H100 GPU Cloud Pricing Comparison: AWS vs GCP vs Azure

Published July 17, 2026 38 min read

Executive Summary

As of July 2026, renting an eight GPU **NVIDIA H100** server from a major hyperscaler costs anywhere from roughly $41 to $98 per hour depending on provider, commitment term, and region, while specialized GPU clouds and marketplaces can bring the effective per-GPU rate down by 70 percent or more. On **Amazon Web Services (AWS)**, the flagship **p5.48xlarge** instance (8x H100, 640GB HBM3) lists at $55.04 per hour on demand (Source: instances.vantage.sh), but AWS's own Capacity Blocks for ML program prices the same hardware at an effective $41.528 per instance hour in US East regions when reserved in advance (Source: aws.amazon.com. On **Google Cloud**, the H100-based **a3-highgpu-8g** machine type runs at a rate that falls to $38.86 per hour under a three year committed use discount (Source: cloud.google.com), while the newer, higher-throughput **a3-megagpu-8g** lists at $93.40 per hour on demand (Source: cloud.google.com). Notably, Google's **a3-ultragpu-8g** SKU, frequently searched alongside H100 pricing, actually ships with 8x NVIDIA H200 GPUs rather than H100 (Source: cloud.google.com) (Source: cloudprice.net), a distinction this report addresses directly. On **Microsoft Azure**, the **ND H100 v5** series (Standard_ND96isr_H100_v5, 8x H100, 96 vCPUs, 1900GB RAM) is estimated by third-party trackers at $98.32 per hour on demand, or roughly $12.29 per GPU, the highest headline on-demand rate among the three hyperscalers (Source: instances.vantage.sh).

Outside the big three, specialized "neocloud" providers consistently undercut hyperscaler on-demand pricing for the same silicon. **CoreWeave** prices its NVIDIA HGX H100 configuration at $49.24 per hour on demand and $19.71 per hour on spot (Source: www.coreweave.com). **Oracle Cloud Infrastructure (OCI)** charges a flat $10.00 per GPU per hour for its BM.GPU.H100.8 bare metal shape (Source: www.oracle.com), equivalent to $80 per hour for an 8-GPU node. **Lambda Labs** lists self-serve H100 SXM instances from $3.99 per GPU per hour and H100 PCIe from $3.29 per GPU per hour (Source: lambda.ai) (Source: lambda.ai), while its reserved 1-Click Clusters fall to $5.54 per GPU per hour at 256-GPU scale (Source: lambda.ai). **RunPod** rents H100 PCIe from $1.99 per hour on Community Cloud and $2.39 per hour on Secure Cloud (Source: www.runpod.io), and the peer-to-peer marketplace **Vast.ai** shows H100 SXM rates as low as $1.92 per hour depending on host tier (Source: vast.ai). An aggregator tracking 47 H100 providers puts the market average at $3.41 per GPU per hour, with spot instances occasionally dipping to $0.34 per hour, and reports that blended on-demand pricing has fallen roughly 4 percent since July 2025 (Source: getdeploying.com.

The pricing gap reflects real differences in networking, support, compliance, and availability rather than pure arbitrage. Hyperscaler H100 clusters ship with proprietary interconnects, in AWS's case up to 3,200 Gbps EFA networking and NVSwitch linking up to 20,000 GPUs in a single UltraCluster (Source: aws.amazon.com), and Azure's ND H100 v5 nodes each carry a dedicated 400 Gb/s NVIDIA Quantum-2 InfiniBand connection per GPU (Source: learn.microsoft.com). These features matter for multi-week distributed training runs but are frequently unnecessary for inference or single-node fine-tuning, exactly where neoclouds and marketplaces compete hardest on price. Demand for H100 capacity remains elevated even as newer Blackwell-generation chips reach the market: Nvidia itself told investors in late 2024 that supply chain constraints would leave demand exceeding supply for several quarters (Source: www.reuters.com), and Goldman Sachs-cited hyperscaler capital expenditure is projected to reach $602 billion in 2026, of which roughly 6 million GPU-equivalents at an average price near $30,000 are purchased directly by the largest cloud operators (Source: introl.com).

For most buyers the practical takeaway is that the cheapest H100 cloud instance is almost never a hyperscaler on-demand rate; it is a neocloud spot instance or a marketplace listing, and the right choice depends on workload duration, networking needs, egress and storage fees layered on top of compute, and tolerance for interruption or capacity-quota friction, all of which this report quantifies in detail below.

Introduction and Background

The **NVIDIA H100 Tensor Core GPU**, built on the Hopper architecture and launched in 2022, remains as of mid-2026 one of the most widely deployed accelerators for large language model training, fine-tuning, and high-throughput inference, even as NVIDIA's newer Blackwell generation ramps into hyperscaler fleets. NVIDIA's own product page states the chip speeds up large language model inference by up to 30 times over the prior generation (Source: www.nvidia.com and, via its fourth-generation Tensor Cores and FP8-precision Transformer Engine, delivers up to 4 times faster GPT-3 175B training over the prior generation (Source: www.nvidia.com. Because a single H100 card costs upwards of $25,000 to purchase outright, according to cloud GPU rental provider RunPod (Source: www.runpod.io), the overwhelming majority of organizations access H100 compute through cloud rental rather than capital purchase, which makes cloud pricing comparisons a first-order business decision rather than a technical footnote.

This report answers the practical question that engineering leads, ML infrastructure teams, and finance stakeholders actually ask: what does an H100 cost per hour, across which providers, under which billing models, and how do those figures compare. It focuses on comparisons across AWS, Google Cloud, and Azure, the three hyperscalers, alongside the specialized GPU clouds and marketplaces (CoreWeave, Lambda Labs, Oracle Cloud Infrastructure, Nebius, RunPod, and Vast.ai) that increasingly set the effective market price. It pays specific attention to instance families that are frequently confused, most importantly Google Cloud's **a3-ultragpu-8g**, which despite its "A3" family name and frequent appearance in H100 comparison searches is actually built around NVIDIA H200 GPUs, not H100.

Several billing dimensions recur throughout this analysis and are worth defining up front. **On-demand pricing** is the standard pay-as-you-go hourly rate with no commitment. **Spot** or **preemptible pricing** offers a discount, sometimes 50 to 70 percent, in exchange for the provider's right to reclaim the instance with short notice, making it suitable for checkpointed or fault-tolerant training but risky for long uninterrupted runs. **Reserved** or **committed use discount (CUD)** pricing requires a one to three year commitment in exchange for a lower effective rate, and AWS additionally offers **Capacity Blocks for ML**, a mechanism to reserve a known quantity of GPU capacity for a defined future window without a long-term contract, as detailed in the AWS section below.Because GPU cloud pricing changes frequently, sometimes multiple times within a single year as **Nebius** demonstrated with a June 1, 2026 rate change for its H100, H200, B200, and B300 lines (Source: docs.nebius.com), every figure in this report is anchored to a specific "as of" date, primarily mid-July 2026, and readers evaluating a purchase decision should re-verify current rates before committing budget. Underlying all of this is a chip market that Nvidia itself has described as demand-constrained: the company told investors in November 2024 that it was selling H100 and successor chips as fast as its chipmaking contractor, Taiwan Semiconductor Manufacturing Co, could produce them (Source: www.reuters.com, and the same reporting noted AI chip rival AMD, along with Intel and Qualcomm, traded lower the same day even as Nvidia's own shares hit a record high, illustrating how tightly investor sentiment tracks Nvidia-specific supply signals (Source: www.reuters.com), a dynamic that continues to shape regional H100 availability and pricing firmness even now.

AWS H100 Pricing: EC2 P5 Family and Capacity Blocks

AWS offers H100 access primarily through the **Amazon EC2 P5** instance family. The flagship **p5.48xlarge** instance pairs 192 vCPUs and 2 TiB of system memory with 8 NVIDIA H100 GPUs delivering a combined 640GB of HBM3 memory, connected by 900 GB/s NVSwitch for single-hop GPU-to-GPU communication and up to 3,200 Gbps of cross-node bandwidth via second-generation Elastic Fabric Adapter (EFA) (Source: aws.amazon.com). AWS states these instances can accelerate time to solution by up to 4x compared to prior-generation GPU instances while reducing the cost to train machine learning models by up to 40 percent (Source: aws.amazon.com), and, as noted above, P5 UltraClusters can scale to as many as 20,000 H100 or H200 GPUs interconnected on a petabit-scale nonblocking network.

On the pricing side, third-party instance tracker **Vantage** confirms p5.48xlarge on-demand pricing starting at $55.04 per hour (Source: instances.vantage.sh), as referenced in the Executive Summary, with a one-year reserved rate of $23.777 per hour (Source: instances.vantage.sh) and a spot price around $19.769 per hour when capacity is available (Source: instances.vantage.sh), reflecting the steep discounts available for interruption-tolerant workloads relative to on-demand billing. A separately priced aggregator lists a single-GPU **p5.4xlarge** on-demand rate of $6.88 per hour (Source: getdeploying.com), useful for teams that need one H100 rather than a full 8-GPU node.

AWS's own **Capacity Blocks for ML** pricing page offers the clearest first-party figures for teams planning ahead. It lists p5.48xlarge at an effective $41.528 per instance hour, or $5.191 per GPU-hour, across most US regions, as referenced in the Executive Summary (Source: aws.amazon.com, dropping to $37.76 per instance hour in Asia Pacific, Australia, Europe, and South America regions. The single-GPU p5.4xlarge Capacity Block runs $5.191 per hour in US regions (Source: aws.amazon.com. AWS also sells H200-based successors on the same P5 chassis family: p5e.48xlarge (8x H200) lists at $47.76 per instance hour under Capacity Blocks (Source: aws.amazon.com and the higher-bandwidth p5en.48xlarge (8x H200, third-generation EFA) runs $54.920 per instance hour in US East and US West regions (Source: aws.amazon.com, giving buyers a clear read on the H100-to-H200 price delta within AWS's own portfolio. Because Capacity Blocks require committing to a specific future time window rather than paying purely on demand, they function as a middle ground between full on-demand flexibility and multi-year reserved pricing, and AWS markets them specifically as a tool for organizations that need guaranteed access during a known training window without negotiating a long-term contract.

Google Cloud H100 and H200 Pricing: A3 High, A3 Mega, and A3 Ultra

Google Cloud's H100 offering sits within its **A3 accelerator-optimized machine family**, but the family has grown to include three tiers that are frequently conflated in search traffic, and only two of them actually run H100 silicon. The original **a3-highgpu-8g** and the networking-upgraded **a3-megagpu-8g** both pair 8 NVIDIA H100 GPUs with either 208 vCPUs and 1,871GB of memory or 208 vCPUs and 1,872GB of memory respectively. The newest tier, **a3-ultragpu-8g**, pairs 8 GPUs with 224 vCPUs and 2,952GB of memory, but, as noted above, those 8 GPUs are NVIDIA H200 units, not H100, a distinction that matters because "A3 Ultra" is one of the most common near-miss queries in any H100 pricing comparison, and buyers searching for H100 capacity on this SKU will actually receive next-generation H200 hardware with 141GB of HBM3e memory per GPU rather than the H100's 80GB of HBM3.

Google Cloud publishes granular pricing across on-demand, Dynamic Workload Scheduler (DWS) flex-start and calendar modes, spot, and one- and three-year committed use discount (CUD) tiers. For **a3-highgpu-8g** (H100), on-demand pricing is $88.49 per hour (Source: cloud.google.com), and a three-year CUD brings the effective rate down to $38.86 per hour, a discount of better than 55 percent versus list price, as referenced in the Executive Summary. For **a3-megagpu-8g** (H100, with enhanced GPU-to-GPU networking), on-demand pricing is higher, at $93.40 per hour as noted in the Executive Summary, reflecting the additional network throughput built into the SKU. By comparison, the H200-based **a3-ultragpu-8g** lists on-demand at $84.81 per hour (Source: cloud.google.com), cheaper on a headline basis than either H100 SKU despite the newer chip, largely reflecting Google's differentiated capacity pricing rather than any inherent value difference; on a monthly basis, a third-party pricing tracker lists a3-ultragpu-8g starting from $63,334.74 per month for continuous use (Source: cloudprice.net), and confirms the 8 GPU, 224 vCPU, 2,952GB memory specification independently, as referenced above.

Google's collaboration with NVIDIA on the A3 line dates to the instances' general availability announcement, when NVIDIA confirmed that "Google Cloud today announced the general availability of its new A3 instances, powered by NVIDIA H100 Tensor Core GPUs" (Source: blogs.nvidia.com), and Google Cloud CEO Thomas Kurian and NVIDIA CEO Jensen Huang discussed how Google was already using H100 and A100 GPUs internally for research and inference across DeepMind and other divisions at the time of that launch (Source: blogs.nvidia.com), a partnership that NVIDIA also credits with bringing the NVIDIA L4 GPU to Google Cloud ahead of other cloud service providers (Source: blogs.nvidia.com). Buyers comparing "Google Cloud A3 Ultra vs AWS P5" should therefore recognize they are frequently comparing an H200 instance against an H100 instance rather than an apples-to-apples GPU generation match, and should instead pair AWS p5.48xlarge against Google's a3-highgpu-8g or a3-megagpu-8g for a true H100-to-H100 comparison, or pair AWS's H200-based p5e.48xlarge against Google's a3-ultragpu-8g for a true H200-to-H200 comparison.

Azure H100 Pricing: ND H100 v5 and NC H100 v5 Families

Microsoft Azure splits its H100 portfolio into two distinct series that serve very different use cases. The **ND H100 v5** series is Azure's flagship scale-out training SKU. Microsoft describes it as starting with a single virtual machine carrying eight NVIDIA H100 Tensor Core GPUs, capable of scaling to thousands of GPUs with 3.2 Tbps of interconnect bandwidth per VM (Source: learn.microsoft.com). Each GPU carries its own dedicated 400 Gb/s NVIDIA Quantum-2 CX7 InfiniBand connection with GPU Direct RDMA support, as referenced in the Executive Summary, and Microsoft further notes that each GPU also features NVLINK 4.0 connectivity for communication within the VM, paired with 96 physical fourth-generation Intel Xeon Scalable processor cores (Source: learn.microsoft.com). The single published size in the series, Standard_ND96isr_H100_v5, pairs 96 vCPUs with 1,900 GiB of memory and 8 H100 GPUs carrying 640GB of combined accelerator memory (Source: learn.microsoft.com. As also noted above, third-party tracker Vantage lists this instance starting at $98.32 per hour on demand or $18.169536 per hour on spot (Source: instances.vantage.sh).

A separate industry pricing comparison independently corroborates the $98.32 headline figure, converting it to roughly $12.29 per GPU per hour (Source: www.spheron.network), and adds detail aggregators built for AWS-style pricing do not typically surface: a one-year reserved commitment brings the effective rate to approximately $63.40 per instance hour (about $7.93 per GPU), while a three-year reserved commitment brings it to roughly $43.79 per instance hour (about $5.47 per GPU) (Source: www.spheron.network, and Azure's low-priority spot tier for this SKU is quoted in a wide $18.00 to $29.50 per instance hour range depending on region and availability (Source: www.spheron.network. Because this source is published by a competing neocloud provider, its comparative framing should be read as directional rather than authoritative, but its underlying Azure list-price figures are internally consistent with the independently sourced Vantage numbers above, and it separately notes that Azure's quota approval process for large ND96isr requests, covering eight or more nodes, can stretch to three or four weeks before capacity is even confirmed as available (Source: www.spheron.network.

Azure's second H100 family, **NC H100 v5**, targets smaller-scale applied AI and inference rather than large distributed training. The **NCads_H100_v5** series is powered by NVIDIA H100 NVL GPUs paired with fourth-generation AMD EPYC Genoa processors, and its virtual machines feature up to 2 NVIDIA H100 NVL GPUs with 94GB of memory each, up to 96 non-multithreaded AMD EPYC Genoa cores, and 640 GiB of system memory (Source: learn.microsoft.com). A related confidential-computing variant, **NCCads_H100_v5**, spans a Trusted Execution Environment across both the CPU and the attached GPU, enabling secure offload of data, models, and computation, and its VMs feature a single NVIDIA H100 NVL GPU with 94 GB of memory, 40 AMD EPYC Genoa cores, and 320 GiB of system memory (Source: learn.microsoft.com). An aggregator lists a single-GPU Azure H100 configuration (NCadsH100v5) at $6.98 per hour on demand (Source: getdeploying.com), roughly on par with AWS's single-GPU p5.4xlarge rate, though the two are not identical hardware configurations since Azure's NC series uses the NVL form factor rather than the SXM form factor found in AWS P5 and Azure ND H100 v5, a distinction Microsoft itself draws by marketing the NC line specifically for real-world applied AI training and batch inference rather than large-scale distributed pretraining (Source: learn.microsoft.com).

The practical implication for anyone searching "nd h100 v5 pricing azure" is that Azure's naming convention separates a small-node applied-AI product line (NC, single or dual GPU, NVL form factor, often confidential-computing capable) (Source: learn.microsoft.com) from a large-node distributed-training product line (ND, eight GPU, SXM form factor, InfiniBand-networked), and confusing the two when requesting a quote can lead to a significant mismatch between expected and delivered price and performance.

Specialized GPU Clouds and Neoclouds: CoreWeave, Lambda, Oracle, Nebius, RunPod, and Vast.ai

A parallel ecosystem of GPU-specialized cloud providers, often called "neoclouds," consistently prices H100 capacity below the three hyperscalers, and understanding this tier is essential for anyone searching for the cheapest H100 cloud instance available in 2026.

CoreWeave, one of the largest independent GPU cloud operators, prices its NVIDIA HGX H100 configuration (8 GPUs, 61.44 TB local storage per node) at $49.24 per hour on demand and $19.71 per hour on spot (Source: www.coreweave.com), which works out to roughly $6.16 per GPU per hour on demand, a figure independently corroborated by a comparison tracker that lists CoreWeave's H100 SXM rate at the same $6.16 per hour per GPU (Source: computeprices.com); the same tracker separately confirms CoreWeave's H200 rate at $6.30 per hour per GPU, only a modest premium over H100 (Source: computeprices.com). CoreWeave's growth has been closely tied to Microsoft: CNBC reported in 2023 that Microsoft's agreement with CoreWeave could be worth billions of dollars over multiple years to secure GPU capacity for OpenAI's workloads on Azure, at a time when NVIDIA had invested $100 million in CoreWeave, which was then valued at $2 billion (Source: www.cnbc.com) (Source: www.cnbc.com), and CoreWeave itself was, at the time, described as offering GPU access at prices its own marketing claimed were up to 80 percent less expensive than legacy cloud providers (Source: www.cnbc.com), a claim CNBC noted CoreWeave was making at the time on its own website (Source: www.cnbc.com).

Oracle Cloud Infrastructure takes the simplest approach to H100 pricing among the providers surveyed: its official price list quotes a flat $10.00 per GPU per hour for the BM.GPU.H100.8 bare metal shape, which pairs 8 NVIDIA H100 80GB Tensor Core GPUs on Hopper architecture with 8x2x200 Gb/sec RDMA networking, the identical rate Oracle charges for its newer BM.GPU.H200.8 shape (Source: www.oracle.com), which is corroborated independently by a third-party GPU price tracker showing the same $10.00 per hour per GPU figure (Source: computeprices.com). At full 8-GPU scale, that works out to $80 per instance hour, positioning OCI between CoreWeave and the AWS on-demand list price, and Oracle separately advertises a B200 bare metal shape at $14.00 per GPU-hour (Source: computeprices.com), giving a sense of the premium the next generation of accelerators commands over H100 within the same provider's price list. Oracle's own infrastructure page states that its OCI Supercluster networking can scale to as many as 16,384 NVIDIA H100 Tensor Core GPUs in a single cluster, alongside much larger allocations of newer Blackwell-generation B200 GPUs, illustrating that H100 fleets are still being provisioned at very large scale even as next-generation capacity comes online (Source: www.oracle.com). Oracle markets itself as the only major cloud provider offering bare metal GPU instances free of virtualization overhead across

NVIDIA and AMD accelerators (Source: www.oracle.com), a positioning aimed squarely at HPC and large-scale training customers who want to eliminate hypervisor overhead entirely, and separately offers AMD Instinct MI300X accelerators at $6.00 per GPU-hour as a lower-cost, non-NVIDIA alternative for organizations willing to port workloads off CUDA (Source: www.oracle.com).

Lambda Labs offers both self-serve, pay-as-you-go instances and reserved "1-Click Clusters." On the self-serve side, as noted above, Lambda prices single H100 SXM instances from $3.99 per GPU per hour and H100 PCIe instances from $3.29 per GPU per hour, varying slightly by configuration size and region. Its reserved 1-Click Clusters, which require a two-week to one-year commitment, price H100 capacity at $6.16 per GPU per hour for 16-GPU clusters, $5.85 per hour at 64-GPU scale (Source: lambda.ai), and, as referenced in the Executive Summary, $5.54 per hour at 256-GPU scale, illustrating a modest volume discount as cluster size grows.

Nebius, a European GPU cloud that has aggressively expanded H100 and H200 capacity, illustrates just how quickly this market moves: its documentation shows H100 NVLink pricing rising from $2.95 per GPU-hour on demand ($1.25 preemptible) before June 1, 2026 (Source: docs.nebius.com) to $3.85 per GPU-hour on demand ($2.15 preemptible) effective that date (Source: docs.nebius.com), alongside simultaneous increases to its B300, B200, and H200 lines. This upward repricing, occurring even as broader market averages trend downward according to aggregator data, underscores that GPU cloud pricing is not monotonically falling and that buyers should check current rates rather than relying on historical figures, however recent.

RunPod splits its offering into Community Cloud, hosted on vetted third-party data centers, and Secure Cloud, hosted in Tier 3 and Tier 4 facilities with enhanced compliance posture. As noted above, RunPod's own product page states it offers H100 PCIe from $1.99 per hour on Community Cloud and $2.39 per hour on Secure Cloud, and the same page notes the underlying card carries 80GB of HBM3 memory and 14,592 CUDA cores based on the Hopper architecture (Source: www.runpod.io). RunPod's general pricing page separately lists a Secure Cloud rate of $2.89 per hour for H100 configurations (Source: www.runpod.io), a discrepancy likely reflecting different H100 variants (SXM versus PCIe) or a pricing update between page refreshes that this report flags rather than silently reconciles, and the same pricing page lists standalone CPU-only instances at $4.39 per hour on Secure Cloud, underscoring that RunPod's infrastructure extends beyond GPU rental alone (Source: www.runpod.io).

Finally, **Vast.ai** operates a genuine marketplace where independent hosts set their own prices, and pricing consequently varies more than at any other provider surveyed. Vast.ai's live pricing page shows H100 SXM at a low of $1.92 per hour, as noted above, with a market range extending well beyond $4 per hour depending on host reliability tier, H100 PCIe at a low of $1.87 per hour (Source: vast.ai), and H100 NVL at a low of $2.00 per hour (Source: vast.ai). Because Vast.ai hosts are not vetted to the same data-center-tier standard as CoreWeave or the hyperscalers, buyers evaluating its listings should weigh the lower price against variable reliability, support, and compliance guarantees; the platform itself describes its pricing as market-driven, with hosts setting their own rates rather than the company publishing a single fixed price (Source: vast.ai).

Table 1 below summarizes on-demand, spot or preemptible, and discounted pricing for 8-GPU H100 (and, where noted, H200) configurations across the three hyperscalers, drawing on the figures already sourced in the sections above.

PROVIDER	INSTANCE / SHAPE	GPU	ON-DEMAND ($/HR)	SPOT / PREEMPTIBLE ($/HR)	BEST DISCOUNTED RATE ($/HR)
AWS	p5.48xlarge (8x H100)	H100 SXM, 640GB total	$55.04	$19.77	$41.53 effective, Capacity Blocks
Google Cloud	a3-highgpu-8g	H100 SXM, 8x80GB	$88.49	~$48.19	$38.86, 3-yr CUD
Google Cloud	a3-megagpu-8g	H100 SXM, enhanced networking	$93.40	~$50.80	~$40.65, 3-yr CUD
Google Cloud	a3-ultragpu-8g *(Not H100; included for comparison)*	**H200** SXM, 8x141GB	$84.81 (Source: cloudprice.net)	~$46.56	~$37.21, 3-yr CUD
Azure	Standard_ND96isr_H100_v5	H100 SXM, 640GB total	$98.32 (Source: instances.vantage.sh)	$18.17 to $29.50	~$43.79, 3-yr reserved

As the table shows, headline on-demand pricing for an 8-GPU H100 node varies by more than 2.4x across the three hyperscalers, from AWS's $55.04 to Azure's $98.32, even before factoring in reserved or committed discounts, which can compress AWS and Google Cloud rates to well under half their on-demand list price. The a3-ultragpu-8g row is included specifically because it is one of the most common near-miss searches in this space and its H200 hardware should not be mistaken for H100 capacity when comparing quotes.

Table 2 ranks specialized providers and marketplaces from cheapest to most expensive advertised per-GPU H100 rate, to directly answer "cheapest H100 cloud instance" and "8x H100 instance cost comparison" style queries, again reflecting figures already cited in the prose above.

PROVIDER	BILLING MODEL	ADVERTISED H100 RATE ($/GPU/HR)	NOTES
Vast.ai	Peer-to-peer marketplace	From $1.87 (PCIe) to $1.92 (SXM) (Source: vast.ai)	Wide range up to ~$4.67/hr; host reliability varies
RunPod	Community Cloud	$1.99 (Source: www.runpod.io)	Third-party vetted data centers
RunPod	Secure Cloud	$2.39 to $2.89 (page discrepancy) (Source: www.runpod.io)	Tier 3/4 facilities, enhanced compliance
Lambda Labs	Self-serve on-demand	$3.29 (PCIe) to $3.99 (SXM) (Source: lambda.ai)	No egress fees advertised
Nebius	On-demand (as of June 2026)	$3.85 (Source: docs.nebius.com)	Single region (eu-north1) for H100 NVLink
CoreWeave	Spot	$19.71 total, 8-GPU node (~$2.46/GPU) (Source: www.coreweave.com)	Kubernetes-native platform
CoreWeave	On-demand	$6.16 (Source: computeprices.com)	8-GPU HGX H100 nodes, InfiniBand
Oracle Cloud (OCI)	Bare metal on-demand	$10.00 (Source: computeprices.com)	No virtualization overhead

The spread in Table 2, from roughly $1.87 to $10.00 per GPU-hour among specialized providers alone, before even reaching hyperscaler list prices, is consistent with an independent aggregator's finding that H100 pricing across 47 tracked providers varies from a $0.34 spot floor to a $3.41 average, a spread the aggregator itself calls significant (Source: getdeploying.com. Community-sourced tracking on Reddit has independently observed price variance of up to 13.8x for H100 rentals across the market, broadly consistent with the range this report documents from primary vendor pages.

Comparative Context and Market Positioning

Positioning these figures against one another clarifies what buyers are actually paying for at each tier. Hyperscaler on-demand pricing, AWS's $55.04, Google's $88.49 to $93.40, and Azure's $98.32 per 8-GPU hour (Source: instances.vantage.sh), buys deep integration with each provider's broader platform: managed Kubernetes, native identity and billing integration, enterprise support SLAs, and, in AWS's case specifically, UltraCluster-scale networking that can span up to 20,000 GPUs in a single nonblocking fabric, a scale enabled by the same 400 Gb/s-class InfiniBand fabrics Azure documents for its own ND H100 v5 series (Source: learn.microsoft.com). Reserved and committed pricing brings hyperscaler rates down substantially, AWS's Capacity Blocks to $41.528 per instance hour and Google's three-year CUD to $38.86 per hour, but both require either advance capacity planning or multi-year budget commitment that many organizations, particularly startups and research labs with uncertain compute trajectories, are unwilling or unable to make.

Neoclouds and marketplaces occupy the opposite end of the flexibility-versus-price spectrum. Oracle's flat $10.00 per GPU-hour bare metal rate (Source: www.oracle.com) and CoreWeave's $6.16 per GPU-hour on-demand rate (Source: www.coreweave.com) sit meaningfully below AWS on-demand pricing while still offering dedicated, non-marketplace infrastructure with InfiniBand or equivalent high-speed networking, and Oracle separately markets this bare-metal approach as eliminating virtualization overhead entirely (Source: www.oracle.com). Lambda Labs and RunPod push further down, into the $2 to $4 per GPU-hour range, generally by offering less networking sophistication, fewer regions, or reduced compliance

certifications relative to the hyperscalers, a positioning RunPod's own pricing page corroborates with its $2.89 per hour Secure Cloud H100 listing (Source: www.runpod.io). Vast.ai sits at the true price floor, but as a marketplace rather than a managed cloud, meaning the buyer inherits variable host reliability, inconsistent support responsiveness, and less predictable capacity availability in exchange for the lowest headline rate.

A useful heuristic that emerges from this research is that the "right" H100 provider depends heavily on job duration and networking sensitivity. For multi-week, multi-node pretraining runs where a single node failure can cascade into lost training time across hundreds of GPUs, the premium for AWS UltraClusters, Azure's InfiniBand-networked ND H100 v5, or CoreWeave's InfiniBand-based HGX H100 nodes is frequently justified by reduced checkpoint-restart overhead alone. For single-node fine-tuning, batch inference, or short-lived experimentation, the price gap between a $98.32 Azure ND H100 v5 node and a $1.99 RunPod Community Cloud H100 is difficult to justify on technical grounds, and buyers optimizing purely for cost per training-hour should default toward the neocloud or marketplace tier unless a specific compliance, support, or networking requirement rules it out.

Hidden costs also matter to a total-cost comparison, and hyperscalers are not always transparent about them in headline pricing. As referenced in the Azure section above, one third-party comparison itemizes outbound data egress at roughly $0.087 per GB and premium managed-disk storage at roughly $0.17 per GB per month on top of Azure's ND H100 v5 compute rate (Source: www.spheron.network), line items that a pure per-GPU-hour comparison omits entirely. Because that estimate comes from a competing neocloud vendor with an incentive to make hyperscaler total cost of ownership look unfavorable, it should be treated as an illustrative order of magnitude rather than a universal constant, but the underlying categories, egress, persistent storage, and support-tier fees, are real components of any hyperscaler GPU bill.

Currency of pricing data matters more in this market than in most cloud comparisons. Nebius raised its H100 on-demand rate by roughly 30 percent (from $2.95 to $3.85 per GPU-hour) effective June 1, 2026, even as the broader market, per aggregator tracking, trended about 4 percent cheaper over the preceding twelve months (Source: getdeploying.com). This divergence, one provider raising prices as the aggregate market falls, illustrates that individual vendor pricing decisions are influenced by regional supply, contractual commitments to large customers, and competitive positioning against next-generation Blackwell capacity, not solely by aggregate H100 supply and demand. Nvidia's chip allocation decisions add a further layer of unpredictability: the company itself acknowledged in November 2024 that no competitor could match the performance specifications of its Blackwell line, a factor analysts say encourages some large buyers to wait for next-generation allocations rather than lock in current-generation H100 capacity at any price (Source: www.reuters.com).

Data Analysis and Evidence

Quantifying the H100's continued relevance requires looking beyond sticker price to the compute capacity actually being deployed and the benchmark performance that capacity delivers. NVIDIA's own datasheet specifies the H100 SXM variant at 3,958 teraFLOPS of FP8 Tensor Core throughput (with sparsity) and the PCIe variant at 3,026 teraFLOPS, alongside 80GB of HBM3 memory on both form factors and up to 900GB/s of NVLink bandwidth on SXM versus 600GB/s on PCIe (Source: www.megware.com). The SXM form factor supports a configurable thermal design power of up to 700 watts, compared to 300 to 350 watts for PCIe cards (Source: www.megware.com, a difference that directly explains why SXM-based instances (AWS P5, Azure ND H100 v5, CoreWeave HGX H100, Lambda's SXM tier) command a price premium over PCIe-based instances (Lambda's PCIe tier, RunPod's PCIe listings) even when GPU counts are identical: the SXM form factor delivers meaningfully higher sustained throughput per card. Up to 256 H100 GPUs can be linked via the NVLink Switch System to accelerate exascale-class workloads as a single fabric, a scale-out capability the same datasheet describes explicitly (Source: www.megware.com), and NVIDIA's own product page adds that the chip triples FP64 double-precision throughput to 60 teraflops for high-performance computing workloads (Source: www.nvidia.com and that new DPX instructions deliver up to 7x higher performance over the prior A100 generation on dynamic programming algorithms used in fields like genome sequencing (Source: www.nvidia.com).

Independent, third-party-verified benchmark results reinforce the H100's continued competitiveness. In the MLCommons MLPerf Training v4.1 round, published in November 2024 and still the most recent large-scale verified comparison spanning Hopper and early Blackwell submissions, NVIDIA's Hopper architecture continued to deliver the highest verified performance among broadly available solutions, and a submission using 11,616 H100 GPUs continued to hold the benchmark record for both overall performance and submission scale on the GPT-3 175B pretraining benchmark (Source: developer.nvidia.com). NVIDIA's early Blackwell preview submissions in that same round delivered per-GPU performance boosts of 2x for GPT-3 pretraining and 2.2x for Llama 2 70B low-rank adaptation fine-tuning relative to Hopper (Source: developer.nvidia.com), giving buyers a concrete reference point for the performance-per-dollar tradeoff between renting H100 today and waiting for or paying a premium for Blackwell-generation capacity. Separately, on the Llama 2 70B LoRA fine-tuning benchmark, an 8-GPU H200 submission delivered about 16 percent better performance than the equivalent H100 submission (Source: developer.nvidia.com), a figure directly relevant to anyone weighing Google's H200-based a3-ultragpu-8g against its H100-based a3-highgpu-8g or a3-megagpu-8g siblings, since it implies the H200 SKU's roughly 4 percent lower on-demand price comes bundled with a meaningful performance uplift on at least some workloads, not merely more memory headroom.

Inference economics have also become part of the H100 value calculation. NVIDIA's own H100 product page cites third-party SemiAnalysis InferenceX benchmarks putting H100 inference cost at approximately $0.09 per million tokens at 66 tokens per second per user for a GPT-OSS-120B model using the vLLM serving framework, as of April 2026 (Source: www.nvidia.com), a data point that helps translate raw per-GPU-hour rental prices into the per-token economics that inference-heavy buyers ultimately care about.

At the market level, the capital flowing into H100 and successor-generation capacity remains enormous. Industry analysis citing Goldman Sachs projects Big Five hyperscaler (Amazon, Microsoft, Google, Meta, Oracle) capital expenditure will exceed $600 billion in 2026 (Source: introl.com), a 36 percent increase over 2025, with roughly 75 percent, or about $450 billion, directed specifically at AI infrastructure. Of that AI infrastructure spend, GPUs and accelerators alone are estimated at roughly $180 billion (Source: introl.com, equating, as noted in the Executive Summary, to something on the order of 6 million GPU-equivalents purchased at an average price near $30,000 apiece, a figure broadly consistent with RunPod's cited $25,000 to $30,000 street price for a single H100 card. At the individual company level, Amazon's 2025 capital expenditure of roughly $125 billion reflected a 61 percent year-over-year increase, with the majority directed to AWS AI infrastructure (Source: introl.com), while Microsoft's 2025 capital expenditure of approximately $95 billion was directed largely at Azure AI infrastructure, underscoring how consistently AI compute has become the dominant driver of hyperscaler spending across providers (Source: introl.com). Nvidia's own commentary to investors reinforces the supply picture behind this spending: in November 2024 the company forecast that demand would exceed supply for several quarters into fiscal 2026, even as it was already selling chips as fast as TSMC's manufacturing lines could turn them out, as referenced in the Introduction above.

Cross-referencing multiple independent sources on cloud GPU pricing consistently surfaces the same qualitative pattern even where exact figures diverge: hyperscaler on-demand rates for 8-GPU H100 nodes cluster in the $55 to $98 per hour range, specialized neoclouds cluster in the $6 to $10 per GPU-hour range, and marketplace or spot pricing can push per-GPU rates under $2. This report treats each individual figure as a snapshot subject to the volatility documented above (Nebius's mid-2026 repricing being the clearest example) rather than a fixed constant, and recommends buyers treat any single-source H100 price quote as a starting point for negotiation and comparison rather than a market-clearing rate.

Case Studies and Real-World Examples

Anthropic on AWS EC2 P5 UltraClusters. Anthropic, an AI safety and research company building large-scale foundation models, was an early adopter of AWS's P5 instances after using P4d instances extensively for prior training runs. Cofounder Tom Brown stated the company expected P5 instances "to deliver substantial price-performance benefits over P4d instances" and that they would be available "at the massive scale required for building next-generation LLMs and related products" (Source: aws.amazon.com), illustrating how a frontier AI lab evaluated the H100-based P5 generation specifically on cost-per-unit-of-training-progress rather than raw hourly price.

Cohere on AWS EC2 P5 for multi-cloud LLM deployment. Cohere, an enterprise language AI company, adopted P5 instances as part of a deliberate multi-cloud strategy that deploys its models across whichever platform best suits a given enterprise customer's existing data environment. CEO Aidan Gomez said NVIDIA H100-powered Amazon EC2 P5 instances would "unleash the ability of businesses to create, grow, and scale faster" when combined with Cohere's language and generative AI capabilities (Source: aws.amazon.com), a case that illustrates H100 cloud pricing decisions being driven by customer data-residency and platform-fit requirements as much as by raw compute economics.

Microsoft and CoreWeave's multi-billion dollar H100 supply agreement for OpenAI. In 2023, Microsoft signed what CNBC reported was potentially a multi-billion dollar, multi-year cloud infrastructure agreement with CoreWeave, specifically to secure sufficient GPU capacity for OpenAI's compute demands running on Microsoft's Azure infrastructure (Source: www.cnbc.com), as referenced in the Neoclouds section above. At the time of that agreement, NVIDIA itself had invested $100 million directly into CoreWeave, valuing the young GPU cloud specialist at $2 billion. This case is instructive because it shows a hyperscaler (Microsoft) turning to a neocloud (CoreWeave) rather than expanding its own H100 fleet alone, precisely because aggregate H100 demand from a single anchor customer (OpenAI) exceeded what Azure's own build-out could satisfy on the required timeline, a pattern that has repeated across the industry as multiple hyperscalers now lease neocloud capacity to supplement proprietary fleets.

Google Cloud's internal DeepMind and research use of H100 and A100 GPUs. At the general availability launch of Google Cloud's A3 instances, as referenced in the Google Cloud section above, Google Cloud CEO Thomas Kurian and NVIDIA CEO Jensen Huang discussed how Google itself was using NVIDIA H100 and A100 GPUs for internal research and inference across DeepMind and other divisions (Source: blogs.nvidia.com). This case demonstrates that hyperscaler H100 pricing decisions are not purely external-facing; internal research consumption at the scale of an organization like DeepMind directly competes for the same physical GPU inventory that external A3 customers rent, a dynamic that helps explain periods of constrained availability and pricing firmness even as list prices for external customers hold steady or decline.

Implications and Future Directions

Several forward-looking implications follow from this pricing landscape. First, the H100 is likely to remain commercially relevant for several more years as a "value tier" accelerator even as H200, B200, and B300 capacity scales, precisely because MLPerf data shows Hopper-generation hardware still holds a genuine performance and cost-efficiency niche for many training and fine-tuning workloads (Source: developer.nvidia.com), and because neocloud and marketplace pricing on H100 will likely continue falling faster than pricing on newer chips as H100 supply depreciates and hyperscalers rotate fleets toward Blackwell-generation silicon. Second, the price gap between hyperscaler on-demand rates and neocloud or marketplace rates, often a factor of five to ten times per GPU-hour, is unlikely to close entirely, because it reflects durable structural differences (proprietary networking fabrics, enterprise support SLAs, compliance certifications, guaranteed capacity) rather than temporary market inefficiency, though committed-use and capacity-block mechanisms will likely continue narrowing the gap for buyers willing to plan ahead.

Third, buyers should expect continued naming and SKU confusion, as illustrated by Google's a3-ultragpu-8g shipping H200 rather than H100 silicon and Azure's parallel NC and ND H100 product lines targeting entirely different workload profiles (Source: learn.microsoft.com); cloud providers have strong incentives to keep family names consistent (A3, ND) even as the underlying silicon generation changes within that family, and procurement teams should always confirm the specific GPU model, not just the family name, before signing a contract or committing budget. Fourth, given the hundreds of billions of dollars in hyperscaler capital expenditure directed at AI infrastructure through at least 2027, and Nvidia's own acknowledgment that supply will trail demand for multiple quarters (Source: www.reuters.com), near-term capacity constraints could periodically reassert upward pricing pressure on any given GPU generation, including bare-metal H100 shapes such as Oracle's BM.GPU.H100.8 (Source: www.oracle.com), in specific regions or availability zones even as the broader multi-provider average continues to trend downward.

Finally, the increasing role of neoclouds in supplying capacity to hyperscalers themselves, as the Microsoft-CoreWeave relationship illustrates, suggests that the boundary between "hyperscaler" and "specialized GPU cloud" pricing tiers will likely continue blurring, with hyperscalers increasingly reselling or white-labeling neocloud capacity for specific customer commitments even as they continue operating higher-margin, fully proprietary infrastructure for their own flagship instance families (Source: www.oracle.com). Organizations planning multi-year AI infrastructure budgets should build in explicit re-evaluation checkpoints, at minimum annually, given the pace of change documented in this report, and should model hidden costs like egress fees, persistent storage, and support tiers alongside the headline per-GPU-hour rate rather than comparing sticker prices alone.

Frequently Asked Questions (FAQs)

What does an H100 GPU cost per hour in the cloud as of mid-2026? Single-GPU H100 rental rates span roughly $1.87 to $10.00 per hour depending on provider and billing model, with hyperscaler 8-GPU on-demand nodes ranging from $55.04 (AWS) to $98.32 (Azure) per hour, and specialized providers like CoreWeave and Oracle offering $6.16 (Source: computeprices.com) to $10.00 per GPU-hour (Source: www.oracle.com) (Source: computeprices.com), as detailed in the sections above.

What is the cheapest H100 cloud instance available today? Marketplace platform Vast.ai advertises the lowest headline rates, with H100 PCIe listings as low as $1.87 per hour, though buyers should weigh the lower price against variable host reliability inherent to a peer-to-peer marketplace model, and RunPod's Community Cloud offers a similar entry point at $1.99 per hour for those wanting a vetted, non-marketplace alternative (Source: www.runpod.io) (Source: www.runpod.io).

Is Google Cloud's a3-ultragpu-8g an H100 instance? No. Despite belonging to the "A3" family alongside the H100-based a3-highgpu-8g and a3-megagpu-8g, the a3-ultragpu-8g SKU is built around 8 NVIDIA H200 GPUs, each with 141GB of HBM3e memory, not H100 GPUs, as established in the Google Cloud pricing section above, and independently confirmed by a third-party pricing tracker's own hardware specification listing (Source: cloudprice.net), a distinction NVIDIA's own account of the A3 launch also frames around the family's H100 origins (Source: blogs.nvidia.com).

How does AWS p5.48xlarge pricing compare to Google Cloud and Azure H100 instances? AWS's p5.48xlarge lists at $55.04 per hour on demand, cheaper on a headline basis than Google Cloud's H100-based a3-highgpu-8g ($88.49) and a3-megagpu-8g ($93.40) and Azure's ND96isr H100 v5 ($98.32) (Source: instances.vantage.sh), though all three providers offer reserved, committed-use, or capacity-block mechanisms that substantially reduce effective rates for planned workloads, as summarized in Table 1 above.

What is AWS's Capacity Blocks for ML program and how does it affect H100 pricing? Capacity Blocks let customers reserve a specific quantity of GPU capacity for a defined future time window without a long-term contract, pricing p5.48xlarge at an effective $41.528 per instance hour in most US regions, a discount of roughly 25 percent versus the $55.04 on-demand rate, as detailed in the AWS pricing section above.

Can H100 pricing increase over time, or does it only fall? It can increase. Nebius raised its H100 NVLink on-demand rate from $2.95 to $3.85 per GPU-hour effective June 1, 2026 (Source: docs.nebius.com), even as broader market tracking showed roughly 4 percent average price declines over the prior year, demonstrating that individual vendor pricing does not always track the aggregate market direction, as discussed in the Neoclouds and Comparative Context sections above.

Does spot or preemptible pricing make sense for H100 training workloads? Spot and preemptible pricing, such as AWS's $19.77 per hour p5.48xlarge spot rate (Source: instances.vantage.sh) or CoreWeave's $19.71 per hour HGX H100 spot rate (Source: www.coreweave.com), offers discounts of roughly 60 to 65 percent versus on-demand, but requires checkpointed, fault-tolerant training pipelines capable of resuming after unplanned reclamation, making it best suited to research and fine-tuning workloads rather than time-critical production training runs.

Are there hidden costs beyond the advertised per-hour H100 rate? Yes. Beyond the headline compute rate, hyperscalers typically bill separately for data egress, persistent block or file storage, and premium support tiers, all of which sit on top of the per-GPU-hour rate and are frequently omitted from simple price comparisons, as discussed in the Azure pricing and Comparative Context sections above.

Does the H100 remain competitive against the H200 and Blackwell-generation chips? Largely yes for many workloads. NVIDIA's own datasheet shows the H100 SXM delivering 3,958 teraFLOPS of FP8 throughput with sparsity (Source: www.megware.com), and MLPerf Training v4.1 results confirm Hopper-based systems, including an 11,616 GPU H100 submission, continued to set overall performance and scale records in the most recent verified round (Source: developer.nvidia.com), even though early Blackwell submissions showed 2x to 2.2x higher per-GPU performance on select benchmarks.

Conclusion

Comparing H100 GPU cloud pricing across AWS, Google Cloud, and Azure in mid-2026 reveals a market with genuine, well-documented price dispersion rather than a single "market rate." Hyperscaler on-demand pricing for an 8-GPU H100 node spans from $55.04 at AWS to $98.32 at Azure, with Google Cloud's H100-based SKUs sitting between the two, while committed-use and capacity-block mechanisms can bring effective rates down toward $38 to $42 per hour for buyers willing to plan ahead. Specialized neoclouds and marketplaces, including CoreWeave, Oracle, Lambda Labs, Nebius, RunPod, and Vast.ai, consistently undercut hyperscaler on-demand pricing, often by a factor of five or more per GPU-hour, in exchange for tradeoffs in networking sophistication, compliance certification, or infrastructure reliability that matter for some workloads and not others.

The single most important buyer takeaway is that the family name on a GPU instance does not guarantee the underlying chip generation: Google Cloud's a3-ultragpu-8g ships H200 rather than H100 silicon despite its A3 branding, and Azure separates its single or dual-GPU NC H100 v5 applied-AI line from its eight-GPU ND H100 v5 distributed-training line under naming that can easily be conflated. Buyers should always confirm the specific GPU model, memory configuration, and networking fabric attached to any quoted instance before treating a price comparison as valid. Given the pace of change already documented within this single research window, Nebius's mid-2026 price increase running counter to a broader market-wide decline being the clearest example, any organization budgeting for sustained H100 usage should build in a recurring pricing review rather than treating today's rate card as fixed, and should weight the decision between hyperscaler, neocloud, and marketplace tiers primarily on job duration, networking sensitivity, hidden ancillary fees, and compliance requirements rather than on advertised hourly rate alone.

Tags: h100 gpu pricing, gpu cloud comparison, aws p5.48xlarge, google cloud a3, azure nd h100 v5, coreweave pricing, gpu cloud cost, nvidia hopper, cloud computing, ai infrastructure

---