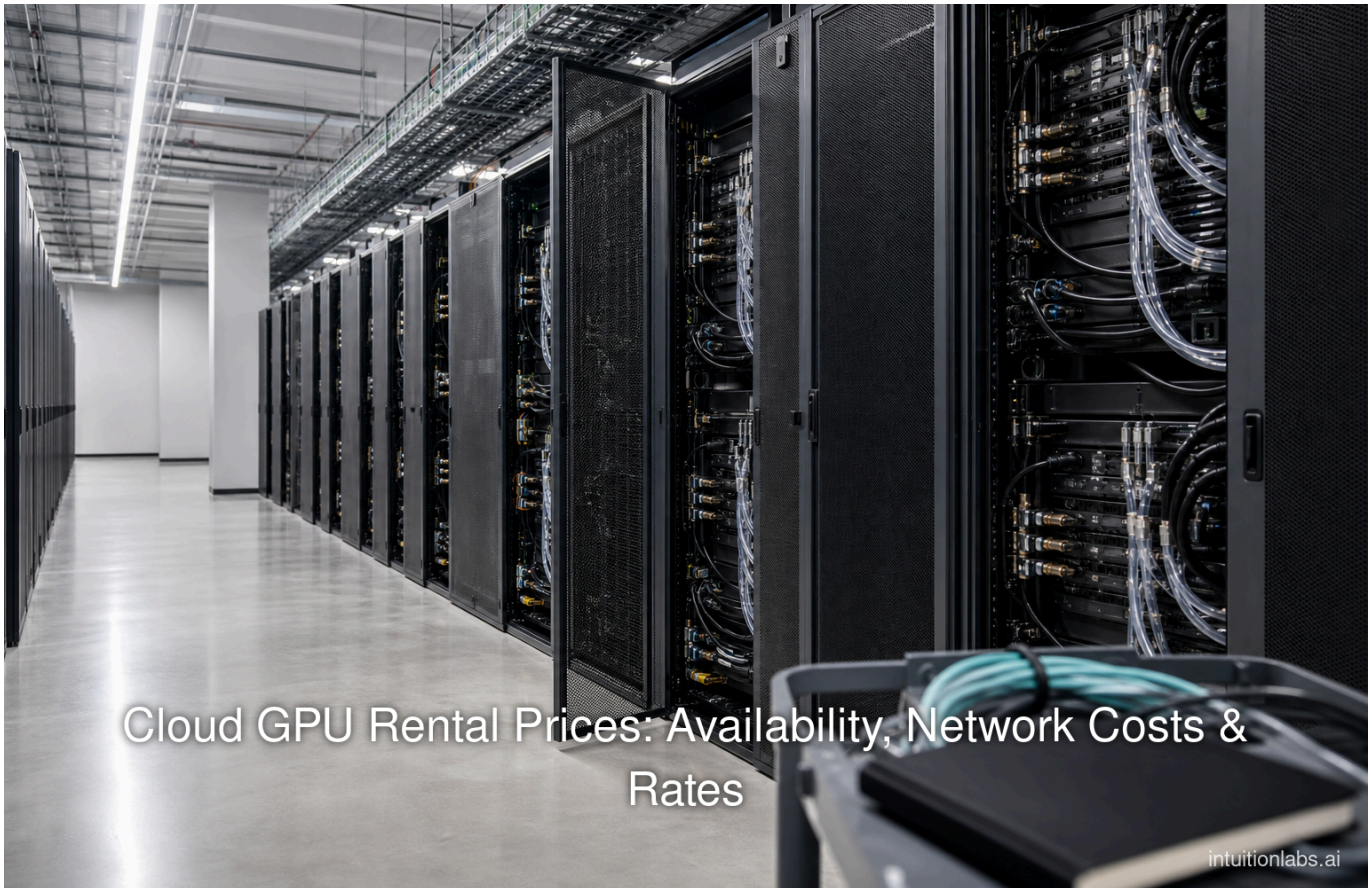
Cloud GPU Rental Prices: Availability, Network Costs & Rates

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · cloud gpu rental prices · gpu cloud pricing comparison · h100 rental price · a100 rental price · gpu cloud providers · cloud gpu availability · gpu network egress costs · on-demand vs spot gpu pricing · gpu as a service · cheapest cloud gpu



Executive Summary

Cloud **graphics processing unit (GPU)** rental prices for AI training and inference diverge sharply depending on which layer of the market a buyer shops in, and the gap has widened rather than narrowed through 2026. On the three hyperscalers, an on-demand 8-GPU **NVIDIA H100** node costs $55.04 per hour on **Amazon Web Services (AWS)** p5.48xlarge in US East (pricing.us-east-1.amazonaws.com), $88.49 per hour on **Google Cloud** a3-highgpu-8g (cloud.google.com), and $98.32 per hour on **Microsoft Azure** ND96isr H100 v5 (prices.azure.com). GPU-cloud specialists price the same silicon far lower: **CoreWeave** lists an on-demand 8x H100 HGX node at $49.24 per hour (www.coreweave.com), while **Lambda** prices a single H100 SXM GPU at $3.99 per hour (lambda.ai) and **RunPod** prices an H100 SXM pod at $3.29 per hour (www.runpod.io). Marketplace platforms go lower still: **Vast.ai** lists H100 SXM rentals from $1.33 per hour with a median of $2.00 per hour (vast.ai), because it auctions spare capacity from independent data centers rather than selling a fixed-rate product.

The market-level data corroborate this stratification. Cloud cost-tracking firm **CloudZero** puts the median on-demand H100 rate at approximately $4.17 per hour on dedicated GPU clouds versus $7.89 per hour on hyperscalers, a roughly 90 percent premium for the large public clouds (www.cloudzero.com). Independent research firm **SemiAnalysis** tracked H100 one-year rental contract prices rising almost 40 percent, from a low of $1.70 per GPU-hour in October 2025 to $2.35 per GPU-hour by March 2026, reflecting acute capacity tightness rather than a falling-price trend that buyers might otherwise expect from maturing hardware (newsletter.semianalysis.com). Reserved and committed-use pricing changes the calculus substantially: Google Cloud's 3-year committed use discount brings the same a3-highgpu-8g node down to $38.86 per hour (cloud.google.com), and GPU-cloud specialists advertise comparable committed-usage savings, detailed in the sections below.

Beyond the headline hourly rate, three variables that this report treats as first-class comparison criteria change the effective cost materially: network egress (ranging from $0 at Crusoe Cloud and CoreWeave to $0.09 to $0.12 per gigabyte on the hyperscalers), storage (from $0.05 per gigabyte-month at RunPod to $0.17 per gigabyte-month for Google Cloud SSD persistent disk), and interconnect quality (400 Gbps dedicated **InfiniBand** per GPU on Azure's ND H100 v5 versus best-effort networking on marketplace platforms). The **GPU as a Service** market itself is forecast very differently by different research firms, a discrepancy this report presents rather than resolves: **Grand View Research** projects $14.46 billion by 2033 (www.grandviewresearch.com), while **Fortune Business Insights** projects $162.54 billion by 2034 (www.fortunebusinessinsights.com). This report lays out dated, provider-by-provider pricing observations as of September 2026, a reproducible method for converting a list price into an effective hourly rate, and the availability and commitment terms that determine which number a given buyer will actually pay.

Introduction and Background

Renting **GPU** compute has become the default way that AI teams access **NVIDIA H100** and **A100** accelerators, since buying and hosting the hardware directly requires capital, data-center space, and power provisioning that only a small number of organizations can justify. The market that has grown around this demand now spans three distinct layers: the hyperscalers (AWS, Google Cloud, Azure) that bundle GPUs into a broader enterprise cloud, GPU-cloud specialists (Lambda, CoreWeave, RunPod) that build infrastructure specifically for AI workloads, and marketplace or neocloud platforms (Vast.ai, Paperspace, Crusoe Cloud) that aggregate capacity from many operators or specialize in cost efficiency. Each layer prices, bills, and provisions GPUs differently enough that a single "cloud GPU price" figure is not meaningful without specifying the provider, the GPU generation, the commitment term, and the date of observation.

This divergence has intensified because underlying GPU supply has tightened rather than eased. SemiAnalysis's rental-price index shows H100 one-year contract pricing rising sharply between late 2025 and early 2026 (newsletter.semianalysis.com), and a component-level shortage is a contributing factor: the three dominant memory manufacturers supplying **high-bandwidth memory (HBM)** for accelerators are reported to be sold out for 2026 (www.cnbc.com). Against that backdrop, **IDC** raised its 2026 AI infrastructure spending forecast to $497 billion, roughly 56 percent year-over-year growth (www.idc.com), meaning demand for rented GPU capacity is growing at the same time supply is constrained.

IntuitionLabs, a life-sciences and AI consultancy, advises pharmaceutical and biotechnology teams on the infrastructure decisions that sit underneath their AI programs, without operating a competing GPU cloud itself; its own public positioning describes work that helps "pharmaceutical and life science organizations shape the future" of their operations (intuitionlabs.ai) through governed, department-level AI adoption rather than raw infrastructure resale (intuitionlabs.ai). For a life-sciences AI team, the GPU rental decision typically follows, rather than precedes, decisions about model architecture, data governance, and validation requirements, which is why this report treats price and availability as inputs to a broader capacity-planning process rather than as the whole decision. A companion IntuitionLabs analysis focuses specifically on H100 rental pricing across AWS Capacity Blocks and RunPod pods; that piece is a useful narrower reference and this report does not restate its findings, instead extending the comparison across the full 2026 provider landscape, additional GPU generations, and the network, storage, and commitment variables that determine effective cost (intuitionlabs.ai).

The remainder of this report is organized by provider layer, followed by a dedicated methodology and data section, a worked example of the effective-rate calculation, and a discussion of where the market is headed. All prices below are on-demand list prices observed as of September 5, 2026, in the region stated, unless otherwise noted; providers frequently adjust rates without notice, so readers using this report for procurement should re-verify the current price against the cited vendor page before committing budget.

Hyperscaler Cloud Pricing: AWS, Google Cloud, and Azure

The three hyperscalers price GPUs as part of much larger compute catalogs, bundling networking, storage, and support into enterprise agreements that specialist providers do not offer. On **AWS**, the flagship H100 instance is the **p5.48xlarge**, an 8-GPU node priced at $55.04 per hour on-demand in US East (N. Virginia), with a single-GPU p5.4xlarge available at $6.88 per hour (pricing.us-east-1.amazonaws.com). AWS's A100-based **P4d** family remains available for buyers who do not need H100-class throughput: the 8-GPU p4d.24xlarge (40GB A100) runs $21.96 per hour, and the memory-upgraded p4de.24xlarge (80GB A100) runs $27.45 per hour, both in us-east-1. A 3-year no-upfront reserved p5.48xlarge drops the effective on-demand-equivalent rate to $23.78 per hour, a discount of roughly 57 percent versus on-demand, illustrating how heavily AWS's list price assumes short-term, unreserved use (pricing.us-east-1.amazonaws.com). AWS states that P5 instances provide up to 640GB of aggregate **HBM3** memory across 8 H100 GPUs (aws.amazon.com) and up to 3,200 Gbps of **Elastic Fabric Adapter (EFA)** networking on P5, P5e, and P5en variants (aws.amazon.com).

Google Cloud prices its **A3 High** (a3-highgpu-8g, 8x H100) instance at $88.49 per hour on-demand, dropping to $61.38 per hour with a 1-year committed use discount and $38.86 per hour with a 3-year commitment, while the higher-throughput **A3 Mega** (a3-megagpu-8g) lists at $93.40 per hour on-demand, falling to $40.65 per hour on a 3-year commitment (cloud.google.com). Google's A100 instances, the **A2 Ultra** (80GB A100) and **A2 Standard** (40GB A100), are priced at $40.55 and $29.39 per hour on-demand respectively (cloud.google.com). Google's GPU documentation states that its H100 configuration provides 80GB of HBM3 per GPU at 3.35 terabytes per second of bandwidth, connected by an **NVLink** full mesh at 900 gigabytes per second (cloud.google.com).

Microsoft Azure's comparable flagship, the **ND96isr H100 v5**, lists at $98.32 per hour on-demand in US East, the highest headline H100 node price among the three hyperscalers observed (prices.azure.com), while its A100-based **NDm A100 v4** instance (8x 80GB A100) runs $32.77 per hour (prices.azure.com). Azure's technical documentation states that each GPU in the ND H100 v5 series receives its own dedicated 400 Gbps **NVIDIA Quantum-2**

InfiniBand connection, independent of the other 7 GPUs in the node (learn.microsoft.com), and that its NDm A100 v4 clusters can scale to thousands of GPUs with 1.6 terabytes per second of aggregate interconnect bandwidth (learn.microsoft.com), a specification aimed squarely at large distributed training jobs.

Ancillary costs differ enough between the three to change a total bill materially. AWS's general-purpose gp3 block storage runs $0.08 per gigabyte-month (aws.amazon.com), Azure's Premium SSD v2 costs $0.081 per gigabyte-month (azure.microsoft.com), and Google Cloud's SSD persistent disk works out to roughly $0.17 per gigabyte-month based on its published $34 monthly cost for a 200GB volume (cloud.google.com). Network egress is the more consequential difference for GPU workloads that move large datasets or model checkpoints: AWS charges $0.09 per gigabyte for the first 10 terabytes of monthly internet egress after a 100GB free allowance (pricing.us-east-1.amazonaws.com) (aws.amazon.com), Azure charges $0.087 per gigabyte at the same tier (azure.microsoft.com), and Google Cloud charges $0.12 per gibibyte for the first 1,024 gibibytes per month (cloud.google.com).

GPU Cloud Specialists: Lambda, CoreWeave, and RunPod

A second tier of providers builds infrastructure specifically for AI training and inference rather than treating GPUs as one SKU among thousands. **Lambda** prices a single on-demand H100 SXM GPU (80GB) at $3.99 per hour, and an A100 SXM GPU (80GB) at $2.79 per hour, with a 40GB A100 SXM variant at $1.99 per hour (lambda.ai). Its PCIe-form-factor H100 starts at $3.29 per GPU-hour for single-GPU configurations (lambda.ai). For multi-week or multi-month training runs, Lambda's **1-Click Clusters** offer dedicated, InfiniBand-connected H100 nodes on 2-week to 1-year terms at $6.16 per GPU-hour for a 16-GPU cluster (lambda.ai) (lambda.ai), a price point above its on-demand SXM rate because the cluster product guarantees dedicated, low-latency multi-node bandwidth. Lambda bills on-demand instances in one-minute increments with no stated minimum spend (docs.lambda.ai), and its persistent filesystem storage carries no minimum storage period and no ingress or egress charge (docs.lambda.ai).

CoreWeave prices an 8-GPU HGX H100 node at $49.24 per hour on-demand, equivalent to about $6.16 per GPU-hour, each node carrying 8 GPUs with 80GB of memory apiece, 128 vCPUs, and 2,048GB of system RAM (www.coreweave.com). Its 8-GPU A100 (80GB) node lists at $21.60 per hour on-demand, and CoreWeave also publishes a spot/interruptible price for the H100 node at $19.71 per hour in North America, roughly 60 percent below its on-demand rate, alongside reserved-capacity discounts of up to 60 percent off on-demand for committed usage (www.coreweave.com). CoreWeave charges nothing for data transfer between its network and the public internet, and its Distributed File Storage runs $0.070 per gigabyte-month with free egress and IOPS on that storage tier (www.coreweave.com).

RunPod prices an on-demand H100 SXM pod (80GB) at $3.29 per hour, an H100 PCIe pod at $2.89 per hour, and an H100 NVL pod (94GB) at $3.19 per hour; its A100 SXM pod (80GB) runs $1.59 per hour and its A100 PCIe pod $1.39 per hour (www.runpod.io). RunPod separates a lower-cost **Community Cloud** tier from a **Secure Cloud** tier, and offers both per-hour and per-second billing depending on the pod type, operating across more than 30 regions (www.runpod.io). Storage on RunPod starts at $0.05 per gigabyte-month, with separate rates for container disk, volume disk, and standard versus high-performance network storage tiers; its pricing page carried a stated update date of July 27, 2026 at the time of this report's research (www.runpod.io).

Marketplace and Neocloud Providers: Vast.ai, Paperspace, and Crusoe Cloud

A third category of provider either aggregates spare capacity from many independent data-center operators or specializes narrowly on cost-efficient dedicated infrastructure. **Vast.ai** is explicitly an auction-style marketplace rather than a fixed-rate platform: its H100 SXM listings range from a low of $1.33 per hour to a median of $2.00 per hour, its H100 PCIe listings from $1.86 to a $3.07 median, and its A100 SXM4 listings from $0.33 to a $0.83 median (vast.ai). Vast.ai's own documentation is explicit that "prices for base rental, storage, and bandwidth vary considerably from machine to machine" because the platform aggregates over 40 independent data centers with supply-and-demand pricing rather than a vendor-set rate card (docs.vast.ai). Instances bill by the second (docs.vast.ai), storage fees continue even on stopped instances (docs.vast.ai), and a reserved tier offers up to 50 percent off the on-demand marketplace rate on 1, 3, or 6-month terms (vast.ai). This variability is the marketplace model's central trade-off: buyers can find the cheapest H100 capacity in the market on Vast.ai, but the price, the host's reliability, and the network path are all less standardized than on a single-operator cloud, since prices are set by supply and demand across more than 40 data centers rather than a single vendor rate card.

Paperspace, operated by DigitalOcean, lists a single H100 GPU at $3.09 per hour as a newly available configuration (www.paperspace.com), separate from a promotional on-demand rate of $5.95 per hour that the vendor states applies under a specific promo pricing structure, a discrepancy this report flags rather than resolves since Paperspace's page presents both figures (www.paperspace.com). Its $2.24 per hour A100 rate is explicitly tied to a 3-year commitment rather than being an on-demand price, and it also offers an 8-GPU HGX H100 multi-GPU node configuration (www.paperspace.com). Paperspace advertises free, unlimited network bandwidth as a platform feature, charges $0.29 per gigabyte for storage beyond each plan's included allotment, and bills instances hourly with no stated per-second or per-minute granularity (www.paperspace.com).

Crusoe Cloud publishes a fixed (non-auction) on-demand rate of $3.90 per GPU-hour for its H100 80GB HGX configuration and $2.30 per GPU-hour for its A100 80GB SXM configuration, billed per-hour or per-second with no minimum commitment (www.crusoe.ai). Crusoe is one of the few providers observed to charge nothing at all for network ingress or egress, in either direction, and prices persistent block storage at $0.08 per gibibyte-month (docs.crusoecloud.com). Notably, Crusoe's newest accelerators, the **NVIDIA GB200 NVL72** and **B200**, carry no public self-serve price on either an on-demand or spot basis and require contacting sales for both, an availability signal in itself: the newest hardware generation is not yet available at a published market rate even from a provider that publishes fixed prices for its H100 and A100 fleet (www.crusoe.ai).

Table 1 below summarizes on-demand H100 pricing, memory, and network egress terms across all nine providers covered in this report.

Provider	On-demand H100 rate	GPU memory	Interconnect	Egress (per GB)	Billing granularity
AWS (p5.48xlarge, 8x)	$55.04/hr ($6.88/hr, 1x)	80GB HBM3	Up to 3,200 Gbps EFA	$0.09 after 100GB free	Per second
Google Cloud (a3-highgpu-8g)	$88.49/hr	80GB HBM3	NVLink @ 900 GBps	$0.12/GiB (tiered)	Per second

Provider	On-demand H100 rate	GPU memory	Interconnect	Egress (per GB)	Billing granularity
Azure (ND96isr H100 v5)	$98.32/hr	80GB HBM3	400 Gbps InfiniBand/GPU	$0.087 after 100GB free	Per minute
Lambda (SXM, per GPU)	$3.99/hr	80GB HBM3	InfiniBand (clusters)	Free on filesystems	Per minute
CoreWeave (HGX, 8x)	$49.24/hr (spot $19.71/hr)	80GB per GPU	Not publicly stated	Free	Not publicly stated
RunPod (SXM pod)	$3.29/hr	80GB	Not publicly stated	Not itemized	Per hour or per second
Vast.ai (SXM, marketplace)	$1.33 to $2.00/hr median	80GB	Host-dependent	Host-dependent	Per second
Paperspace	$3.09 to $5.95/hr	80GB (HGX node)	Not publicly stated	Free (unlimited)	Per hour
Crusoe Cloud (HGX, per GPU)	$3.90/hr	80GB HBM3	Not publicly stated	Free	Per hour or per second

The table makes the layered structure of the market visible at a glance: hyperscaler on-demand rates cluster between $55 and $98 per hour for an 8-GPU node, GPU-cloud specialists cluster between roughly $3.29 and $6.16 per GPU-hour, and the Vast.ai marketplace undercuts both by pricing the same H100 SXM hardware from $1.33 per hour. None of these figures is the "right" price in isolation; each reflects a different bundle of reliability, interconnect quality, region coverage, and support that the next section's methodology attempts to make comparable.

Comparative Context and Market Positioning: A Reproducible Method

Because the nine providers above price GPUs on different axes (per-node versus per-GPU, fixed-rate versus auction, hourly versus per-second), a defensible comparison requires converting each listed price into a common **effective hourly rate** rather than comparing headline numbers directly. This report defines that rate as:

Effective hourly rate = ((on-demand GPU-hour price) + (prorated storage cost per hour x expected GB retained) + (egress cost per GB x expected GB transferred per hour)) x (1 / expected uptime fraction, for spot or interruptible capacity)

Applying this method to a single GPU-hour of use with a modest 50GB of retained checkpoint storage and 20GB of hourly data movement changes the ranking in instructive ways. On AWS, the storage term adds roughly $0.0005 per hour and the egress term adds up to $1.80 per hour at the $0.09 per gigabyte rate, meaningfully increasing the effective cost of a data-movement-heavy workload relative to its $6.88 per hour headline single-GPU price (pricing.us-east-1.amazonaws.com). On Crusoe Cloud, the same workload's egress term is zero, so the effective rate equals the headline rate, a policy CoreWeave and Paperspace also follow on their own pricing pages cited above (docs.crusoecloud.com). For interruptible or spot capacity, the uptime-fraction term matters most: CoreWeave's spot H100 price of $19.71 per hour is roughly 60 percent below its on-demand rate, but a workload

that tolerates only 70 percent effective uptime because of interruptions has an effective rate closer to $28.16 per hour once idle-restart time is accounted for, which is still below the $49.24 on-demand rate but by a narrower margin than the sticker price suggests (detailed above).

Independent price-tracking corroborates that reserved and spot discounts, not on-demand list prices, are where most of the achievable savings sit. CloudZero's analysis, drawing on the **GetDeploying** index of 76 providers and 98 GPU models, found on-demand GPU rental rates rose only 1.7 percent over a recent 12-month period even as AI API prices fell sharply elsewhere in the stack, and separately reports that Google Cloud's spot and preemptible discounts of 60 to 91 percent off on-demand are the steepest among the hyperscalers it tracked (www.cloudzero.com) (www.cloudzero.com). The same analysis separately confirms CoreWeave's committed-usage discounts of up to 60 percent off on-demand (www.cloudzero.com). For availability, buyers should treat "on-demand" as a spectrum rather than a binary: AWS and Azure sell capacity reservation products for guaranteed access to scarce H100 supply, Crusoe requires a sales conversation for its newest GB200 and B200 accelerators rather than offering self-serve on-demand pricing (as detailed above), and Vast.ai's marketplace model means availability at any given price point fluctuates with which independent hosts currently have idle capacity listed (vast.ai).

For AI teams inside regulated industries such as life sciences, IntuitionLabs' advisory perspective is that the GPU rental decision should follow, not precede, a data-governance and validation plan, since moving training data across regions or providers to chase a lower headline rate can create compliance overhead that exceeds the compute savings; this is a "how to choose" consideration this report surfaces rather than a specific product recommendation, consistent with IntuitionLabs' role as an advisor rather than an infrastructure vendor (intuitionlabs.ai).

Data Analysis and Evidence

Quantifying "how much cheaper" the GPU-cloud and marketplace tiers are than the hyperscalers, and what independent performance data underlies the pricing gap between H100 and A100 hardware, requires triangulating vendor specifications, market-research estimates, and third-party price trackers, since no single source publishes a comprehensive cross-provider dataset.

Hardware specifications set the ceiling on what a given price buys. NVIDIA's own datasheet states the H100 SXM5 delivers roughly 3 terabytes per second of memory bandwidth per GPU (www.nvidia.com) and up to 3,958 teraFLOPS of FP8 Tensor Core throughput on the SXM5 variant (www.nvidia.com); a dual-GPU H100 NVL configuration links two GPUs via NVLink to reach 188GB of combined HBM3 for large-model inference (www.nvidia.com). By comparison, NVIDIA's A100 datasheet lists over 2 terabytes per second of memory bandwidth (www.nvidia.com), with the 80GB SXM variant specifically rated at 2,039 gigabytes per second (www.nvidia.com). NVIDIA has reported that its H100 submission to the independent **MLCommons MLPerf Training v3.0** benchmark round delivered up to 3.1 times more performance per accelerator than its prior A100 submission (web.archive.org), a vendor-reported figure validated by MLCommons' independent submission and audit process rather than an internal-only claim; MLCommons' own methodology notes that its "Available" category is defined specifically as systems "available for purchase or for rent in the cloud," making the benchmark suite explicitly usable for cloud price-performance comparisons rather than only for evaluating owned hardware

(mlcommons.org). Trade coverage of the September 2025 MLPerf Inference v5.1 round noted the benchmark suite had by then accumulated 90,000 cumulative submitted results, evidence of the scale of independent, reproducible cross-vendor testing the suite now represents (www.hpcwire.com).

On market sizing, three named research firms publish materially different **GPU as a Service** forecasts, and this report presents the spread rather than picking a winner: Grand View Research projects the market reaching $14.46 billion by 2033 at a 16.0 percent compound annual growth rate (www.grandviewresearch.com); Precedence Research values the market at $4.96 billion in 2025 (www.precedenceresearch.com); and Fortune Business Insights forecasts growth from $8.66 billion in 2026 to $162.54 billion by 2034, a 44.3 percent compound annual growth rate that implies far steeper adoption than either of the other two estimates (www.fortunebusinessinsights.com). The wide spread across three named, methodologically distinct research firms is itself a data point for readers: it indicates the GPU-as-a-service category is still young enough, and its boundaries defined loosely enough (whether it includes bundled MLOps services, for instance), that market-size figures in this space should be treated as directional rather than precise.

On pricing dynamics specifically, CloudZero's synthesis of the GetDeploying tracker's 76-provider, 98-model dataset found on-demand GPU rental rates essentially flat (up only 1.7 percent) over a recent 12-month window (as cited above), which appears to sit in tension with SemiAnalysis's finding of a nearly 40 percent rise in H100 one-year contract prices over a similar period (newsletter.semianalysis.com). The discrepancy is best explained by measurement scope rather than a genuine contradiction: SemiAnalysis's index tracks a narrower one-year contract structure specifically for H100, where committed-capacity scarcity has been most acute, while the GetDeploying-based figure spans a much broader basket of 98 GPU models and on-demand (not contract) pricing across the wider market, including older and less contested GPU generations. Readers should treat both figures as accurate within their own defined scope, not as competing measurements of the same thing.

Case Studies and Real-World Examples

(Hypothetical Example) Consider an AI team planning a 7-day fine-tuning run on a single H100 GPU, expecting to retain 200GB of checkpoint and dataset storage for the duration and to transfer approximately 500GB out of the cloud environment to an on-premises system once the run completes. Applying the effective-hourly-rate method from the prior section to four representative providers illustrates how the ranking changes once storage and egress are included, rather than comparing headline GPU-hour prices alone.

At Lambda's $3.99 per hour on-demand SXM rate, the 168-hour run costs $670.32 in raw compute; storage and egress on Lambda's filesystem carry no separate ingress or egress charge, so the effective total is unchanged at $670.32 (lambda.ai) (docs.lambda.ai). At RunPod's $3.29 per hour H100 SXM rate (cited above), raw compute costs $552.72, plus a modest storage charge of roughly $0.05 per gigabyte-month prorated for the week, since RunPod's egress is not itemized separately on its pricing page and is therefore excluded from this estimate rather than assumed. At AWS's $6.88 per hour single-GPU p5.4xlarge rate, raw compute costs $1,155.84, and the 500GB egress at $0.09 per gigabyte (after the 100GB free allowance) adds $36.00, bringing the effective total to roughly $1,191.84, nearly double the Lambda figure for the same workload once the higher headline rate is combined with a non-zero egress fee (pricing.us-east-1.amazonaws.com) (pricing.us-east-1.amazonaws.com). At Vast.ai's median H100 SXM marketplace rate of $2.00 per hour, raw compute costs $336.00, the lowest of the four, though the vendor's own documentation cautions that the exact figure will vary by which host is selected and that storage and bandwidth terms are set per-machine rather than platform-wide (vast.ai) (docs.vast.ai).

This worked example is illustrative only, built from list prices as of September 2026 and simplified assumptions about storage duration and data movement; it is not a benchmark of actual training throughput, and a real fine-tuning run's total cost will also depend on GPU utilization, checkpoint frequency, and whether the job completes without interruption. Readers should substitute their own workload's storage and egress profile into the same formula rather than treating the dollar figures above as universal.

Implications and Future Directions

Three forces will likely keep cloud GPU pricing volatile through the remainder of 2026 and into 2027. First, component-level supply: with the major HBM memory manufacturers reported effectively sold out for 2026, any near-term easing in GPU rental prices is more likely to come from demand plateauing than from a supply surge (www.cnbc.com). Second, capital deployment: IDC's raised forecast of $497 billion in 2026 AI infrastructure spending, roughly 56 percent above the prior year, signals that hyperscalers and neoclouds alike are still in a capacity-building phase rather than a mature, price-competitive equilibrium (www.idc.com). Third, newest-generation hardware remains gated behind sales conversations rather than self-serve pricing even at providers that publish fixed rates for older generations, as Crusoe Cloud's "contact sales" listings for its GB200 NVL72 and B200 accelerators illustrate (www.crusoe.ai); buyers evaluating a migration to next-generation silicon should expect availability, not price, to be the binding constraint for at least the next several quarters.

For buyers, the practical implication is that the "best" provider depends on workload shape more than on any single price list. Short, bursty, cost-sensitive jobs are well served by marketplace pricing such as Vast.ai's, where the lowest headline rates are available but reliability and interconnect quality are less standardized, as detailed above. Large, multi-node distributed training benefits from dedicated InfiniBand fabrics such as Azure's 400 Gbps per-GPU connection or Lambda's InfiniBand-connected clusters (both detailed above), where the higher headline rate buys guaranteed network topology rather than best-effort bandwidth. Enterprises with existing hyperscaler commitments and predictable, sustained usage are likely to find the largest savings not in switching providers but in converting on-demand hyperscaler spend to reserved or committed-use pricing, given the roughly 55 to 60 percent discounts observed on AWS's 3-year reserved p5 pricing (detailed in the Hyperscaler section above) and Google Cloud's 3-year committed use discount on A3 instances (cloud.google.com).

For regulated industries such as life sciences, where data residency and validation requirements often constrain which region or provider is even eligible for a workload, IntuitionLabs' advisory view is that the effective-rate method described above should be applied only after governance constraints have narrowed the eligible provider set, since the cheapest theoretical rate is irrelevant if the provider or region cannot meet a compliance requirement (intuitionlabs.ai).

Conclusion

Cloud GPU rental prices in September 2026 should be compared only after normalizing the unit and stating the configuration, region, commitment term, and service tier. Azure's ND H100 v5 series starts with a VM containing eight NVIDIA H100 Tensor Core GPUs. Hyperscalers charge a premium for enterprise integration, committed-use discounting, and dedicated interconnect fabrics; GPU-cloud specialists price closer to the underlying hardware cost while still offering dedicated networking for multi-node training; and marketplace platforms undercut both by trading price for standardization. The effective-hourly-rate method presented in this report, list price plus prorated storage

plus egress adjusted for uptime, gives buyers a reproducible way to compare across these structurally different pricing models rather than relying on headline rates that are not directly comparable. Given that independent trackers disagree on whether rental prices are rising sharply (SemiAnalysis) or holding roughly flat across the broader market (CloudZero via GetDeploying), and that named research firms disagree by an order of magnitude on the GPU-as-a-service market's future size, buyers should treat every figure in this space as dated and provider-specific, re-verify prices against the cited vendor pages before committing budget, and weight commitment-term and egress terms as heavily as the sticker price when comparing providers.

Frequently Asked Questions (FAQs)

What is the cheapest way to rent a cloud GPU? As of September 2026, the lowest observed H100 prices come from Vast.ai's auction marketplace, from $1.33 per hour with a $2.00 per hour median (as detailed above), though prices vary by host and are less standardized than a single-operator cloud's rate card. Among fixed-rate providers, Lambda's $3.99 per hour on-demand H100 SXM rate is among the lowest published, as detailed in the GPU Cloud Specialists section above.

How does GPU cloud pricing compare across providers in 2026? On-demand 8-GPU H100 nodes range from roughly $49 per hour at GPU-cloud specialists to $98.32 per hour on Azure (prices.azure.com), while single-GPU on-demand rates at specialist providers range from roughly $3.29 to $3.99 per hour, as detailed in the Hyperscaler and GPU Cloud Specialists sections above.

How do I calculate an effective hourly GPU rate? Add the prorated hourly storage cost and expected hourly egress cost to the listed GPU-hour price, then divide by the expected uptime fraction for interruptible or spot capacity, as detailed in the Comparative Context section above.

How does GPU availability differ by provider? Hyperscalers offer capacity-reservation products for guaranteed access; Crusoe Cloud requires a sales conversation for its newest GB200 and B200 accelerators rather than self-serve pricing (as detailed above); and Vast.ai's marketplace availability fluctuates with which independent hosts currently list idle capacity, per the Comparative Context section above.

What does GPU network egress typically cost? Hyperscaler egress runs $0.087 to $0.12 per gigabyte after a free monthly allowance (detailed in the Hyperscaler section above), while CoreWeave, Crusoe Cloud, and Paperspace advertise free or unlimited egress on their own pricing pages (www.paperspace.com).

Should I rent on-demand or spot/interruptible GPU capacity for AI training? Spot and preemptible instances save 60 to 91 percent versus on-demand on some hyperscalers, per CloudZero's analysis, but require a workload tolerant of interruption; reserved or committed-use contracts (up to 57 to 60 percent off on AWS and GPU-cloud specialists, as detailed above) are generally the better fit for predictable, sustained workloads (www.cloudzero.com).

Which provider is best for large-scale LLM training? Providers with dedicated, high-bandwidth interconnect, Azure's 400 Gbps per-GPU InfiniBand or Lambda's InfiniBand-connected 1-Click Clusters (both detailed in the Hyperscaler and GPU Cloud Specialists sections above), are best suited to multi-node distributed training where cross-node bandwidth, not just per-GPU compute, determines throughput.

How do H100 and A100 rental prices compare? Lambda offers H100 and A100 instances; compare prices only among matched GPU counts, memory capacities, regions, and commitment terms.
