GPU Cloud Rental Prices 2026: H100 vs H200 Cost Comparison

Published July 10, 2026 37 min read

Executive Summary

As of July 2026, renting an **NVIDIA H100** GPU in the cloud costs between roughly $1.38 and $6.98 per GPU-hour depending on provider, with a cross-provider median near $2.99/hr, while the newer **H200** (141GB HBM3e memory) runs from about $2.30 to $13.78/hr with a median around $4.00 (Source: aimultiple.com) (Source: aimultiple.com). The single biggest driver of price is provider category, not the chip itself: specialized "neocloud" providers such as **RunPod** ($1.99 to $2.39/hr for H100 PCIe) (Source: runpod.io), **Lambda** ($3.99/hr on-demand H100 SXM) (Source: lambda.ai), **Vast.ai** (from $1.38/hr on a peer-to-peer marketplace), and **Nebius** ($3.85/hr H100, $4.50/hr H200 as of June 2026) (Source: docs.nebius.com) undercut the three major hyperscalers by roughly 40% to 400%. **Microsoft Azure** lists its NC40ads H100 v5 instance at $6.98/hr on-demand (Source: instances.vantage.sh), **AWS** prices its p5.4xlarge (1x H100) at $6.88/hr (Source: getdeploying.com), and **Google Cloud**'s A3 Mega H100 instance runs $93.40/hr for 8 GPUs, or roughly $11.68/GPU-hour (Source: cloud.google.com).

The pricing gap between "neocloud" specialists and hyperscalers reflects fundamentally different business models rather than different hardware: the underlying silicon is identical NVIDIA Hopper-generation compute in both cases. Analysis from **Spheron** found that "an H100 on Spheron starts at $1.33/hr compared to $6.98/hr on Azure and $7.00/hr on AWS (as of Q1 2026), that's 40 to 60% less than hyperscalers" (Source: spheron.network). The independent research firm **SemiAnalysis** documents a separate and more volatile trend in the committed-capacity market: 1-year H100 rental contract pricing "shot up almost 40% from a low of $1.70/hr/GPU in October 2025 to $2.35/hr/GPU by March 2026" (Source: newsletter.semianalysis.com), driven by surging inference demand from multi-agent AI workloads even as on-demand spot pricing has continued to soften.

On the H100 versus H200 question specifically: the H200 costs roughly 20% to 40% more per GPU-hour than the H100 at most providers, in exchange for 141GB of HBM3e memory (versus 80GB on the H100) and 4.8 terabytes per second of memory bandwidth, "nearly double the capacity of the NVIDIA H100 GPU... with 1.4X more memory bandwidth" (Source: nvidia.com). That extra memory matters concretely: workloads such as **Meta's Llama 4 Maverick** (400 billion parameters) that need two full 8-GPU H100 nodes fit on a single 8-GPU H200 node (Source: gmicloud.ai). For teams that do not need the extra memory headroom, on-demand H100 capacity remains the more cost-efficient default; for large-context inference and models above roughly 70 billion parameters, the H200's memory bandwidth advantage typically pays for its price premium.

Buyers evaluating GPU cloud rental prices should treat "H100 pricing" and "H200 pricing" as ranges bounded by provider category rather than single numbers: marketplace and neocloud on-demand rates cluster in the $1.38 to $4.50 range for both chips, while hyperscaler on-demand list prices run $6.88 to over $13 per GPU-hour (Source: getdeploying.com). Reserved and multi-year contracts, spot/preemptible instances, and per-second versus per-hour billing further widen the effective spread, and this report walks through each of these levers in detail, alongside named deployments from **CoreWeave**, **Corvex**, and **Mistral AI** that illustrate how enterprises are actually buying this capacity in mid-2026.

Introduction and Background

The market for renting **NVIDIA H100** and **H200** graphics processing units (GPUs) by the hour has become one of the most closely watched line items in enterprise AI budgets. Both chips are built on NVIDIA's Hopper architecture and remain the dominant workhorses for large language model (LLM) training and inference as of mid-2026, even as newer Blackwell-generation chips (B200, B300) enter the market at a premium. Understanding **gpu cloud rental prices** requires separating three distinct markets that are frequently conflated in headline figures: on-demand hourly rental from specialized GPU clouds ("neoclouds"), on-demand rental from the three major hyperscalers (AWS, Microsoft Azure, Google Cloud), and longer-term reserved or contract pricing negotiated directly with providers.

The H100, launched in 2022, ships in **SXM** and **PCIe** form factors with 80GB of HBM3 memory and up to 3.35 terabytes per second (TB/s) of memory bandwidth per GPU ("3 terabytes per second (TB/s) of memory bandwidth per GPU" per NVIDIA's own architecture documentation) (Source: nvidia.com). The H200, which began broad cloud availability in 2024, is built on the same Hopper compute die but introduces HBM3e memory, expanding capacity to 141GB and bandwidth to 4.8 TB/s (Source: nvidia.com). Because compute throughput (measured in FP8 TFLOPS) is nearly identical between the two chips, the H200's price premium over the H100 is almost entirely a function of its larger, faster memory pool.

Pricing volatility has been the defining story of this market since late 2025. **SemiAnalysis**, an independent semiconductor and AI infrastructure research firm, reports that "on-demand GPU rental capacity is sold out across all GPU types" as of early 2026, with providers unwilling to relinquish locked-up on-demand instances even as list prices rise (Source: newsletter.semianalysis.com). At the same time, the **AIMultiple Cloud GPU Rental Price Index**, which tracks 63 providers and 17 GPU models monthly, finds that posted on-demand median pricing for mainstream Hopper-generation cards has held in a relatively tight band even as newer Blackwell cards saw sharper increases (Source: aimultiple.com). These two data sources are not contradictory: they measure different segments of the same market, on-demand posted list pricing versus negotiated committed-capacity contracts, and the divergence between them is itself one of the most important facts a buyer needs to understand before committing budget.

This report answers the question "how much does it cost to rent an H100 or H200 GPU in the cloud" by walking through pricing at every major provider category, comparing H100 and H200 economics head to head, presenting the underlying market data in aggregate, and reviewing named real-world deployments. All prices are anchored "as of July 2026" unless a different date is explicitly cited, since posted GPU cloud rates have historically moved by double-digit percentages within a single quarter.

Buyers researching **nvidia h100 price per hour** or **h200 gpu rental cost** figures online will frequently encounter conflicting numbers from different sources, and this is rarely a sign that one source is wrong. A single provider can post several different rates for the same chip depending on form factor (SXM versus PCIe), billing granularity (per-second versus per-hour minimums), cluster size (single GPU versus multi-node), and commitment length (on-demand, 1-month, 1-year, or 3-year). GetDeploying's own methodology note is instructive here: it tracks "46 cloud providers for the H100," with "the pricing spread is significant. While the average sits at $3.49/hr, the lowest price is currently $0.61/hr per GPU (spot instance)" (Source: getdeploying.com), a nearly 6x spread within a single tracker's dataset. The same tracker finds an even wider spread for H200: "there is a 92% difference between the highest and lowest listings for the H200. You can pay up to $13.78/hr, but the market floor is currently $1.00/hr per GPU (spot instance)" (Source: getdeploying.com). This report cites specific figures with their source and access conditions attached, rather than collapsing the market into a single average, because the average obscures the decision that actually matters to most buyers: which provider category fits a given workload's tolerance for interruption, its need for guaranteed capacity, and its budget ceiling.

Amazon Web Services (AWS)

Capabilities

AWS offers H100 access through its **EC2 P5 instance family**, with three variants: p5 (standard H100 SXM5, 80GB), p5e, and p5en (both using an "H100e" variant with 192GB HBM3e per GPU) (Source: spheron.network). On-demand pricing for a single-GPU p5.4xlarge instance is $6.88/hr as tracked live by GetDeploying (Source: getdeploying.com), while the full 8-GPU p5.48xlarge instance lists at $55.04/hr total, the same $6.88/GPU-hour rate scaled to 8 GPUs. AWS also sells **Capacity Blocks for ML**, a reservation product for guaranteed short-term access: a single H100 (p5.4xlarge) Capacity Block in US East (N. Virginia) is priced at $5.191/hr per accelerator, and an 8x H200 (p5e.48xlarge) Capacity Block runs $47.76/hr total, or $5.97/GPU-hour (Source: aws.amazon.com).

Adoption

AWS cut P5 on-demand pricing by 44% in June 2025, a move widely covered in trade press as a response to intensifying competition from neocloud providers (Source: www.spheron.network). Despite the cut, AWS remains the default choice for organizations already standardized on AWS services such as SageMaker, Bedrock, or EFA-optimized EKS clusters, where migrating workloads to a neocloud carries real integration costs.

Strengths and Limitations

AWS spot pricing for P5 instances is "structurally scarce," discounting roughly 44% off on-demand (to about $3.83/GPU-hour) when available, but AWS prioritizes on-demand and reserved customers for P5 allocation, so spot pools "fill only with what remains" (Source: www.spheron.network). Hidden costs compound the headline rate: EBS storage adds $40 to $80/month for a typical checkpoint volume, data egress runs $0.09/GB, and cross-AZ transfer adds $0.02/GB in each direction. Three-year reserved AWS pricing can reach roughly $1.90/hr per H100, competitive with neocloud on-demand rates, but only for buyers able to commit multi-year capacity (Source: awesomeagents.ai).

AWS's own EC2 On-Demand pricing page notes that "each partial instance-hour consumed will be billed per-second for Linux, Windows, Windows with SQL Enterprise, Windows with SQL Standard, and Windows with SQL Web Instances" (Source: aws.amazon.com), a billing-granularity detail that matters for short-lived jobs since it eliminates the round-up-to-the-hour waste common on older cloud billing models. New AWS accounts additionally start with a default P-instance vCPU quota of zero, and because each p5.48xlarge instance consumes 192 vCPUs, first-time GPU renters must request and receive a quota increase, a process that can take several business days with a written business justification, before they can launch any P5 capacity at all. This procurement friction is a real, if non-monetary, cost that does not appear in any hourly rate comparison but materially affects how quickly a team can actually access AWS H100 capacity relative to a neocloud sign-up flow that typically provisions in minutes.

Microsoft Azure

Capabilities

Azure's flagship H100 offering is the **NC40ads H100 v5** virtual machine series, which pairs a fractional or full H100 GPU allocation with up to 40 vCPUs and 320 GiB of memory. On-demand pricing starts at $6.98/hr, with spot pricing available at $1.40298/hr (Source: instances.vantage.sh). Azure also offers the **ND H100 v5** series for full 8-GPU nodes targeted at large-scale training. For H200, Microsoft's ND H200 v5 SKU is sold primarily through direct sales contact rather than posted self-serve pricing (Source: awesomeagents.ai).

Adoption

Azure carries the highest posted on-demand H100 list price among the three hyperscalers reviewed in this report, and multiple independent trackers flag it as the ceiling of the market: AIMultiple notes that "Microsoft Azure and Google Cloud carry the upper tail past $10" for H100 pricing (Source: aimultiple.com), and for H200 specifically, GMI Cloud's provider comparison lists Azure at the top of its range at $13.78/GPU-hour (Source: www.gmicloud.ai). A separate live-pricing table normalizing "NC40ads_H100_v5 VM" to a "Single H100 GPU VM in East US region" lists the rate at $8.30/hr, moderately above the $6.98/hr figure Vantage tracks, illustrating how Azure's regional and configuration variance alone can shift the quoted H100 rate by close to $1.30/hr depending on which specific SKU and region a source samples (Source: www.thundercompute.com).

Strengths and Limitations

Azure's value proposition rests less on raw GPU-hour price and more on integration with the Microsoft ecosystem, Azure OpenAI Service, Azure Machine Learning, and Windows-dependent workloads (Source: awesomeagents.ai). One notable exception to Azure's premium pricing pattern is its A100 spot tier, which at roughly $0.74/hr is "among the cheapest A100 availability you will find" across any provider (Source: awesomeagents.ai), suggesting Azure's pricing strategy varies significantly by GPU generation and instance tier rather than applying a uniform premium.

Azure's own documentation describes the ND H100 v5 series as "a new flagship addition to the Azure GPU family," designed "for high-end Deep Learning training and tightly coupled scale-up and scale-out Generative AI and HPC workloads" (Source: learn.microsoft.com). Each ND H100 v5 GPU carries "its own dedicated, topology-agnostic 400 Gb/s NVIDIA Quantum-2 CX7 InfiniBand connection," and deployments "can scale up to thousands of GPUs with 3.2 Tbps of interconnect bandwidth per VM" (Source: learn.microsoft.com), positioning the SKU specifically toward large-scale multi-node training clusters where networking quality, not just per-GPU price, determines total training cost.

Google Cloud Platform (GCP)

Capabilities

Google Cloud sells H100 access through its **A3** family (A3 High and A3 Mega, using `a3-highgpu-8g` and `a3-megagpu-8g` machine types) and H200 through its **A3 Ultra** family (`a3-ultragpu-8g`). Per Google's own pricing page, the A3 Mega 8-GPU H100 configuration lists at $93.400712807/hour on-demand, or roughly $11.68 per GPU-hour, while the A3 Ultra 8-GPU H200 configuration lists at $84.806908493/hour on-demand, or roughly $10.60 per GPU-hour (Source: cloud.google.com). A3 High, the lower-networking variant, is somewhat cheaper at $88.49/hour for 8 H100 GPUs. GCP also lists sustained-use and 1- and 3-year committed-use discount tiers directly on the same pricing page, with 3-year commitments cutting the A3 Mega H100 rate to roughly $40.65/hour for 8 GPUs.

Adoption

AIMultiple's index flags a methodological wrinkle worth noting for buyers comparing sources: "Google Cloud added its A3z Mega H100 variant to the standard-A3 listing, lifting the H100 cohort median from ~$2 to ~$3," and separately observes that "the Google Cloud row is itself a mix of three SKUs (a3-highgpu, a3-megagpu, a3-edgegpu) collapsed under one nvidia-h100 label, which lifts its cohort median" (Source: aimultiple.com). This is a useful caution: headline "GCP H100 price" figures can vary substantially depending on which A3 SKU a given source is quoting.

Strengths and Limitations

GCP's A3 instances are frequently cited in community discussion as carrying a steep premium relative to AWS for equivalent hardware, reflecting a broader pattern where GCP's list pricing runs meaningfully above AWS for comparable configurations, though GCP's committed-use discounts can substantially close that gap for predictable, sustained workloads. GCP's own pricing page shows this discount structure directly: the A3 Mega 8x H100 configuration drops from $93.40/hour on-demand to $50.79/hour under a 1-year commitment and to $40.65/hour under a 3-year commitment (Source: cloud.google.com), a roughly 57% reduction from list price, which brings GCP's effective 3-year H100 rate to roughly $5.08/GPU-hour, closer to (though still above) the neocloud on-demand range.

GCP's principal strength is native integration with Vertex AI, BigQuery, and the broader Google Cloud data and machine learning stack, which matters for teams whose data pipelines already live in GCP. Its principal limitation for pure GPU-rental cost efficiency is that A3's smallest published unit is an 8-GPU node; GCP does not offer a self-serve single-GPU A3 instance comparable to AWS's p5.4xlarge or Azure's NC40ads, so teams needing only one or two GPUs for experimentation typically pay for capacity they do not use unless they turn to GCP's older, smaller A2 (A100) instance family instead. A separate cross-provider comparison normalizing GCP's single-GPU equivalent rate lists "a3-highgpu-1g" at "$11.06" per hour in US-central, on-demand, corroborating that GCP anchors the expensive end of the H100 hyperscaler range even before accounting for its 8-GPU minimum footprint (Source: www.thundercompute.com).

Neocloud and Specialized GPU Providers

Capabilities

The "neocloud" category, providers built specifically around GPU rental rather than general-purpose cloud services, spans a wide range of pricing models. **RunPod** offers H100 PCIe "from $1.99/hr on Community Cloud and $2.39/hr on Secure Cloud" (Source: www.runpod.io), with H200 SXM starting at $3.59/hr on Community Cloud (Source: getdeploying.com). **Lambda** prices on-demand H100 SXM at $3.99 to $4.29/hr depending on cluster size and region, with its 1-Click Cluster product offering $6.16/hr per GPU for a 16-GPU reserved H100 cluster on a 2-week to 1-year term, dropping to $5.54/hr at 256-GPU scale (Source: lambda.ai). **Vast.ai**, a peer-to-peer GPU marketplace, rents H200 GPUs "for $3.75/hr" (Source: vast.ai) and lists H100 on-demand pricing in the $1.38 to $1.87/hr range according to independent tracking (Source: awesomeagents.ai). **CoreWeave** lists an 8-GPU HGX H100 node at $49.24/hr on-demand ($6.16/GPU-hour) with spot pricing at $19.71/hr ($2.46/GPU-hour), and an 8-GPU HGX H200 node at $50.44/hr on-demand ($6.31/GPU-hour) (Source: www.coreweave.com) (Source: www.coreweave.com). **Nebius** raised prices effective June 1, 2026, moving on-demand H100 NVLink from $2.95/hr to $3.85/hr and H200 NVLink from $3.50/hr to $4.50/hr, while preemptible H100 rose from $1.25/hr to $2.15/hr (Source: docs.nebius.com). **Jarvislabs** advertises H100 at $2.69/hr and H200 at $3.80/hr, described independently as an "under-the-radar provider with competitive pricing" offering "under-90-second instance spin-up" and per-minute billing (Source: awesomeagents.ai).

Adoption

Spheron, another neocloud entrant, prices H100 spot at $1.66/hr and on-demand at $2.64/hr per GPU, versus AWS's $55.04/hr for an equivalent 8-GPU node (Source: www.spheron.network). Running an 8x H100 training job for 720 hours (roughly one month), AWS on-demand costs $39,628, AWS's deepest 3-year reserved tier costs $17,834, and Spheron's on-demand rate costs $15,206, cheaper than AWS even at AWS's maximum multi-year commitment discount, with Spheron spot bringing the same workload down to roughly $9,562.

Strengths and Limitations

Vast.ai's marketplace model is explicitly described by independent trackers as "the cheapest option on the market, period," but with an explicit trade-off: "hosts can reclaim their machines, instances can disappear mid-job," making it suitable for checkpointed batch work but not recommended for production inference (Source: awesomeagents.ai). CoreWeave's reserved enterprise pricing for H100 can reach roughly $1.45/hr per GPU-hour, "genuinely the cheapest per-GPU-hour for H100 at scale, but only accessible with enterprise commitments" (Source: awesomeagents.ai).

Beyond the providers profiled above, several additional neoclouds round out the competitive set. **TensorDock** operates its own data centers and prices H100 at "~$2.50/hr on-demand, ~$1.60/hr spot," with A100 at roughly $1.30/hr and RTX 4090 at roughly $0.25/hr (Source: awesomeagents.ai). **Salad** taps a network of "60,000+ consumer gaming PCs" and offers RTX 4090 at $0.20/hr, "the cheapest legitimate option" the source identified for that card, though it carries no data-center-grade GPUs and "reliability is consumer-grade: expect occasional node failures" (Source: awesomeagents.ai). Crusoe lists 8x H200 nodes at $34.32/hr total ($4.29/GPU-hour) (Source: getdeploying.com), while Nebius lists H200 at $4.50/GPU-hour on a single-GPU instance through third-party tracking (Source: getdeploying.com). This long tail of smaller providers is precisely why aggregated dashboards report averages substantially above the true achievable floor: a buyer willing to shop across a dozen neoclouds, rather than defaulting to the first familiar name, can typically beat the $3.49/hr H100 average GetDeploying reports by 30% or more.

Serverless and function-call inference platforms occupy a distinct pricing tier worth separating from raw GPU rental. **Modal** bills H100 access at $3.95/hr "per-second on a true serverless model, so idle time costs nothing and bursty inference scales cleanly," while **Fal** rents dedicated H100 compute at $1.89/hr with per-second billing, "near the floor of the table for raw GPU pricing" (Source: comparegpuclouds.com). **Together AI** prices multi-GPU H100 clusters at $5.49/hr per GPU, pairing "cluster compute with their open-model serverless inference stack, so a team training a model can move it to production on the same platform" (Source: comparegpuclouds.com). These platform-layer providers intentionally price above raw-compute neoclouds because the rate bundles auto-scaling, observability, and cold-start management that a self-managed RunPod or Vast.ai deployment would otherwise require an engineering team to build.

Billing Models: On-Demand, Spot, and Reserved Pricing

How a provider bills for GPU time changes the effective price as much as the headline rate does. Three billing models dominate the market. **On-demand** billing charges a fixed rate with no commitment and no interruption risk (outside of provider-side outages), and is the default rate quoted throughout this report unless otherwise noted. **Spot** or **preemptible** billing offers a substantial discount, typically 40% to 60%, in exchange for the provider's right to reclaim the instance with little or no notice when it needs the capacity back; AIMultiple's spot-discount tracker finds that "over the past six months, modern [GPU category] saves ~50%" versus on-demand pricing (Source: aimultiple.com). **Reserved** or **committed-use** billing locks in a lower rate in exchange for a fixed term, typically 1, 3, or 5 years, and is where the largest single discounts appear, GCP's 3-year commitment cuts its A3 Mega H100 rate by roughly 57%, as shown above, and Nebius's preemptible H100 tier prices at $2.15/hr against a $3.85/hr on-demand rate, a 44% discount (Source: docs.nebius.com).

Billing granularity is a related but distinct variable. Several neoclouds, including Spheron and Jarvislabs, bill per minute or per second rather than rounding up to the nearest hour, a meaningful advantage for short, iterative experimentation workloads where a traditional hourly-minimum billing model can silently double the effective cost of a 20-minute test run. Minimum commitment periods also vary: Vast.ai bills per second with no minimum, while some hyperscaler reserved products (such as AWS Capacity Blocks) require fixed-duration blocks rather than allowing early termination. Buyers modeling total cost of ownership should multiply their expected utilization pattern, not just their expected peak GPU count, against each provider's specific billing unit before comparing headline hourly rates, since two providers quoting the same $/hr figure can produce meaningfully different bills once minimum billing increments and idle-time charges are factored in.

Feature Comparison

Table 1 below summarizes on-demand hourly pricing for a single H100 and H200 GPU across the providers examined in this report, drawn directly from each provider's own pricing page or from live third-party trackers that cite the provider's posted rate.

PROVIDER	CATEGORY	H100 ON-DEMAND $/HR	H200 ON-DEMAND $/HR	NOTES
Vast.ai	Marketplace	$1.38 to $1.87 (Source: awesomeagents.ai)	$3.75 (Source: vast.ai)	Per-second billing; reliability varies by host
RunPod	Neocloud	$1.99 to $2.39 (Source: runpod.io)	$3.59 (Source: getdeploying.com)	Community vs. Secure Cloud tiers
Jarvislabs	Neocloud	$2.69 (Source: jarvislabs.ai)	$3.80 (Source: awesomeagents.ai)	Per-minute billing, ~90 second startup
Spheron	Neocloud	$1.33 to $2.64 (Source: spheron.network)	$5.34 spot (Source: spheron.network)	Per-minute billing, no lock-in
Lambda	Neocloud	$3.99 to $4.29 (Source: lambda.ai)	Not listed on-demand single-GPU	$6.16/GPU on 16-GPU 1-Click Cluster (Source: lambda.ai)
Nebius	Neocloud	$3.85 (from Jun 1, 2026) (Source: docs.nebius.com)	$4.50 (from Jun 1, 2026) (Source: docs.nebius.com)	Preemptible H100 $2.15, H200 $2.45
CoreWeave	Neocloud/Hyperscale	$6.16 (8-GPU node) (Source: coreweave.com)	$6.31 (8-GPU node) (Source: coreweave.com)	Spot $2.46/GPU (H100); reserved from ~$1.45/GPU
AWS	Hyperscaler	$6.88 (Source: getdeploying.com)	$5.97 (Capacity Block) (Source: aws.amazon.com)	44% price cut in June 2025 (Source: spheron.network)
Microsoft Azure	Hyperscaler	$6.98 (Source: instances.vantage.sh)	Up to $13.78 (Source: gmicloud.ai)	ND H200 v5 sold via direct sales
Google Cloud	Hyperscaler	~$11.68 (A3 Mega, per-GPU) (Source: cloud.google.com)	~$10.60 (A3 Ultra, per-GPU) (Source: cloud.google.com)	3-year committed use discounts available

Independent trackers not already reflected in Table 1 corroborate this same pattern. Thunder Compute's own July 2026 comparison lists its H100 at "$2.19" as the "lowest fixed-price on-demand rate in this comparison," with Vast.ai at $2.01, CoreWeave's 8-GPU node normalizing to $6.16/GPU, AWS p5.4xlarge at $6.88, Azure's NC40ads at $8.30, and Google Cloud's a3-highgpu-1g at $11.06, the widest single-provider spread in that dataset (Source: www.thundercompute.com). BestGPUCloud's side-by-side comparison separately finds "NVIDIA H100 is the more affordable option, saving you $0.90/hr compared to NVIDIA H200," which "over a month of continuous use" amounts to "approximately $648.00" in savings for a single GPU (Source: www.bestgpucloud.com), a tracker that separately lists best available H100 pricing "from $2.89/hr" against H200 "from $3.79/hr" across its monitored provider set (Source: www.bestgpucloud.com). CompareGPUClouds.com, tracking 12 providers, finds "Baseten" pricing H100 inference at $6.50/hr with the platform layer for auto-scaling and observability built into the rate, while noting that at the low end "three providers sit within twelve cents of each other at the floor," Vast.ai at $1.87/hr, Fal at $1.89/hr, and RunPod at $1.99/hr (Source: comparegpuclouds.com).

This table makes the core finding of this report visible at a glance: the price a buyer pays for an H100 or H200 GPU-hour is determined far more by which category of provider they choose than by the chip itself. A buyer moving from AWS on-demand ($6.88/hr) to RunPod Community Cloud ($1.99/hr) for the same H100 SXM/PCIe silicon can cut compute spend by roughly 71%, before accounting for AWS's additional egress, storage, and support fees. The H100-to-H200 premium, by contrast, is far more consistent across providers, typically 15% to 30% at neoclouds and roughly comparable at hyperscalers, reflecting the fact that H200 is fundamentally the same compute die with more memory rather than a separate architecture generation.

Performance and Benchmarks

Raw hourly price is only half the cost equation; the other half is how much useful work a dollar of GPU-hour buys, which depends on workload type. For LLM inference on models in the 30 to 70 billion parameter range, AIMultiple's provider-workload matrix recommends H100 on a neocloud "for tight latency SLA," reflecting the H100's balance of compute and cost (Source: aimultiple.com). For models above 70 billion parameters that are memory-bound rather than compute-bound, the same analysis recommends H200 or AMD's MI300X specifically because "141-192 GB HBM enables larger KV-cache" (Source: aimultiple.com), a category where the H200's memory-per-dollar, not its raw FP8 throughput, is the deciding factor.

Independent inference benchmarking from **Baseten** on H200 GPUs found meaningful throughput gains over H100 specifically on long-context workloads, aligning with NVIDIA's own claim that H200 delivers "up to 1.6x higher inference performance" relative to H100 on large models (Source: www.nvidia.com). On the compute side, the H100 and H200 share nearly identical tensor core throughput: NVIDIA's own H100 specification sheet lists 3,958 teraFLOPS of FP8 Tensor Core performance for the H100 NVL configuration (Source: www.nvidia.com), essentially matching the H200's listed FP8 Tensor Core performance of 3,958 TFLOPS (Source: www.nvidia.com). This convergence in raw compute confirms that the H200's real-world performance edge for large models comes almost entirely from its memory subsystem rather than from additional arithmetic throughput, which explains why the two chips' cloud rental prices track relatively closely rather than diverging the way H100 and Blackwell-generation B200 pricing does.

For training workloads, CoreWeave reported setting "new AI training records in MLPerf Training v6.0, training DeepSeek-V3 in approximately two minutes," an industry-standard benchmark result that underscores how cluster-scale networking (not just individual GPU specification) materially affects the price-performance a buyer actually realizes on large training runs. This is a critical caveat for readers comparing single-GPU hourly rates: a cheaper GPU-hour on a provider with weaker interconnect (standard Ethernet rather than InfiniBand) can produce a higher total cost per completed training run, since "if you are running a multi-node training job on Llama 3 70B or larger, InfiniBand can cut training time by 50%+ compared to standard Ethernet" (Source: awesomeagents.ai).

Data Analysis and Evidence

Table 2 below summarizes on-demand median pricing across GPU generations, from the prior-generation A100 through the newest Blackwell-generation cards, as tracked by the AIMultiple Cloud GPU Rental Price Index.

GPU	GENERATION	ON-DEMAND MEDIAN $/GPU-HR	RANGE
A100	Ampere (prior)	$1.79 (Source: aimultiple.com)	Neocloud band, plus serverless outliers to $5.04
H100	Hopper (current)	$2.99 (Source: aimultiple.com)	Down from above $7 in early 2024
H200	Hopper (current)	$4.00 (Source: aimultiple.com)	$2.30 to $13.78
B200	Blackwell (newest)	$6.11 (Source: aimultiple.com)	$3.44 (Vast.ai) to $16.11 (Google Cloud)
B300	Blackwell (newest)	$7.92 (Source: aimultiple.com)	$5.44 (Vast.ai) to $18.00 (Oracle Cloud)

This progression makes the generational pricing pattern explicit: each new NVIDIA architecture launches at a premium roughly double the prior generation's median, then compresses toward that prior generation's price band as neocloud supply catches up with hyperscaler-only early availability. H100 today sits almost exactly where A100 sat at a comparable point in its own adoption curve, which is the basis for this report's expectation, discussed further in the Implications section below, that H100 and H200 pricing will continue drifting toward a "value tier" position as Blackwell-generation capacity scales through 2026 and 2027.

The most comprehensive cross-provider dataset available is the **AIMultiple Cloud GPU Rental Price Index**, compiled monthly from "63 providers and 17 GPU models" spanning "on-demand, spot, and 1-year reserved tiers" (Source: aimultiple.com). As of the most recent monthly snapshot referenced in this index, "H100 is listed by 46 providers, the broadest of any current accelerator," with a cohort median "now around $2.99/GPU-hour, down from above $7 in early 2024" (Source: aimultiple.com), a roughly 57% decline over roughly two years. H200 pricing spans "from $2.30 (FluidStack) to $13.78 (Microsoft Azure), with a cohort median around $4.00" (Source: aimultiple.com). By comparison, A100 (the prior-generation Ampere chip) holds "a tight neocloud band around $1.79," and RTX 4090 is "the cheapest training-class card on the index at $0.52 median" (Source: aimultiple.com).

Reserved and spot pricing tiers materially change the economics for buyers who can plan capacity in advance. The index finds that "the 1-year reserved discount typically runs 16 to 39% off the posted on-demand rate," but specifically that "H100 and H200 see modest single-digit-to-low-teens discounts" on reserved terms because "their on-demand market is competitive enough that providers do not sacrifice margin for commitments"

(Source: aimultiple.com). On spot pricing, "over the past six months, modern [GPU category] saves ~50%" versus on-demand (Source: aimultiple.com), broadly consistent with the roughly 44% AWS spot discount and the roughly 55% Vast.ai/RunPod spot discounts documented above.

A live snapshot from GPUFinder, refreshed daily, corroborates the wide H200 spread cited above: as of July 9, 2026 the "cheapest H200 confirmed in stock right now" was "$1.32/hr on Vast," against a "6-month range $2.93 to $15.97/hr" and a current floor price GPUFinder characterizes as "mid-range vs the last 6 months (up 6% vs last month)" (Source: gpufinder.dev). The same snapshot separately reports H200 on-demand pricing outside the marketplace tier "starts around $2.14/hr and runs up to $3.59/hr" for dedicated bare-metal capacity, framing the H200 as "settling between H100 and B200 rates" (Source: gpufinder.dev).

Availability, not just price, is a live constraint in 2026. AIMultiple reports that "H100, A100, and H200 cluster near 63 to 70%" confirmed-stock rates, "where roughly two-thirds of the catalog is confirmed stock and the rest is provisioning-dependent" (Source: aimultiple.com). This matches SemiAnalysis's on-the-ground survey finding that "hunting for even 8 nodes (64 GPUs) of H100s or H200s is not easy, half the providers we asked were completely sold out" as of early 2026, and that "market-wide, all capacity coming online until August to September 2026 has already been booked" (Source: newsletter.semianalysis.com). SemiAnalysis's own H100 1-year contract index, built from "direct survey data across a pool of 100+ market participants" and validated against actual transactions, tracks the 25th to 75th percentile range for GPU rental contracts and shows 1-year H100 pricing rising from $1.70/hr in October 2025 to $2.35/hr by March 2026 (Source: newsletter.semianalysis.com). This is a striking divergence from AIMultiple's flat-to-declining on-demand median over the same window, and buyers should read the two figures as describing genuinely separate markets rather than reconcile them into a single number.

Historical pricing context from **Jarvislabs** corroborates the broader downtrend in on-demand rates even as contract pricing tightened: H100 hourly rates moved from "$8.00 to $10.00 per hour (peak scarcity pricing)" in Q4 2024, down to "$5.50 to $7.00 per hour" in Q1 2025, "$3.50 to $4.50 per hour" in Q2 2025, and "$2.85 to $3.50 per hour (market stabilization)" by Q3 to Q4 2025, a cumulative decline the source characterizes as a "64 to 75% decrease from peak prices" (Source: jarvislabs.ai).

Cost-per-workload figures published alongside these price trends give a concrete sense of what current rates buy. At a representative $2.99/hr H100 rate, running a summarization pipeline requiring roughly "2 to 3 hours of GPU time per day" costs "$8.97/day or $269/month" for a single GPU (Source: jarvislabs.ai), while a lighter chatbot-style inference workload requiring roughly 1 to 1.5 hours per day of GPU time costs approximately $3.74/day, or $112/month. These figures underscore that for many production inference workloads, the effective monthly compute bill is a function of actual GPU-seconds consumed rather than the number of GPUs nominally provisioned, which is why per-second and per-minute billing granularity, discussed in the billing models section above, can materially change realized cost even when the headline $/hr rate is identical across two providers.

Case Studies and Real-World Examples

CoreWeave and Mistral AI: Purpose-Built Infrastructure for Frontier Model Training

CoreWeave, a Nasdaq-listed (CRWV) GPU cloud specialist, published a case study titled "Mistral AI Unlocks 2.5x Faster Training Speeds," documenting how the French AI lab used CoreWeave's purpose-built GPU infrastructure to accelerate large-scale model development (Source: www.coreweave.com). The engagement illustrates a broader pattern among frontier AI labs: rather than defaulting to a hyperscaler, Mistral selected a GPU-specialized provider explicitly for training-throughput reasons, a decision consistent with the price-performance argument neoclouds make across this market. CoreWeave has separately disclosed similarly structured infrastructure relationships with Cohere and IBM for training and deploying reasoning and agentic AI models on NVIDIA GB200 NVL72 rack-scale systems. An independent provider profile summarizes CoreWeave's positioning succinctly: it "runs one of the larger NVIDIA fleets in the market, with capacity spanning GB200 NVL72, B200, H200, H100, A100, and L40S," and notes that "reserved capacity discounts run up to 60% for committed terms" for buyers willing to sign multi-month or multi-year contracts (Source: comparegpuclouds.com).

CoreWeave and OpenAI: Scaling a Multi-Billion-Dollar Compute Relationship

CoreWeave's relationship with **OpenAI** is the largest publicly disclosed GPU cloud contract examined in this report and illustrates how committed-capacity pricing operates at the top of the market. In March 2025, CoreWeave announced an initial agreement with OpenAI worth up to $11.9 billion, expanded by $4 billion in May 2025, and expanded again on September 25, 2025 by up to $6.5 billion, bringing "the total contract value with OpenAI up to approximately $22.4 billion" (Source: www.coreweave.com). This scale of commitment sits at the opposite end of the market from the hourly, no-commitment rentals available on Vast.ai or RunPod, and demonstrates why "posted on-demand price" and "market clearing price for large committed capacity" diverge as sharply as the SemiAnalysis and AIMultiple data show.

Corvex and a Battery-Technology AI Provider: H200 Deployment for Production Inference

On January 22, 2026, **Corvex, Inc.**, an AI cloud computing company, announced a long-term GPU lease agreement for a dedicated cluster of NVIDIA H200 GPUs with "an established AI-driven provider of high-performance battery technologies," per its PR Newswire announcement (Source: www.prnewswire.com). The customer selected Corvex specifically "for its superior overall value, confidential AI enablement to unlock market expansion, and hyperscaler-class operations as compared to alternative AI cloud infrastructure providers" (Source: www.prnewswire.com). This case is notable for two reasons directly relevant to gpu cloud rental pricing decisions: it demonstrates a specific, named production use case selecting H200 (rather than H100) for a data-sovereignty-sensitive workload, and it illustrates the "hyperscaler-grade GPU management experience at a meaningfully lower cost" positioning that neoclouds consistently use to compete against AWS, Azure, and GCP on price (Source: www.prnewswire.com).

Cost Modeling Example: Fine-Tuning a Model on Rented H100s (Hypothetical Example)

To illustrate how hourly rental prices translate into project-level budgets, consider a hypothetical mid-sized AI team fine-tuning a 70-billion-parameter open-weight model using LoRA (low-rank adaptation) on 4x rented H100 GPUs. At a representative neocloud rate of $2.99/hr per GPU and a "Fine-Tuning Time: ~15 hours (typical for domain adaptation with LoRA)" (Source: jarvislabs.ai), total compute cost works out to roughly $179.40 (4 GPUs times $2.99 times 15 hours). By contrast, training a comparable model from scratch on 8x H100 GPUs over a median 840-hour (five-week) run at the same $2.99/hr rate costs approximately $20,093 in compute alone, versus an estimated $250,000 to purchase and provision the equivalent hardware outright, implying a break-even point of roughly 10,450 GPU-hours of continuous 8-GPU use, or about seven weeks. At smaller scale, a single dedicated H100 purchased outright costs roughly $25,000 plus $5,000 in supporting infrastructure, against a cloud cost of $2.99/hr times 720 hours per month, or $2,152/month, implying a break-even timeline of roughly 14 months of continuous 24/7 usage before ownership becomes cheaper than renting. For usage patterns under roughly 40 hours per month, the same source estimates cloud rental remains "20x more economical" than purchase (Source: jarvislabs.ai), reinforcing that the buy-versus-rent decision is overwhelmingly a function of utilization rate rather than absolute project size. (Hypothetical Example)

Baseten and H200 Inference Benchmarking: Memory Bandwidth in Production

Baseten, an inference infrastructure provider, worked with Lambda to independently benchmark NVIDIA H200 GPUs against H100 for large language model inference using **Mistral Large**, a 123-billion-parameter model, on an 8xH200 cluster (Source: www.baseten.co). The results were workload-dependent rather than uniformly favoring H200: for long input sequences, "the 8xH200 cluster offers 3.4X higher performance than the 8xH100 cluster for this sequence length and batch size," which the authors characterize as "certainly a cost-effective use for H200 GPUs" (Source: www.baseten.co). For high-throughput batch inference, H200 delivered "47% higher performance in BF16 and 36% higher performance in FP8," a gain the authors call "likely a cost-effective use of H200 GPUs" (Source: www.baseten.co). Critically, for short-context and short-output workloads, the two chips performed nearly identically: "the 8xH200 cluster offers approximately equal performance in BF16 and 11% higher performance in FP8" (Source: www.baseten.co).

This case is directly relevant to the H100-versus-H200 purchasing decision this report addresses: the benchmark data indicates the H200's price premium over H100, typically 15% to 40% depending on provider as shown in the feature comparison table above, is most easily justified for workloads with long context windows or large batch sizes, and least justified for short-context, low-batch inference, where paying the H200 premium buys little to no measured performance gain. Baseten's own guidance to buyers reflects exactly this split: "if H100s are a better fit for your inference needs, you can create a free Baseten account and get immediate on-demand access to H100 GPUs" (Source: www.baseten.co), underscoring that the two chips are complementary tools rather than a strict upgrade path.

Implications and Future Directions

Three structural forces are likely to shape gpu cloud rental prices through the remainder of 2026 and into 2027. First, the tension between falling on-demand posted prices and rising committed-capacity contract prices, documented above via the AIMultiple and SemiAnalysis indices respectively, appears likely to persist as long as inference demand from multi-agent AI workloads continues to grow faster than new datacenter capacity comes online. SemiAnalysis argues that "if the return on investment from using AI tools is 5 to 10x, then there is clearly a long way to go in GPU rental pricing before prices rise enough to curtail demand" (Source: newsletter.semianalysis.com), a view that, if it holds, implies continued upward pressure on committed and reserved H100/H200 pricing even as on-demand spot rates stay comparatively soft.

Second, the H100-to-Blackwell transition will keep placing downward pressure on Hopper-generation (H100/H200) pricing over time, even as near-term demand keeps rates elevated. B200 pricing currently carries a median of $6.11/GPU-hour with a range of $3.44 (Vast.ai) to $16.11 (Google Cloud) (Source: aimultiple.com), roughly double the H100 median, following "the pattern repeats from H100's earlier curve: hyperscalers carry new accelerators at 3 to 5x neocloud floors during the first year" (Source: aimultiple.com). As Blackwell capacity scales, H100 and H200 are likely to follow the same trajectory A100 followed after H100's launch, becoming a value tier for workloads that do not require the newest silicon.

Third, buyers should expect the gap between hyperscaler and neocloud pricing to persist rather than close, because it reflects structurally different cost bases (hyperscaler bundled networking, storage, compliance, and support overhead) rather than a temporary promotional discount. Teams with predictable long-term GPU needs and enterprise compliance requirements will continue to find value in hyperscaler reserved and committed-use contracts, while teams optimizing purely for compute cost per token or per training step will continue migrating toward neocloud and marketplace providers, a bifurcation this report's data confirms is already well established as of July 2026.

A fourth, more operational implication concerns procurement timing. Because SemiAnalysis documents that "all capacity coming online until August to September 2026 has already been booked" (Source: newsletter.semianalysis.com), teams planning large training runs several months out should treat committed-capacity negotiation as a scheduling problem, not merely a pricing negotiation. The practice SemiAnalysis describes of renters "subdividing their clusters and subletting the compute" (Source: newsletter.semianalysis.com) is itself a signal of how tight committed Hopper-generation capacity has become, and buyers who wait until a training run is imminent to shop for multi-node H100 or H200 clusters risk finding no inventory at any price, a materially different risk than simply paying a higher rate.

Frequently Asked Questions (FAQs)

How much does it cost to rent an H100 GPU in the cloud? As of mid-2026, on-demand H100 pricing ranges from approximately $1.38/hr on marketplace providers like Vast.ai up to $6.98/hr on Microsoft Azure, with a cross-provider median around $2.99/GPU-hour (Source: aimultiple.com).

How much does an H200 GPU cost to rent? H200 on-demand pricing spans roughly $2.30 to $13.78 per GPU-hour, with a median around $4.00, and neocloud on-demand rates typically landing between $3.50 and $4.50/hr (Source: aimultiple.com).

Is the H100 or H200 cheaper? The H100 is consistently cheaper, typically by 15% to 40% per GPU-hour depending on provider, because it is a lower-memory variant of the same underlying Hopper compute die used in the H200 (Source: nvidia.com).

What is the cheapest way to rent an H100 GPU? Vast.ai spot pricing, around $1.00/hr, is consistently cited as the market floor, with Vast.ai on-demand starting around $1.38/hr and RunPod Community Cloud at $1.99/hr as the more reliable on-demand alternative (Source: awesomeagents.ai).

How much does AWS charge for an H100? AWS's p5.4xlarge (single H100) lists at $6.88/hr on-demand (Source: getdeploying.com), and Capacity Blocks for a single H100 run $5.191/hr for guaranteed short-term reservations (Source: aws.amazon.com).

What does Lambda Labs charge for H100 GPUs? Lambda's on-demand H100 SXM instances price at $3.99 to $4.29/hr depending on region, and its 1-Click Cluster product prices reserved 16-GPU H100 clusters at $6.16/GPU-hour on 2-week to 1-year terms (Source: lambda.ai).

Why is cloud GPU pricing so much higher on hyperscalers than on specialized providers? Hyperscaler pricing bundles broader ecosystem services, enterprise support, compliance certifications, and integrated networking, while neoclouds compete purely on GPU-hour price; the gap reflects "50 to 400% more than GPU-first clouds for equivalent hardware" (Source: awesomeagents.ai).

Are cloud GPU prices going up or down in 2026? Both, depending on segment: posted on-demand list prices have declined roughly 57% to 75% from their early-2024/late-2024 peaks according to AIMultiple and Jarvislabs data respectively, while 1-year committed-capacity contract pricing rose roughly 40% between October 2025 and March 2026 per SemiAnalysis (Source: newsletter.semianalysis.com).

Should I buy an H100 or rent one? The purchase-versus-rent breakeven documented in the Case Studies section above lands around 14 months of continuous 24/7 use for a single H100, or roughly 500 GPU-hours per month before ownership overtakes rental economics; below that utilization level, cloud rental is consistently cheaper once power, cooling, and depreciation are factored into the purchase side of the comparison.

Does GPU cloud pricing include networking and storage? Rarely at the headline rate. As detailed in the AWS section above, storage, cross-AZ networking, and data egress are billed separately on top of the base instance rate at most hyperscalers, and buyers comparing quoted $/GPU-hour figures across providers should confirm whether networking, storage, and egress are bundled or itemized separately before treating any single number as the true all-in cost.

Conclusion

Renting an NVIDIA H100 or H200 GPU in mid-2026 costs anywhere from roughly $1.38 to nearly $14 per GPU-hour, and the single most important variable in that range is not the GPU generation but the provider category. Neocloud and marketplace providers, RunPod, Vast.ai, Lambda, Nebius, Jarvislabs, and Spheron among them, consistently price both H100 and H200 access at 40% to 400% below the three major hyperscalers for functionally identical NVIDIA silicon. The H100 remains the more cost-efficient default for most training and inference workloads, while the H200's roughly 15% to 40% price premium buys meaningfully more memory headroom (141GB versus 80GB) that becomes decisive for the largest open-weight models and long-context inference. Buyers evaluating GPU cloud rental prices should treat posted on-demand rates, spot pricing, and committed-capacity contract pricing as three separate markets moving in partially independent directions, since 2026 data shows on-demand list prices declining even as 1-year contract pricing for the same chips has risen sharply. For most organizations, the practical path forward is to match workload characteristics, batch and checkpointed jobs tolerate marketplace volatility for maximum savings, while production inference and long training runs justify paying a premium for reliability, to the provider tier that fits, rather than chasing the single lowest headline number. Whichever tier a team lands on, the data reviewed in this report supports one durable rule of thumb: revisit the pricing decision at least quarterly, since the spread between the cheapest and most expensive publicly listed H100 or H200 rate has moved by double digits in nearly every quarter since early 2024, and a provider that was the best deal six months ago is not guaranteed to hold that position today.

Tags: gpu cloud rental prices, h100 pricing, h200 pricing, nvidia h100, nvidia h200, cloud gpu comparison, aws h100, gpu rental cost, neocloud pricing

---