocloud pricing method keeps hyperscalers and neoclouds in separate segments and puts two prices on one basis before they are compared.

Limits of this analysis

- Nine providers, one GPU and one day. The ranges and medians describe this sample, not either segment as a whole.
- All prices are posted list prices. No transaction or contract price was found or used.
- Lambda's price in force on 7 November 2026 was not visible without logging in. The 25 October capture and the price posted for 16 November are used instead [27, 29].
- Regions differ. Nebius's H100 is documented only in eu-north1 [32], and Lambda, Together AI and Oracle state no region for their prices [22, 27, 35].
- Azure's reservation totals were converted at an assumed 8,760 hours a year. Its NVMe size was read as a total, which its page does not confirm [17].
- Network figures come from specifications. No throughput benchmark was run or reviewed.
- No service level, support or compliance terms were reviewed.
- No egress price was found for Together AI. Crusoe says it does not charge for egress "at this time" [33], and CoreWeave's no-egress statement comes from its storage section [25].
- Oracle's and Together AI's placements are classification choices [3, 4]. IBM and Nscale were not priced.
- Silicon Data's and SemiAnalysis's figures are their own and were not re-derived [6, 7]. The prepayment remark is anecdotal.
- The egress and storage examples assume decimal terabytes, one account, a U.S. region and 730 hours a month.
- Spot and preemptible prices were captured once, and AWS updates its table every 5 minutes [12].

Sources

1. CoreWeave, Inc. Annual Report on Form 10-K for the fiscal year ended December 31, 2025. CoreWeave, Inc., via SEC EDGAR, filed 2 March 2026, viewed 7 November 2026. https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/crwv-20251231.htm

2. NVIDIA Corporation Annual Report on Form 10-K for the fiscal year ended January 25, 2026. NVIDIA Corporation, via SEC EDGAR, filed 25 February 2026, viewed 7 November 2026. https://www.sec.gov/Archives/edgar/data/1045810/000104581026000021/nvda-20260125.htm

3. Cloud services market investigation: Summary of final decision. UK Competition and Markets Authority, 31 July 2025, viewed 7 November 2026. https://assets.publishing.service.gov.uk/media/688b20e6ff8c05468cb7b120/summary_of_final_decision.pdf

4. Cloud services market investigation: Final decision report. UK Competition and Markets Authority, 31 July 2025, viewed 7 November 2026. https://assets.publishing.service.gov.uk/media/688b8891fdde2b8f73469544/final_decision_report.pdf

5. It's the People, Not the Placement: Rethinking Allocations in Post-Moore Clouds. Tejas Harith and Antoine Kaufmann, arXiv:2501.11185, 19 January 2025, viewed 7 November 2026. https://arxiv.org/abs/2501.11185

6. ClusterMAX 3.0: The Industry Standard GPU Cloud Rating System Returns. SemiAnalysis, 23 September 2026, viewed 7 November 2026. https://newsletter.semianalysis.com/p/clustermax-30-the-industry-standard

7. H100 Rental Prices by Region (Q3 2026): The Hyperscaler Premium Has Halved. Yang Gao, Silicon Data, 31 July 2026, viewed 7 November 2026. https://www.silicondata.com/blog/h100-rental-prices-by-region-q3-2026

8. NVIDIA H100 Tensor Core GPU. NVIDIA Corporation, viewed 7 November 2026. https://www.nvidia.com/en-us/data-center/h100/

9. EC2 On-Demand Instance Pricing, with the on-demand price file for US East (N. Virginia), Linux, published 7 November 2026, and the data transfer price file, published 3 November 2026. Amazon Web Services, viewed 7 November 2026. https://aws.amazon.com/ec2/pricing/on-demand/ 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 https://b0.p.awsstatic.com/pricing/2.0/meteredUnitMaps/datatransfer/USD/current/datatransfer.json

10. Amazon EC2 P5 Instances. Amazon Web Services, viewed 7 November 2026. https://aws.amazon.com/ec2/instance-types/p5/

11. Amazon EC2 Capacity Blocks for ML Pricing. Amazon Web Services, viewed 7 November 2026. https://aws.amazon.com/ec2/capacityblocks/pricing/

12. Amazon EC2 Spot Instances Pricing, with its Spot price data file, retrieved about 09:07 UTC. Amazon Web Services, viewed 7 November 2026. https://aws.amazon.com/ec2/spot/pricing/ https://website.spot.ec2.aws.a2z.com/spot.json

13. Savings Plans. Amazon Web Services, viewed 7 November 2026. https://aws.amazon.com/savingsplans/

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

15. Azure Retail Prices API responses for Standard_ND96isr_H100_v5 and Bandwidth in East US (api-version 2023-01-01-preview), retrieved about 09:03 UTC. Microsoft, viewed 7 November 2026. https://prices.azure.com/api/retail/prices?api-version=2023-01-01-preview&$filter=armRegionName%20eq%20%27eastus%27%20and%20armSkuName%20eq%20%27Standard_ND96isr_H100_v5%27 https://prices.azure.com/api/retail/prices?api-version=2023-01-01-preview&$filter=serviceName%20eq%20%27Bandwidth%27%20and%20armRegionName%20eq%20%27eastus%27

16. Azure Retail Prices overview. Microsoft Learn, updated 6 August 2024, viewed 7 November 2026. https://learn.microsoft.com/en-us/rest/api/cost-management/retail-prices/azure-retail-prices

17. ND-H100-v5 size series. Microsoft Learn, updated 25 September 2026, viewed 7 November 2026. https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/ndh100v5-series

18. Accelerator-optimized VM Pricing. Google Cloud, viewed 7 November 2026. https://cloud.google.com/products/compute/pricing/accelerator-optimized

19. GPU machines in accelerator-optimized machine family. Google Cloud documentation, last updated 6 November 2026, viewed 7 November 2026. https://docs.cloud.google.com/compute/docs/accelerator-optimized-machines

20. About Flex-start VMs. Google Cloud documentation, last updated 26 October 2026, viewed 7 November 2026. https://docs.cloud.google.com/compute/docs/instances/about-flex-start-vms

21. Network pricing. Google Cloud, viewed 7 November 2026. https://cloud.google.com/vpc/network-pricing

22. Oracle list price API response, all products in USD, last updated 2 November 2026, with the single-part request for B98415. Oracle, viewed 7 November 2026. https://apexapps.oracle.com/pls/apex/cetools/api/v1/products/?currencyCode=USD https://apexapps.oracle.com/pls/apex/cetools/api/v1/products/?partNumber=B98415&currencyCode=USD

23. Compute Shapes. Oracle Cloud Infrastructure documentation, Oracle, viewed 7 November 2026. https://docs.oracle.com/en-us/iaas/Content/Compute/References/computeshapes.htm

24. Oracle Cloud Price List, networking section. Oracle, viewed 7 November 2026. https://www.oracle.com/cloud/price-list/#pricing-networking

25. CoreWeave Cloud Pricing. CoreWeave, viewed 7 November 2026. https://www.coreweave.com/pricing

26. H100 (InfiniBand), CoreWeave Docs. CoreWeave, last modified 23 June 2026, viewed 7 November 2026. https://docs.coreweave.com/platform/instances/gpu/gd-8xh100ib-i128

27. Instances, On-Demand Cloud. Lambda, viewed 7 November 2026. https://lambda.com/instances

28. 1-Click Clusters. Lambda, viewed 7 November 2026. https://lambda.com/1-click-clusters

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

30. NVIDIA GPU Pricing, Nebius AI Cloud. Nebius, viewed 7 November 2026. https://nebius.com/prices

31. Nebius prices, archived captures of 15 September 2026 and 1 October 2026. Nebius, via the Internet Archive Wayback Machine, viewed 7 November 2026. https://web.archive.org/web/20260915165115/https://nebius.com/prices https://web.archive.org/web/20261001131140/https://nebius.com/prices

32. Types of virtual machines and GPUs in Nebius AI Cloud. Nebius documentation, viewed 7 November 2026. https://docs.nebius.com/compute/virtual-machines/types

33. GPU pricing for AI compute and inference, Crusoe Cloud pricing. Crusoe, viewed 7 November 2026. https://www.crusoe.ai/cloud/pricing

34. Overview, Crusoe Cloud virtual machines. Crusoe Cloud documentation, viewed 7 November 2026. https://docs.crusoe.ai/compute/virtual-machines/overview/index.html

35. Pricing. Together AI, viewed 7 November 2026. https://www.together.ai/pricing

About this report

Published by Rillor on 8 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. Neocloud and hyperscaler GPU prices, compared like for like. 8 November 2026. https://rillor.com/insights/neocloud-vs-hyperscaler-gpu-pricing

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.

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