Nvidia H100 Rental Price History and H200 Cost Trends 2026

Published July 10, 2026 34 min read

Executive Summary

As of July 2026, renting a single **Nvidia H100** GPU by the hour costs roughly $1.49 to $12.29 depending on provider, with a market median around $2.95 to $3.46 per GPU-hour (Source: alatirok.com) (Source: getdeploying.com). Specialist "neoclouds" such as **RunPod** ($2.89 to $2.99 per hour) (Source: www.runpod.io), **Vast.ai** (marketplace rates from about $1.49 to $2.01) (Source: gpucloudcost.com), and **Hyperstack** (from $2.60) undercut the three big hyperscalers by a wide margin: **AWS** lists its P5 instance at $6.88 per H100-hour (Source: www.computetape.com), **Microsoft Azure** charges $6.98 to $12.29 depending on the source and commitment terms (Source: instances.vantage.sh) (Source: www.computetape.com), and **Google Cloud** runs approximately $10.98 per GPU-hour on its a3-highgpu instances (Source: www.spheron.network). The gap between the cheapest neocloud and the priciest hyperscaler listing is commonly 6x to 9x for functionally identical Hopper-generation silicon (Source: gpucloudcost.com).

The **Nvidia H200**, a memory-upgraded Hopper variant with 141 gigabytes (GB) of HBM3e versus the H100's 80GB, commands a modest premium over the H100 on most neoclouds, roughly $2.40 to $4.54 per GPU-hour (Source: aimultiple.com), but a much steeper one from hyperscalers, up to $13.78 per GPU-hour on Azure (Source: alatirok.com). H100 rental rates launched at approximately $4.70 per hour and spiked above $8 during the initial 2023 shortage (Source: www.latent.space), fell to roughly $2.85 by early 2024 (Source: www.latent.space), and continued sliding through 2025 as AWS cut on-demand P5 (H100) pricing by 44 percent and P5en (H200) pricing by 25 percent in June 2025 (Source: aws.amazon.com). AIMultiple's Cloud GPU Rental Price Index, which tracks 63 providers and 17 GPU models monthly, pegs the H100 on-demand cohort median at approximately $2.99 per GPU-hour as of mid-2026, down from above $7 in early 2024, a roughly 58 percent two-year decline (Source: aimultiple.com).

That decline, however, has not been uniform. SemiAnalysis's newly launched H100 1-year rental price index shows committed capacity moving in the opposite direction of spot: 1-year contract pricing rose roughly 40 percent, from a low of $1.70 per GPU-hour in October 2025 to $2.35 by March 2026, as on-demand capacity sold out across nearly every GPU generation and holders of locked-up compute refused to release it back into the pool (Source: newsletter.semianalysis.com). This split market, cheap and abundant spot capacity alongside tight and rising committed pricing, is the central fact any buyer needs going into a 2026 procurement decision.

Beyond raw hourly rates, the report below documents the full spread of on-demand, spot, and reserved pricing for both chips across more than 20 providers; explains why hyperscaler prices run 3x to 6x above neocloud floors for identical Hopper silicon (Source: alatirok.com); quantifies the GPU-as-a-service market's growth toward double-digit billions of dollars by the early 2030s (Source: www.marketsandmarkets.com); and walks through named case studies of companies that cut infrastructure spend by 60 percent or more by moving off hyperscaler GPU contracts (Source: vast.ai). The practical conclusion for most buyers: for short-lived, spot-tolerant, or experimental workloads, the H100 remains available for $2 to $3 per hour on multiple reputable neoclouds; for committed, multi-month capacity, buyers should expect 2026 rates to be flat-to-rising rather than falling, and should lock in pricing sooner rather than later.

Introduction and Background

The **Nvidia H100 Tensor Core GPU**, built on the Hopper architecture and launched in March 2023, became the default accelerator for large language model (LLM) training and inference during the generative AI boom (Source: gpufinder.dev). It ships with 80GB of HBM3 memory, 3.35 terabytes per second (TB/s) of memory bandwidth, fourth-generation NVLink offering 900 gigabytes per second (GB/s) of GPU-to-GPU interconnect, and a dedicated Transformer Engine that Nvidia states delivers up to 30x higher AI inference performance on the largest models compared to the prior A100 generation (Source: www.nvidia.com). Its successor, the **Nvidia H200**, released to general availability in the cloud in mid-to-late 2024, is not a new architecture but a memory refresh: it uses the identical Hopper GH100 die but ships with 141GB of HBM3e at 4.8TB/s, nearly double the memory capacity and 43 percent more bandwidth than the H100 (Source: www.nvidia.com) (Source: blog.barrack.ai).

Neither chip is sold to most buyers as retail hardware; the overwhelming majority of usage is rented by the GPU-hour from cloud providers, ranging from the three dominant hyperscalers (AWS, Microsoft Azure, Google Cloud) to a fast-growing category of specialist "neoclouds" (CoreWeave, Lambda, RunPod, Crusoe, Nebius, Hyperstack) and peer-to-peer GPU marketplaces (Vast.ai, Hyperbolic). This report answers the question "what does it cost to rent an H100 or H200 in 2026" by compiling verified, currently published on-demand, spot, and reserved pricing across more than 20 named providers, tracing the price history of both chips from their 2023 launch through mid-2026, and explaining the structural reasons behind the persistent 3x to 9x spread between the cheapest and most expensive listed rate for the same silicon (Source: gpufinder.dev).

Buying outright is largely a non-option for the audience this report serves: a single H100 GPU costs approximately $25,000 to $40,000 depending on the SXM versus PCIe form factor and configuration (Source: deploybase.ai). For nearly every AI team below hyperscale, renting by the hour, day, month, or year is the only economically sensible way to access Hopper-class compute, which is why hourly rental pricing, not sticker price, is the number that matters for budgeting AI infrastructure in 2026.The remainder of this report is organized around: the current on-demand and spot rate landscape for both chips; the historical price trajectory since 2023; the quantitative data underpinning the market; named real-world procurement outcomes; and forward-looking implications for buyers planning 2026-2027 GPU budgets.

Methodology. The figures compiled in this report come from two categories of sources, cross-checked against each other wherever possible. The first category is provider pricing pages fetched directly during research for this report, including RunPod, Lambda, CoreWeave, Crusoe Cloud, Hyperstack, and Nvidia's own specification pages, which are treated as authoritative for that provider's own listed rate. The second category is independent, methodologically transparent aggregators, specifically AIMultiple's Cloud GPU Rental Price Index (63 providers, 17 GPU models, monthly snapshots since July 2024) (Source: aimultiple.com), ComputeTape's reviewed provider-row methodology, which labels every figure as either a sourced quote or an explicitly flagged estimate (Source: www.computetape.com), and SemiAnalysis's H100 rental index, built from monthly surveys of institutional buyers and validated against real transaction data (Source: newsletter.semianalysis.com). Because posted list prices for hyperscalers frequently differ from normalized per-GPU node prices published by third-party trackers, this report notes both figures where they diverge rather than silently picking one.

Current On-Demand H100 Rental Pricing by Provider

On-demand H100 pricing in July 2026 spans roughly an 8x range between the cheapest and most expensive publicly listed rate for a single 80GB H100 GPU. Independent aggregator GPU Cloud Cost, which states it is independent of every vendor it lists and carries no affiliate links, put the spread at $1.49 per hour (Vast.ai) to $12.29 per hour (AWS EC2), an 8.2x spread, as last verified in June 2026 (Source: gpucloudcost.com). GPU Tracker's own snapshot corroborates a similarly wide spread, from $0.80 per hour on the low end to $97.44 on the high end across its full catalog of listings, though the extreme high figure reflects atypical regional and multi-GPU-node edge cases rather than a representative single-GPU rate (Source: gputracker.dev). GetDeploying's tracker, which covers 45-plus providers for the H100 specifically, similarly frames the market as wide but bounded, noting that an 8x H100 node priced at $24.00 per hour on-demand works out to $3.00 per GPU-hour once normalized, a figure it uses as a worked example of why cross-provider comparisons must normalize multi-GPU nodes before drawing conclusions (Source: getdeploying.com).

Table 1 below compiles verified, single-GPU-normalized on-demand H100 pricing directly from provider pricing pages and cross-checked aggregators, all captured in July 2026.

PROVIDER	TYPE	H100 SKU / INSTANCE	$ PER GPU-HOUR (ON-DEMAND)	SOURCE NOTE
Vast.ai	Marketplace	Third-party hosted H100	$1.49 to $2.01	Marketplace rate, varies by host and region (Source: gpucloudcost.com)
Thunder Compute	Neocloud	H100	$2.19	Fixed on-demand rate (Source: www.thundercompute.com)
Hyperstack	Neocloud	H100 (PCIe)	$2.60	Listed on-demand price (Source: www.hyperstack.cloud)
RunPod	Neocloud	H100 PCIe / SXM	$2.89 / $2.99	Public pricing page, per-second billing (Source: www.runpod.io)
CoreWeave	Neocloud	NVIDIA HGX H100 (8x node)	$6.16	$49.24/hr node price normalized per GPU (Source: www.coreweave.com)
Crusoe Cloud	Neocloud	H100 HGX	$3.90	Public listed hourly pricing (Source: www.crusoe.ai)
Lambda	Neocloud	H100 SXM (1x)	$3.99	Public 1-GPU rate (Source: lambda.ai)
AWS EC2 (P5)	Hyperscaler	p5.4xlarge	$6.88	On-demand, U.S. regions (Source: www.computetape.com)
Microsoft Azure	Hyperscaler	NC40ads_H100_v5	$6.98 to $12.29	Range across sources; East US regional list price $6.98 (Source: instances.vantage.sh); higher figure from ComputeTape's eastus row (Source: www.computetape.com)
Oracle Cloud	Hyperscaler	BM.GPU.H100.8	$10.00	8-GPU bare-metal node at $80.00/hr (Source: www.thundercompute.com)
Google Cloud	Hyperscaler	a3-highgpu-1g	$10.98	us-central1 listing (Source: www.spheron.network)

The pattern that emerges from *Table 1* is consistent across every independent tracker consulted for this report: neoclouds and GPU marketplaces cluster between $1.49 and $4.00 per GPU-hour, while the three major hyperscalers cluster between $6.88 and $11.06. AIMultiple's cross-provider index, compiled monthly from 63 providers, confirms the same shape at the aggregate level, finding that Microsoft Azure and Google Cloud carry the upper tail of H100 pricing past $10 per hour while Thunder Compute, Vast.ai, and RunPod sit at the bottom of the spread (Source: aimultiple.com). ComputeTape's independently sourced June 2026 snapshot, which reviews only public list-price rows with a documented source link and observation date, put the median across its five reviewed on-demand rows at $4.29 per H100-hour, with a band of $3.29 (RunPod Secure) to $12.29 (Azure) (Source: www.computetape.com).

Beyond the providers in *Table 1*, several second-tier neoclouds add further texture to the market. GPU Cloud Cost's June 2026 rate card lists **Together AI** at $3.09 per hour for a reserved 91-to-180-day cluster, landing between the neocloud floor and the hyperscaler ceiling (Source: gpucloudcost.com). Regional and commitment structure also matters: several published "cheapest" tables mix genuinely spot-available capacity with long-term contract pricing that only normalizes to a low per-GPU figure once a multi-year commitment is factored in, so a buyer should always confirm whether a headline rate is truly on-demand before budgeting against it. European buyers with data-residency requirements have at least one flat-rate alternative: **IONOS** offers dedicated H100 and H200 servers at a flat $3,990 per month with European Union data residency, a fixed-cost model that sidesteps the hourly-rate comparison entirely for teams that can commit to steady utilization (Source: aimultiple.com).

Nvidia H200 Rental Pricing and How It Compares to the H100

The H200's added memory carries a real but moderate premium on neoclouds and a much steeper one on hyperscalers. AIMultiple's index puts the H200's on-demand range from $2.30 (FluidStack) to $13.78 (Microsoft Azure), with a cohort median around $4.00 and a "working median" of $3 to $4 once community-tier and instance-share listings are set aside (Source: aimultiple.com). ByteCosts' provider-by-provider catalog corroborates the low end of that range, showing Vast.ai's cheapest confirmed on-demand H200 listing at $2.35 per GPU-hour in a French data center, with Nebius at $4.50, Google Cloud at $4.58, AWS at $7.91, and Oracle Cloud at $10.00 (Source: bytecosts.com).

RunPod's own pricing page lists H200 SXM at $4.39 per hour directly (Source: www.runpod.io), while CoreWeave's HGX H200 node lists at $50.44 per hour for 8 GPUs, or $6.31 normalized per GPU, with a spot price of $20.93 per hour for the node ($2.62 per GPU) (Source: www.coreweave.com). Crusoe Cloud lists H200 HGX at $4.29 per GPU-hour on its own pricing page (Source: www.crusoe.ai). Alatirok's May 2026 analysis, sourced from the AIMultiple index, summarized the cross-generation comparison bluntly: an H100 runs roughly $2.69 to $2.95 on-demand on a neocloud, an H200 adds "a small premium for a big memory jump" at $3.39 to $3.50, while a Blackwell-generation B200 anchors the top of the modern-GPU tier at $5.24 (Source: alatirok.com).

Whether the premium is worth paying depends entirely on workload memory footprint. According to GMI Cloud, running Meta's Llama 4 Maverick (400 billion parameters) on H100s requires two full 8-GPU nodes, while the same model fits on a single 8-GPU H200 node because of the added HBM3e capacity (Source: www.gmicloud.ai). Barrack AI's April 2026 analysis notes that median on-demand H200 pricing had actually climbed roughly 28 percent year-over-year as inference workloads for large open-weight models like Llama 3 70B and Mixtral 8x22B kept demand for the extra memory headroom stubbornly high, even as H100 pricing broadly declined (Source: blog.barrack.ai).

Spot, Reserved, and Multi-Year Pricing Tiers

Beyond the on-demand rate, three other billing tiers materially change the effective price of Hopper-generation compute: spot/preemptible instances, short-term reservations, and multi-year committed contracts.

Spot pricing. AIMultiple's spot discount tracker, which pairs each provider's spot rate against its own same-month on-demand rate for the same GPU, found that over the past six months the "modern" GPU category (which includes H100 and H200) has saved a median of approximately 50 percent versus on-demand, with the newest-generation cards saving around 49 percent and legacy cards around 75 percent (Source: aimultiple.com). CoreWeave's own spot pricing for its HGX H100 node is $19.71 per hour, versus $49.24 on-demand, roughly a 60 percent discount (Source: www.coreweave.com), and Hyperstack lists an H100 PCIe spot rate of $2.00 against a $2.60 on-demand list price (Source: www.hyperstack.cloud). Alatirok cites a Spheron H100 spot rate of $1.03 per hour, about 6.7 times cheaper than AWS's on-demand rate for the same silicon (Source: alatirok.com). Spot capacity is not free of tradeoffs: it can be reclaimed by the provider on short notice, so it suits checkpointed, interruption-tolerant batch workloads far better than long, uninterrupted training runs. GPU Finder's live spot tracker corroborates the general shape of this discount, showing H100 spot pricing starting from $0.34 per hour against an on-demand floor several multiples higher, and H200 spot starting from $0.66 per hour (Source: gpufinder.dev) (Source: gpufinder.dev).

Reserved and 1-year pricing. AIMultiple found that a 1-year reserved commitment typically nets 16 to 39 percent off posted on-demand rates, but that H100 and H200 specifically see only "modest single-digit-to-low-teens discounts," because on-demand H100/H200 markets are already competitive enough that providers do not need to sacrifice much margin to win a 1-year commitment (Source: aimultiple.com). Hyperstack's own reserved pricing lists H100 SXM at $2.72 per hour and H100 PCIe at $1.75, both below its on-demand list price of $3.99 and $2.60 respectively (Source: www.hyperstack.cloud).

The 1-year contract anomaly. The most important pricing signal in this report is that the reserved-capacity market is moving in the opposite direction of the on-demand/spot market. SemiAnalysis's newly launched H100 1-Year Rental Price Index, built from monthly surveys of more than 100 market participants and validated against real transaction data, shows the 1-year contract rate rising almost 40 percent from a low of $1.70 per GPU-hour in October 2025 to $2.35 by March 2026 (Source: newsletter.semianalysis.com). SemiAnalysis attributes this to on-demand capacity being effectively sold out market-wide, with renters who locked in earlier contracts refusing to release capacity back into the pool even as rates rise, and reports hearing of some H100 contracts being renewed for four-year terms through 2028 at the same rates they were originally signed (Source: newsletter.semianalysis.com). AIMultiple corroborates the divergence, noting that "the contract market moved differently: 1-year H100 commitments trended up over the same window, while our on-demand H100 median was roughly flat," which the firm describes as "a widening difference between month-to-month and 1-year committed pricing" (Source: aimultiple.com).

H100 vs H200 Rental Cost: Which to Choose

The headline hourly rate tells only part of the H100-versus-H200 story; the more relevant metric for many buyers is cost per unit of useful work. Alatirok's analysis argues that because the Blackwell-generation B200 delivers roughly 3x to 4x the inference throughput of an H100 for only about 2x the hourly rate, "the price-per-hour ranking inverts once you normalize to work done," a caution that applies equally to the H100-versus-H200

comparison for memory-bound workloads (Source: alatirok.com).

AIMultiple's workload-selection guidance is specific: for LLM inference on 30 to 70 billion parameter models, either an A100 80GB or an H100 is appropriate, with the H100 preferred when tight latency service-level agreements (SLAs) matter; for inference on 70-billion-plus-parameter, memory-bound models, the H200 (or AMD's MI300X) is preferred because its 141 to 192GB of HBM enables a larger key-value (KV) cache; and for fine-tuning models in the 7 to 13 billion parameter range, an A100 or H100 remains the cost-efficient default (Source: aimultiple.com) (Source: aimultiple.com). Runcrate's own H200 product page positions the chip explicitly as "the cost/perf sweet spot for production LLM inference" given its 141GB HBM3e (Source: www.runcrate.ai).

For teams uncertain which chip fits, the practical decision rule that recurs across the independent trackers surveyed for this report is: default to the H100 unless a specific model or context-length requirement genuinely overflows 80GB of VRAM, since the H200's added memory carries a real dollar premium on nearly every neocloud and a very large one on hyperscalers, and that premium is only recovered when the extra memory eliminates model-parallelism overhead or unlocks a materially larger batch size.

The two chips also differ in raw compute despite sharing the same GH100 die: Nvidia's own specification sheet lists identical FP8 Tensor Core throughput of 3,958 teraFLOPS (with sparsity) for both the H100 SXM and the H200 SXM, confirming that the H200 is a memory-and-bandwidth upgrade rather than a compute upgrade (Source: www.nvidia.com). The practical consequence is that any workload bottlenecked purely on compute throughput, rather than memory capacity or bandwidth, gains little from paying the H200 premium and should default to the cheaper H100. Conversely, workloads bottlenecked on key-value cache size for long-context inference, or on fitting a large model's weights without resorting to multi-node parallelism, are exactly the cases where the H200's 4.8TB/s of memory bandwidth (versus the H100's 3.35TB/s) and 141GB of capacity (versus 80GB) translate directly into fewer GPUs needed per deployment, which can offset or exceed the per-GPU-hour premium once the full cluster cost is compared rather than the single-GPU rate.

Cheapest H100 Rental Options and What They Trade Off

The cheapest listed H100 rates come from GPU marketplaces and boutique neoclouds, but "cheapest" and "usable" are not the same claim. A synthesis of developer discussion on Reddit's r/LocalLLaMA and r/StableDiffusion communities, gathered for this report, converges on a three-tier mental model of the market. At the ultra-cheap end, marketplaces such as Vast.ai and dynamic-pricing neoclouds such as Verda and DataCrunch offer the lowest headline rates but carry meaningfully higher variance in reliability, including reports of host-side reboots, oversubscribed RAM, and inconsistent PCIe bandwidth; developers describe this tier as suited to short, non-critical experiments rather than long uninterrupted training runs. A middle tier of specialized neoclouds, including RunPod and Lambda, offers a balance of price and professional-grade reliability, though Lambda in particular has drawn recurring developer complaints about on-demand availability of its high-end GPUs, with the community half-jokingly describing checking for open capacity as "refresh roulette." The hyperscaler tier is broadly seen as worth the premium only when the buyer needs a strict SLA, enterprise compliance certification, or very large committed clusters that only a hyperscaler can reliably provision.

A concrete illustration of the availability question comes directly from a fetched Reddit thread on r/LocalLLaMA, where a commenter with roughly eight years of enterprise cloud purchasing experience wrote: "How many of them have lower prices but no available capacity? For ex: lambda prices generally look decent but I've yet to see any capacity available," and separately noted that businesses that negotiate rarely pay the advertised hyperscaler sticker price, reporting "at least a 30% discount vs advertised price without any commitment requirements" and deeper discounts, down to "30% of the advertised price," for three-year commitments (Source: www.reddit.com). This corroborates the broader finding in this report that published list prices, on both ends of the market, are a starting point for negotiation rather than a final transaction price, particularly for larger or longer-term deployments.

AIMultiple's supply data adds a further caveat: across its dataset, H100, A100, and H200 cluster near 63 to 70 percent confirmed-stock availability, meaning roughly a third of the catalog for these chips is provisioning-dependent rather than instantly available, a materially tighter supply picture than commodity cards like the RTX 4090 or RTX 5090, which reach 93 to 97 percent confirmed availability (Source: aimultiple.com). Buyers evaluating "cheapest H100 rental" listings should therefore treat the headline dollar figure alongside a genuine availability check before committing budget to any one provider.

GPU Finder's live snapshot for July 2026 adds a directly comparable confirmed-in-stock figure: the cheapest H100 it could confirm as actually available for immediate rental, as opposed to merely listed, was $1.33 per hour on Vast.ai, alongside a note that pricing sits "mid-range vs the last 6 months" with a trailing six-month range of $2.00 to $15.20 per hour across the tracked catalog (Source: gpufinder.dev) (Source: gpufinder.dev). The same tracker identifies Vast.ai, PrimeIntellect, and DigitalOcean as the providers most reliably in stock over the trailing 30 days, while flagging Vultr and AceCloud as frequently waitlisted despite appearing in headline price comparisons (Source: gpufinder.dev). A buyer who filters strictly on price without cross-checking a provider's recent stock-confirmation history therefore risks selecting a listing that reads as cheap on a comparison page but is not actually orderable when the workload needs to start.

Data Analysis and Evidence

This section consolidates the quantitative backbone of the report: hourly rate distributions, market-size figures, and the mechanics behind the persistent hyperscaler premium.

Cross-provider spread. GPU Tracker's April 2026 snapshot, covering 54-plus providers and 5,213-plus tracked instance listings, put the H100 range at $0.80 (Verda) to $97.44 per hour and the H200 range at $1.19 (Verda) to $169.60 per hour, though the extreme high end of these ranges reflects multi-GPU node prices not normalized per GPU and edge-case regional listings rather than typical market rates (Source: gputracker.dev) (Source: gputracker.dev). GetDeploying's July 2026 tracker, covering 45 providers for the H100 specifically, reports an average of $3.46 per hour with a low of $0.34 per hour on spot instances (Source: getdeploying.com).

The hyperscaler premium, quantified. Alatirok's May 2026 analysis, drawing on the AIMultiple index, states the pattern in the clearest terms available in this research: "Neoclouds rent the exact same Nvidia GPUs as hyperscalers for 3x to 6x less: an H100 is roughly $2.69/hr on a neocloud versus $6.88 on AWS, $10.98 on Google Cloud, and $12.29 on Azure" (Source: alatirok.com). AIMultiple's own methodology note explains part of the mechanism: on Google Cloud specifically, the a3-highgpu, a3-megagpu, and a3-edgegpu SKUs are collapsed under a single "nvidia-h100" catalog label, which mechanically lifts the reported cohort median for that provider even though cheaper individual SKUs exist (Source: aimultiple.com).

Historical price trajectory. *Table 2* below reconstructs the H100 price history from launch through mid-2026 using figures independently verified from Latent.Space's October 2024 market analysis, Nvidia's own 2023 investor materials as cited in that analysis, the AWS price-cut announcement, and the AIMultiple and SemiAnalysis indices.

PERIOD	APPROXIMATE H100 RENTAL RATE	SOURCE AND CONTEXT
March 2023 (launch)	$4.70/hr typical, spiking above $8/hr	Original market rate at launch amid acute shortage (Source: www.latent.space)
2023 Nvidia investor deck	$4/hr	Nvidia's own "market opportunity" projection for renting H100s, later proven far too conservative (Source: www.latent.space)
Early 2024	~$2.85/hr	Multi-provider average as new supply came online (Source: www.latent.space)
August 2024	$1 to $2/hr for auctioned/spot capacity	Small-slice auction pricing as oversupply grew (Source: www.latent.space)
Early 2024 (hyperscaler median, per AIMultiple)	above $7/hr	AIMultiple's H100 cohort baseline before the 2024-2026 decline (Source: aimultiple.com)
June 2025	AWS cuts P5 (H100) on-demand 44%, P5en (H200) 25%	Official AWS price-reduction announcement, effective June 1, 2025 (Source: aws.amazon.com)
October 2025 (1-year contract low)	$1.70/hr	SemiAnalysis 1-year rental index trough (Source: newsletter.semianalysis.com)
March 2026 (1-year contract)	$2.35/hr	SemiAnalysis 1-year rental index, up ~40% from the October 2025 low (Source: newsletter.semianalysis.com)
Mid-2026 (on-demand cohort median)	~$2.95 to $2.99/hr	AIMultiple and Alatirok cross-provider medians (Source: aimultiple.com) (Source: alatirok.com)

Table 2 shows a market that fell roughly 58 percent on a headline on-demand basis between early 2024 and mid-2026 (Source: aimultiple.com), while its committed-capacity segment moved in the opposite direction over just the five months from October 2025 to March 2026. The interpretive takeaway is that "GPU prices are falling" and "GPU prices are rising" are both currently true statements, depending entirely on which billing tier is

being described, and any procurement plan that only checks the on-demand headline risks being blindsided by tight, rising reserved-capacity pricing.

Short-term volatility beneath the multi-year trend. The two-year decline documented in *Table 2* has not been a smooth line. Silicon Data's GPU Rental Index recorded a sharp, isolated 10 percent spike in H100 on-demand pricing over just four weeks, from a $2.00 per GPU-hour baseline on December 9, 2025 to a $2.20 peak on January 6, 2026, while A100 and B200 pricing over the same window stayed essentially flat (Source: www.silicondata.com). The firm attributes the isolated move to a combination of hyperscalers prioritizing reserved and long-term contract allocation over on-demand spot inventory, regionally scarcer spot capacity in Europe and Asia, and a "use-it-or-lose-it" surge in enterprise spending as organizations with December fiscal year-ends rushed to exhaust infrastructure budgets before they lapsed (Source: www.silicondata.com). The episode illustrates a broader point underlying this report's price-history data: even as the multi-year on-demand trend points down and the multi-year reserved-contract trend points up, month-to-month volatility within either trend can still produce a double-digit swing in a matter of weeks, and buyers timing a large procurement decision should treat any single snapshot price as one data point in a noisy series rather than a stable benchmark.

Market size. Independent research firms differ meaningfully on the total addressable GPU-as-a-service (GPUaaS) market size, which is a useful reminder that market-sizing methodology varies by scope (GPUaaS narrowly, versus the broader "cloud GPU rental" category, versus the full accelerated-computing hardware market). MarketsandMarkets valued the global GPUaaS market at $8.21 billion in 2025, projected to reach $26.62 billion by 2030 at a 26.5 percent compound annual growth rate (CAGR) (Source: www.marketsandmarkets.com). Analysys Mason forecasts GPUaaS revenue growing from $21 billion worldwide in 2024 to $134 billion by 2030 (Source: www.analysysmason.com). Mordor Intelligence estimates a smaller current base, $5.73 billion in 2025 growing to $7.38 billion in 2026 and $26.09 billion by 2031 (Source: www.mordorintelligence.com). These figures should not be treated as reconcilable to a single number; they reflect different scope definitions and methodologies across research firms, but all point in the same directional conclusion of sustained double-digit-percentage annual growth through the early 2030s. Two further estimates illustrate the same directional story at different scope: WiseGuyReports sizes the "cloud GPU rental" category specifically (as distinct from the broader GPUaaS category) at $3.91 billion in 2025, growing to $12.0 billion by 2035 at an 11.8 percent CAGR (Source: www.wiseguyreports.com), while Precedence Research puts the broader GPU-as-a-Service market at $4.96 billion in 2025, rising to approximately $6.10 billion in 2026 and $37.10 billion by 2035 at a 22.29 percent CAGR (Source: www.precedenceresearch.com).

Segment analysis: neocloud versus hyperscaler share of demand. Analysys Mason's GPUaaS forecast highlights a structural shift underway beneath the aggregate growth figures: hyperscale cloud providers currently account for the majority of GPUaaS revenue, carrying over their dominance from traditional cloud computing, but the firm forecasts that "their market share will gradually fall" as neoclouds and specialized GPU providers continue taking share on price and provisioning speed (Source: www.analysysmason.com). This dovetails with the pricing data in *Table 1* and *Table 2*: if buyers increasingly route price-sensitive training and inference workloads to neoclouds while reserving hyperscaler capacity for compliance-bound or enterprise-integration workloads, the revenue-share shift documented by Analysys Mason is the natural downstream consequence of the persistent 3x to 6x price gap this report documents at the unit-economics level.

Case Studies and Real-World Examples

Krnl: cutting infrastructure cost 65 percent by moving off AWS to serverless GPUs. According to a case study published by RunPod, the AI application company Krnl transitioned from AWS to RunPod's serverless GPU offering after repeatedly hitting infrastructure bottlenecks when its consumer AI tools went viral on social media. The published case study states the migration helped the company "scale to millions" of users while "slashing idle cost and scaling more efficiently," and the case study's own headline figure is a 65 percent cut in infrastructure costs (Source: www.runpod.io).

Creatix Technology: 60-percent-plus ongoing cost reduction on Vast.ai's GPU marketplace. Creatix Technology, the developer of the consumer apps Magic Eraser and AI Video Editor with more than 10 million downloads and over 200,000 daily active users, published a case study through Vast.ai describing how it combined on-premise servers with Vast.ai's distributed GPU marketplace to scale its AI-driven training and inference workloads. The case study reports "60%+ Ongoing Cost Reduction" and "5x+ Increased Daily User Support" as the two headline outcomes of the move (Source: vast.ai).

Foundry BioSciences: H100-cluster protein-language-model training via Canopy Wave GPUaaS. Foundry BioSciences, a startup using protein language models to accelerate research in protein evolution for longevity applications, is documented in a Canopy Wave case study as using 16x-H100 GPU instances through Canopy Wave's GPU-as-a-Service platform. The case study states the arrangement "reduced training times, minimized operational waste, and benefited from superior support" and reports the company found Canopy Wave outperforming AWS and Google Cloud specifically on stability and support responsiveness (Source: canopywave.com).

LayerJot: scaling from 5 to 256 GPUs across three cloud providers in one week. LayerJot, described as a med-tech startup building computer-vision and multimodal AI for medical equipment workflows, is documented in a Strong Compute case study as having been limited to 5 on-premises Nvidia GPUs before engaging Strong Compute. The published outcome states the company ran "44 experiments run across 6 separate AI projects"

and logged "6.5 hours total training time on 256 GPUs in 90 cloud machines across 3 different cloud providers, including H100 and A100 instances," and that the equivalent scaling effort would otherwise have required "2 full time engineers 3-6 months" of dedicated DevOps work (Source: words.strongcompute.com) (Source: words.strongcompute.com).

Runware: 5x to 10x lower generative-media inference pricing via flexible multi-provider GPU access. Together AI's published customer story on Runware, a generative image and video API provider, describes how the company combined custom infrastructure with Together AI's on-demand access to H100, H200, and B200 capacity to achieve "5-10x lower pricing" for its customers alongside same-day GPU scaling, rather than being limited to a single provider's available inventory (Source: www.together.ai). This case illustrates a pattern distinct from the single-provider migrations above: rather than switching wholesale from one cloud to another, some AI-native companies now treat multi-provider GPU access itself as the cost lever, routing workloads to whichever provider currently has the cheapest available capacity for a given chip generation.

(Hypothetical Example) A 12-person startup training a 70-billion-parameter model on spot capacity. To illustrate how the spot-pricing tier discussed earlier in this report translates into a real budget, consider a hypothetical 12-person AI startup that needs to fine-tune a 70-billion-parameter open-weight model. At an H100 spot rate near the $1.03 to $2.00 per GPU-hour range documented above (Source: alatirok.com) (Source: www.hyperstack.cloud), an 8-GPU node running for roughly 140 hours of checkpointed training would cost in the broad neighborhood of $1,150 to $2,250 in compute alone, versus several times that on a comparable hyperscaler on-demand rate, illustrating why spot capacity paired with robust checkpointing has become the default cost lever for budget-constrained teams, provided their workload can tolerate mid-run interruption.

Implications and Future Directions

Three structural forces will most likely shape H100 and H200 rental pricing through the remainder of 2026 and into 2027. First, the divergence between falling on-demand rates and rising committed-capacity rates documented by SemiAnalysis appears to be a genuine supply constraint rather than a temporary anomaly: the firm reports that all capacity coming online through August to September 2026 has already been booked, and that finding even a modest 8-node (64-GPU) cluster of H100s or H200s has become difficult, with about half the providers surveyed reporting complete sellouts (Source: newsletter.semianalysis.com) (Source: newsletter.semianalysis.com). Buyers planning multi-month or multi-year workloads should treat 2026 as a market in which locking in pricing sooner is likely to be cheaper than waiting, a reversal of the "wait for it to get cheaper" logic that dominated 2024.

Second, rising component costs are a documented contributor to tight rental supply, not just AI demand. SemiAnalysis reports that a parabolic rise in DRAM and NAND memory contract pricing in early 2026 forced original equipment manufacturers (OEMs) to reprice AI servers well beyond the increase in underlying component costs, which compressed the financial case for new cluster buildouts and caused some operators to slow-roll or abandon planned deployments, tightening the rental market further just as demand from agentic and multi-step AI workflows was accelerating (Source: newsletter.semianalysis.com).

Third, Blackwell-generation supply (B200, B300, GB200) is expanding from neocloud-only availability onto mainstream hyperscaler price sheets, and AIMultiple's data shows this expansion, not a genuine price increase in the underlying chips, is what has driven the "last released" GPU category's headline median roughly 25 percent higher over the past year: newly listed hyperscaler Blackwell rows simply carry a 2x to 3x markup over existing neocloud listings and pull the category average upward as they enter the tracked dataset (Source: aimultiple.com). As Blackwell capacity broadens, it is a reasonable expectation, though not a certainty documented in this research, that some inference workloads currently anchored to H100 or H200 capacity for cost reasons will migrate to B200 once its price-per-token advantage becomes more widely accessible outside large committed contracts, which could ease demand pressure on Hopper-generation on-demand pricing even as reserved Hopper pricing remains elevated through existing long-term contracts.

For procurement teams, the practical implication is to separate two distinct purchasing decisions: short-lived, spot-tolerant, or exploratory workloads should continue to shop the on-demand and spot markets aggressively, where genuine competition keeps neocloud rates well under half of hyperscaler list prices; but any workload requiring guaranteed multi-month capacity should be priced and negotiated now, since the reserved-capacity market has already demonstrated it can move 40 percent in five months when supply tightens.

A related implication concerns negotiating leverage at the hyperscaler tier specifically. As the enterprise-buyer commentary quoted earlier in this report indicates, published hyperscaler list prices are rarely the price a business actually pays once volume or term commitments enter the conversation, with discounts commonly reaching 30 percent off list with no commitment and considerably more for multi-year terms (Source: www.reddit.com). This means the 3x-to-6x neocloud-versus-hyperscaler gap documented throughout this report is best read as a ceiling on the realistic hyperscaler premium for any buyer with genuine negotiating scale, not a fixed multiplier every buyer will actually pay. Smaller teams and individual developers without that leverage, however, face the full list-price gap, which is precisely why the neocloud and marketplace tier has captured such a disproportionate share of the price-sensitive end of the market.

Finally, buyers should treat egress, storage, and networking fees as a meaningful addition to the headline hourly rate rather than a rounding error. AWS's own H200 documentation notes egress fees of $0.09 per gigabyte apply after the first 100GB of free transfer each month (Source: www.thundercompute.com), and RunPod's own pricing page shows persistent network storage running $0.05 to $0.07 per gigabyte per month, with a higher-performance storage tier at $0.14 per gigabyte per month (Source: www.runpod.io). For a data-intensive training run moving terabytes of checkpoints or downloading large open-weight model files, these secondary costs can meaningfully change the effective per-hour economics of an otherwise cheap listed rate, reinforcing the Reddit-sourced community guidance cited earlier in this report that a low advertised hourly rate can be deceptive without checking the full bill of materials.

Frequently Asked Questions (FAQs)

What is the average H100 GPU cost per hour in 2026? Cross-provider medians from independent trackers cluster between $2.95 and $3.46 per GPU-hour for on-demand rentals as of mid-2026, though individual providers range from about $1.49 up to $12.29 depending on whether the provider is a marketplace, specialist neocloud, or hyperscaler (Source: alatirok.com) (Source: getdeploying.com) (Source: gpucloudcost.com).

What is the cheapest way to rent an H100? Peer-to-peer marketplaces such as Vast.ai post the lowest headline on-demand rates, typically $1.49 to $2.19 per hour, with spot-tier pricing on providers like Spheron and Hyperstack falling as low as roughly $1.00 to $2.00 per hour (Source: gpucloudcost.com) (Source: alatirok.com).

How much does the H200 cost per hour compared to the H100? The H200 typically carries a 15 to 40 percent premium over the H100 on neoclouds, roughly $2.35 to $4.54 per GPU-hour versus the H100's $1.49 to $4.00, but the premium widens to nearly 5x on some hyperscaler listings, where H200 pricing can reach $13.78 per hour (Source: bytecosts.com) (Source: alatirok.com).

Why is Nvidia H100 cloud pricing so much higher on AWS, Azure, and Google Cloud than on specialist providers? Hyperscalers bundle the same underlying Hopper silicon with virtual private cloud (VPC) networking, compliance certifications, managed services, and enterprise support commitments that specialist neoclouds typically do not provide, and this bundling, not a difference in the chip itself, accounts for the bulk of the 3x to 6x price gap documented across independent trackers in this report (Source: alatirok.com).

Is GPU rental price history trending up or down in 2026? Both, depending on billing tier. On-demand and spot pricing has fallen roughly 58 percent from its early-2024 peak and remained roughly flat through mid-2026, while 1-year committed contract pricing rose approximately 40 percent between October 2025 and March 2026 as available on-demand capacity sold out market-wide (Source: aimultiple.com) (Source: newsletter.semianalysis.com).

Can I rent a single H100 or H200 GPU, or do I need to rent a full 8-GPU node? Specialist neoclouds and marketplaces, including RunPod, Lambda, Hyperbolic, Vast.ai, Modal, and Nebius, generally offer single-GPU access suited to individual developers and smaller workloads, while the major hyperscalers, AWS, Azure, and Oracle Cloud, typically sell H200 capacity only in full 8-GPU node configurations (Source: www.thundercompute.com).

How big is the cloud GPU rental market? Estimates vary by scope and methodology across research firms: MarketsandMarkets sizes global GPUaaS at $8.21 billion in 2025 growing to $26.62 billion by 2030 (Source: www.marketsandmarkets.com), while Analysys Mason projects GPUaaS revenue reaching $134 billion worldwide by 2030 (Source: www.analysysmason.com).

Does GPU rental price vary meaningfully by region? Yes. Thunder Compute's provider comparison explicitly excludes regional variation from its headline U.S. figures because "regional variation can add 5-20%" on top of the base rate depending on data center location (Source: www.thundercompute.com), and buyers with data-residency requirements in the European Union may find fixed-rate options such as IONOS's flat $3,990-per-month dedicated server more predictable than hourly billing across regions (Source: aimultiple.com).

What does an 8-GPU H100 node cost versus a single GPU? Multi-GPU nodes are the default unit of sale for hyperscalers and several neoclouds. CoreWeave's HGX H100 node lists at $49.24 per hour for 8 GPUs, which normalizes to $6.16 per GPU-hour (Source: www.coreweave.com), while Oracle Cloud's bare-metal 8-GPU H100 node runs $80.00 per hour total, or $10.00 per GPU-hour as shown in *Table 1* above. Buyers comparing single-GPU marketplace listings against multi-GPU hyperscaler nodes should always normalize to a per-GPU rate before concluding one option is cheaper, since headline node prices and single-GPU prices are not directly comparable without that adjustment.

Conclusion

Renting an Nvidia H100 in mid-2026 costs, on a market-median basis, somewhere between $2.95 and $3.46 per GPU-hour, with real published rates ranging from as low as $1.49 on peer-to-peer marketplaces to more than $12 on the priciest hyperscaler listing for identical silicon. The Nvidia H200 costs a moderate premium above the H100 on most specialist providers, typically $2.40 to $4.54 per hour, but a much steeper premium, up to $13.78

per hour, on hyperscalers. Both figures represent a dramatic decline from the H100's 2023 launch price of $4.70 to over $8 per hour, driven by a rapid buildout of neocloud and marketplace supply, aggressive hyperscaler price cuts including AWS's 44 percent P5 reduction in June 2025, and intensifying competition among more than 60 tracked providers.

That decline, however, is no longer the whole story. The reserved and multi-year contract market has moved in the opposite direction since late 2025, rising roughly 40 percent in five months as on-demand capacity sold out and rising component costs slowed new cluster construction. Buyers evaluating H100 or H200 rental options in the second half of 2026 should treat the on-demand and spot markets as continuing to favor the buyer, particularly for interruption-tolerant workloads on reputable neoclouds, while treating any multi-month or multi-year committed capacity decision as time-sensitive, since the data compiled in this report shows that segment of the market has already demonstrated it can tighten and reprice faster than most procurement cycles can react. The single most actionable finding for any team budgeting GPU compute for the remainder of 2026 is to separate these two purchasing decisions explicitly rather than benchmarking every workload against the same headline "H100 price."

Tags: h100 rental price, h200 rental price, gpu cloud pricing, nvidia h100 cost, gpu rental market, cloud gpu comparison, h100 vs h200, gpu-as-a-service

---