



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

H100 Rental Prices Compared: \$1.49-\$6.98/hr Across 15+ Cloud Providers (2026)

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2025-10-24 · nvidia h100 · gpu rental · cloud gpu pricing · ai hardware · gpu-as-a-service · aws h100 · machine learning cost

Original article: <https://intuitionlabs.ai/articles/h100-rental-prices-cloud-comparison>



In this article

1. [H100 Rental Prices: AWS Capacity Blocks and Runpod Pods Compared \(2026\)](#)
 2. [Executive Summary](#)
 3. [Introduction and Background](#)
 4. [Product and Pricing Context](#)
 5. [Market Projections](#)
 6. [Pricing by Provider](#)
 7. [Detailed Analysis by Vendor](#)
 8. [Spot and Term Discounts](#)
 9. [Comparison Method](#)
 10. [Case Studies and Examples](#)
 11. [Economic and Adoption Implications](#)
 12. [Technological Currency and Inventory](#)
 13. [Future GPUs \(H200 and beyond\)](#)
 14. [Broader Impact](#)
 15. [Conclusion](#)
-

H100 Rental Prices: AWS Capacity Blocks and Runpod Pods Compared (2026)

Executive Summary

H100 pricing cannot be expressed as one universal on-demand rate: providers package H100 GPUs in different node sizes, regions, capacity products, and service tiers. On its current Capacity Blocks page, AWS lists P5.48xlarge Capacity Blocks at \$47.757 per hour (\$5.970 per H100) in every listed region, including US East (N. Virginia), US East (Ohio), and US West (Oregon). AWS Capacity Blocks are prepaid capacity products, not on-demand EC2 pricing; AWS's on-demand rate for the same p5.48xlarge is \$55.04 per hour (\$6.88 per H100) in US East (N. Virginia). Comparisons below identify the product and commitment basis rather than treating prices as interchangeable.

Update (October 2026): First-party price lists checked on October 10, 2026 show H100 list prices rising at several providers, reversing the 2025 declines described later in this article. AWS's Capacity Blocks rate for p5.48xlarge rose from \$5.191 to \$5.970 per H100-hour between this article's August 2026 check and October. Runpod's Secure Cloud H100 SXM Pod rose from \$3.29 to \$3.99 per hour over the same period. Lambda's 8-GPU H100 SXM instance now lists at \$3.99 per GPU-hour, against \$2.99 in the 2025 snapshot cited below, and Nebius raised its on-demand HGX H100 rate from \$3.85 to \$4.50 per GPU-hour effective October 1, 2026. Across the providers covered here, published on-demand list prices per H100-hour ranged from \$1.99 (Runpod Community Cloud, H100 PCIe) to \$12.29 (Azure ND96isr H100 v5, per GPU, East US); marketplace and spot offers can be lower. Prices change often, so confirm them on the provider's page before budgeting.

Key findings: The cheapest current H100 rental rates (per GPU-hour) by vendor include:

- **AWS EC2 Capacity Blocks (P5.48xlarge):** \$5.970 per H100-hour (\$47.757 per instance-hour) in every listed region, including US East (N. Virginia), US East (Ohio), and US West (Oregon); this is prepaid capacity, not on-demand EC2 pricing ([AWS Capacity Blocks pricing](#)).
- **AWS EC2 On-Demand (p5.48xlarge):** \$55.04 per instance-hour, or \$6.88 per H100-hour, on Linux in US East (N. Virginia), US East (Ohio), and US West (Oregon); \$68.80 (\$8.60 per H100) in US West (N. California) ([AWS EC2 On-Demand pricing](#)).
- **Google Cloud (a3-highgpu-8g):** \$88.49 per eight-H100 VM-hour on demand, or about \$11.06 per H100-hour, in Iowa (us-central1); Spot, DWS, and committed-use prices on the same page are lower ([Google Cloud pricing](#)).
- **Microsoft Azure (NC H100 v5 and ND H100 v5):** \$6.98 per hour for the single-GPU NC40ads H100 v5 and \$98.32 per hour (about \$12.29 per H100) for the eight-GPU ND96isr H100 v5, Linux pay-as-you-go in East US ([Azure VM pricing](#)).
- **Oracle Cloud (BM.GPU.H100.8):** \$10.00 per H100-hour pay-as-you-go on Oracle's USD price list, or \$80.00 per hour for the eight-GPU shape ([Oracle Cloud price list](#)).
- **Runpod Pods:** Secure Cloud rates of \$2.89 per hour for H100 PCIe, \$3.99 for H100 SXM, and \$3.19 for H100 NVL; Community Cloud rates of \$1.99, \$2.69, and \$2.59 for the same SKUs ([Runpod pricing](#)). These are SKU-specific Pod rates, not a market-wide benchmark.
- **Lambda, CoreWeave, and Nebius:** \$3.99 per GPU-hour for an 8×H100 SXM Lambda instance ([Lambda pricing](#)); \$49.24 per hour, or \$6.16 per GPU, for a CoreWeave 8×H100 HGX node ([CoreWeave pricing](#)); and \$4.50 per GPU-hour for Nebius HGX H100 from October 1, 2026 ([Nebius pricing](#)). Other vendors are omitted from the current-price summary when a matching public first-party hourly rate was not verified.

Prices vary materially by region, capacity product, operating system, service tier, and commitment. Historical tracker snapshots and promotional rates are not current-price evidence. AWS Capacity Blocks are prepaid reservations, while Runpod's displayed Pod rates are SKU-specific; neither should be used to infer a market-wide on-demand range.

Despite the 2025 price cuts, H100 rent remains a premium compared to older GPUs (A100s are now sub-\$1/GPU-hr open-market (www.thundercompute.com)). The sustained price reductions reflect oversupply and competition, as well as long-term commitments and spot markets putting downward pressure (www.thundercompute.com) (go4hosting.in). We examine the historical context, current pricing data, multi-provider comparisons (with extensive citations), market dynamics, and projected trends below.

Introduction and Background

The **NVIDIA H100 Tensor Core GPU** is a Hopper-architecture accelerator offered in 80 GB configurations and used for large-language-model training, HPC, and cloud AI workloads. It is no longer NVIDIA's newest data-center architecture: NVIDIA positions B200 systems as Blackwell-based products. Released in 2022–2023, the

H100 offered **unprecedented performance** (e.g. multi-Exaflop on DGX systems) and swiftly became the gold standard for cutting-edge AI research (go4hosting.in). However, its *sticker price* at retail (~\$25–40K/GPU (cyfuture.cloud)) makes ownership difficult for many, driving demand for rental solutions.

In parallel, the **GPU-accelerated cloud computing market** has **exploded** in recent years. The global “GPU-as-a-Service” market was only \$3.34 billion in 2023 but is projected to reach \$33.9 billion by 2032 (www.thundercompute.com). This growth is driven by **AI/ML needs**, making high-end GPUs like the H100 in continuously tight supply. Industry analysts note that “*GPU prices remain sky-high in 2025*” and that renting is often the only viable option for smaller players (cyfuture.cloud). As adoption grows, vendors have been incentivized to lower prices: Thunder Compute reports H100 rental rates **falling from \$8/hr to \$2.85–3.50/hr** in 2025 (www.thundercompute.com).

Rental models: Cloud providers typically charge by the hour (or minute) for GPU instances. Prices depend on *instance type*, region, and commitment. On-demand (no-commitment) is highest; spot/preemptible instances can be 60–90% cheaper (cloud.google.com) (go4hosting.in); and 1–3 year commitments (Reserved/Savings Plans) offer up to ~45–50% further discounts (aws.amazon.com) (go4hosting.in). In practice, many users mix and match: e.g. burst on-demand or spot when needed, while using reserved instances for steady-state workloads. The market spans major hyperscalers (AWS, Google, Azure, etc.), specialist AI/cloud vendors (Lambda Labs, CoreWeave, Runpod, etc.), and even AI-focused marketplaces (Vast.ai, etc.).

This report compares a limited set of publicly displayed H100 prices, last checked on October 10, 2026. It does not cover every vendor or establish a market-wide price range. Several providers require a calculator, account access, or a sales quote for current pricing. The comparison uses current first-party pricing only where a matching hourly product is publicly displayed; historical material is context only, not current-price evidence.

Product and Pricing Context

H100 Release and Early Pricing

When NVIDIA announced the H100 in March 2022 (nvidianews.nvidia.com) and it entered general availability in 2023 (aws.amazon.com), it was the fastest GPU for AI training. At launch, public on-demand rental pricing was extremely high. For instance, in mid-2023 AWS’s P5 instances (8×H100) were often listed above \$60/hr (~\$7.50/GPU-hr) (www.thundercompute.com) and Google’s A3 instances around \$88/hr (~\$11/GPU-hr) (www.thundercompute.com). These figures match NVIDIA’s historical pattern of high launch prices, as seen with prior A100/A6000 tiers. Academia & industry narratives at the time noted that while H100 power justified its cost-performance, *cost remained a major barrier* for widespread use.

However, by 2025 the supply of H100s increased substantially (due to added production, alternate chips, and enterprise-focused deals). Cloud providers have also advanced specialized hardware (e.g. AWS Trainium, Google TPUs) which competes with H100. To maintain competitiveness, many began cutting prices. A notable turning point occurred in June 2025: AWS announced a **~44% price reduction** on P5 instances (H100) across regions (aws.amazon.com) (www.datacenterdynamics.com). This brought AWS H100 GPU rental to roughly half its former rate. A DataCenterDynamics report confirms: “*On-demand price ... 44% for the H100 instance*”

(www.datacenterdynamics.com). Azure did not follow with list-price cuts: its retail price data shows the current NC40ads H100 v5 and ND96isr H100 v5 rates in effect since March 2024 and December 2023 respectively ([Azure VM pricing](#)).

Beyond hyperscalers, many smaller GPU-cloud vendors instituted competition-driven price wars. Throughout 2025, daily tracking sites (e.g. GPUCompare) documented multiple providers slashing H100 rates (see “Temporal Trending” under Comparison Method below). By late 2025, aggregated analyses expect H100 rental to settle around **\$2–4/GPU-hr** in most markets (www.thundercompute.com) (go4hosting.in). For perspective, ThunderCompute’s September 2025 snapshot still showed Azure H100 at \$6.98 and GCP at \$11.06 (www.thundercompute.com), but as of fall 2025 current rates are much lower (go4hosting.in).

A key market consideration is **price volatility and regional variation**. Indicators include GPUCompare’s mid-2025 reports of providers continuously lowering H100 prices by up to 20–25% in single moves (gpucompare.com) (gpucompare.com). Reported insights note “*continued price reductions across board*” (gpucompare.com) and H100 reaching “*new lows*” (gpucompare.com). The H100 price volatility surpasses that of previous-gen GPUs, reflecting fierce competition and excess capacity. Regionally, even within the same cloud provider, pricing differs; for example, the U.S. West Coast regions are consistently ~10–30% above East US rates (www.linkedin.com). Users can often reduce costs by choosing slower regions or leveraging preemptible instances.

Market Projections

Looking forward, experts anticipate this downward trend to continue, albeit with fluctuations due to demand surges. The ThunderCompute analysis projects global GPU rental market growth from \$3.34B (2023) to \$33.91B by 2032 (www.thundercompute.com), which should force pricing down over time. The arrival of newer GPUs (NVIDIA H200, and Blackwell B200 and B300 systems that AWS, Lambda, and Nebius now list) may further depress H100 prices as customers upgrade. On the other hand, sustained AI hype could keep demand high, partially offsetting price drops. In 2025, the evidence pointed to H100 rentals easing toward a competitive equilibrium of roughly \$2–3/GPU-hr for on-demand usage (www.thundercompute.com) (go4hosting.in). First-party list prices checked in October 2026 sit above that range at most providers covered here (see the update at the top of this article).

Pricing by Provider

This article presents selected H100 pricing examples from large clouds and specialist providers. Values are shown only where a current first-party page provides an applicable hourly price. The entries use different purchase options, including prepaid capacity, on-demand VMs, and Pods, so they are not a like-for-like market comparison.

Provider	Instance / SKU	GPUs/Instance	Price (\$/GPU-hr)	Source
AWS	p5.48xlarge Capacity Block (8×H100)	8	\$5.970 in every listed region	AWS Capacity Blocks (effective reservation rate per accelerator; prepaid capacity)

Provider	Instance / SKU	GPUs/Instance	Price (\$/GPU-hr)	Source
AWS	p5.48xlarge On-Demand (8×H100)	8	\$6.88 (N. Virginia, Ohio, Oregon); \$8.60 (N. California)	AWS EC2 On-Demand pricing (Linux)
Google Cloud	a3-highgpu-8g (A3 High, 8×H100)	8	\$11.06 on demand (Iowa, us-central1)	Google Cloud pricing
Azure	NC40ads H100 v5	1	\$6.98 (East US, Linux)	Azure VM pricing (region-dependent)
Azure	ND96isr H100 v5 (8×H100)	8	\$12.29 (East US, Linux)	Azure VM pricing (region-dependent)
Oracle Cloud	BM.GPU.H100.8 (8×H100)	8	\$10.00 (pay-as-you-go, USD)	Oracle Cloud price list (currency and region dependent)

Note: The AWS Capacity Block entry is a prepaid reservation rate, not on-demand EC2 pricing. AWS currently lists a reservation-component effective rate of \$47.757 per P5.48xlarge hour (\$5.970 per H100) in every listed region, up from \$41.528 (\$5.191 per H100) in listed U.S. regions and \$37.76 (\$4.720 per H100) elsewhere at this article's August 2026 check. These rates exclude an operating-system charge when applicable; Linux is listed at \$0, while other supported operating systems have separate charges. The Google Cloud, Azure, and Oracle figures are pay-as-you-go list prices for the stated region; other regions and purchase models differ.

Vendor / Service	Instance / Config	GPUs	Price (\$/GPU-hr)	Source
Runpod Pods	H100 PCIe	1	\$2.89 (Secure Cloud); \$1.99 (Community Cloud)	Runpod pricing
Runpod Pods	H100 SXM	1	\$3.99 (Secure Cloud); \$2.69 (Community Cloud)	Runpod pricing
Runpod Pods	H100 NVL	1	\$3.19 (Secure Cloud); \$2.59 (Community Cloud)	Runpod pricing
Lambda	H100 SXM instance (8×H100)	8	\$3.99	Lambda pricing
Lambda	H100 SXM / H100 PCIe instance	1	\$4.29 / \$3.29	Lambda pricing
CoreWeave	HGX H100 node (8×H100)	8	\$6.16 (\$49.24 per node-hour)	CoreWeave pricing
Nebius	HGX H100	Per GPU	\$4.50 from October 1, 2026 (previously \$3.85)	Nebius pricing

Notes: The Runpod figures are displayed Pod rates for the named SKUs; Secure Cloud is the default view on Runpod's page. The Lambda, CoreWeave, and Nebius figures are on-demand list prices from each provider's pricing page. Other specialist-cloud, marketplace, promotional, and historical offers are omitted because they are not a verified, like-for-like current comparison.

These figures show that **many non-hyperscaler platforms are cheaper** per GPU-hour than the large clouds' on-demand list prices. For instance, Lambda's 8×H100 SXM rate (\$3.99 today; \$2.99 in a 2025 snapshot (www.thundercompute.com)) and Runpod's H100 PCIe (\$1.99 on Community Cloud today, the same rate cited in that snapshot (www.thundercompute.com)) are well below AWS's \$6.88 on-demand rate. (A Go4Hosting analysis puts AWS bulk-reservation prices near \$1.90–\$2.10/GPU-hr (go4hosting.in), which requires 1–3 year commitments; AWS Savings Plans rates were not rechecked for this update.) This limited table does not establish a current market-wide on-demand range or ranking: its entries use different regions, service tiers, and purchase models.

Detailed Analysis by Vendor

AWS (Amazon EC2)

AWS offers H100 under its **P5 instance family**. The flagship is p5.48xlarge (192 vCPU, 8×H100 80GB); AWS also sells a single-GPU p5.4xlarge. Prior to mid-2025, a Thunder Compute snapshot put AWS's on-demand price for p5.48xlarge at about \$60.54/hour total (www.thundercompute.com) (i.e. \$7.57/GPU-hr), one of the highest H100 rental rates. However, on June 5, 2025 AWS announced an “*up to 45% price reduction*” for GPU instances, specifically 44% off H100 (P5) on-demand (www.datacenterdynamics.com). AWS's on-demand price list now shows p5.48xlarge at \$55.04 per hour (\$6.88 per GPU-hour) on Linux in US East (N. Virginia), US East (Ohio), and US West (Oregon), and \$68.80 per hour in US West (N. California); the single-GPU p5.4xlarge is \$6.88 per hour in US East (N. Virginia) ([AWS EC2 On-Demand pricing](#)). Applying the 44% cut, the current US East rate implies a May 2025 baseline near \$98 per hour, well above the Thunder Compute figure. Some third-party trackers cite about **\$3.90 per GPU-hour** for P5 on demand (go4hosting.in), which does not match AWS's own list price.

For a prepaid AWS reference, the Capacity Blocks page lists effective *reservation-component* rates for P5 H100 capacity: \$5.970 per H100 (\$47.757 per p5.48xlarge hour) in every listed region, from US East (N. Virginia) to Europe (London) and South America (Sao Paulo), and \$5.970 per hour for a single-GPU p5.4xlarge block. These are prepaid Capacity Blocks reservation rates, not on-demand EC2 rates. AWS bills the operating system separately while instances run; Linux is listed at \$0, but other supported operating systems add a charge ([AWS Capacity Blocks pricing](#)).

AWS Spot and Reservation Pricing

AWS Spot and Savings Plans can have different pricing and interruption or commitment terms from Capacity Blocks. Obtain the current rate for the exact P5 instance, region, operating system, and purchase option from AWS before budgeting; this article does not establish a universal AWS spot or committed-use H100 rate.

Microsoft Azure

Azure offers H100-capable VM families. Azure's retail price list for East US shows the single-GPU NC40ads H100 v5 at \$6.98 per hour and the eight-GPU ND96isr H100 v5 at \$98.32 per hour (about \$12.29 per GPU-hour), both Linux pay-as-you-go; Spot prices on the same list were \$1.29 and \$18.17 per hour on October 10, 2026. Prices vary by region, so use Azure's pricing page or calculator with the intended region and VM configuration before making a comparison ([Azure VM pricing](#)).

Box-by-box, Azure's ND and NC families have historically been priced above AWS, reflecting differing market positioning. Users report ~\$7/GPU-hr as typical (www.thundercompute.com). Azure also sells 1-year and 3-year reservations: in East US, the listed 1-year reservation works out to about 25% (NC40ads H100 v5) to 36% (ND96isr H100 v5) below pay-as-you-go, and the 3-year reservation to about 45% to 56% below. Azure Spot VMs are available for both sizes at the rates given above.

Google Cloud Platform (GCP)

Google Cloud offers H100 GPUs in A3 VM configurations. Prices depend on configuration, region, and purchase option. For Iowa (us-central1), the accelerator-optimized pricing page lists a3-highgpu-8g (A3 High, 8×H100) at \$88.49 per hour on demand (about \$11.06 per GPU-hour), \$50.32 for Spot, \$38.32 for DWS Flex-start, \$41.60 for DWS Calendar mode, and \$61.38 and \$38.86 with 1-year and 3-year resource committed-use discounts; a3-megagpu-8g (A3 Mega) is \$93.40 per hour on demand. Do not compare a Google Cloud on-demand or Spot selection directly with AWS Capacity Blocks, marketplace offers, or other service tiers ([Google Cloud accelerator-optimized VM pricing](#); [Google Cloud Spot VM pricing](#)).

Oracle Cloud

Oracle offers H100 GPUs via **bare-metal instances** (the BM.GPU or BMH series). A standard **BM.GPU.H100.8** configuration packs 8×H100 GPUs into one machine. Oracle lists BM.GPU.H100.8 as an eight-H100 Hopper shape. Oracle's USD price list shows Compute GPU H100 at \$10.00 per GPU-hour pay-as-you-go, or \$80.00 per hour for the eight-GPU shape, before any negotiated discounts ([Oracle Cloud price list](#)).

Other large clouds are omitted: **Alibaba Cloud** had announced new GPU instances but published **no official H100 on-demand** figures as of late 2025. Similarly, **IBM Cloud** started offering H100 in late 2024 but pricing details are not widely published, so we omit them here.

Lambda, CoreWeave, and Paperspace

Lambda Labs, a popular AI cloud, rents H100 GPUs by the unit or in 8-GPU clusters. Lambda's 8×H100 "Lambda Cloud" instances listed at **\$2.99/GPU-hr** in 2025 (www.thundercompute.com) for H100 SXM, or \$23.92/hr for the 8-GPU instance. Lambda's pricing page now lists on-demand H100 SXM instances at \$3.99 per GPU-hour for 8 GPUs, \$4.09 for 4, \$4.19 for 2, and \$4.29 for 1, with a single H100 PCIe at \$3.29; its H100 1-Click Clusters (16 GPUs and up, 2 weeks to 1 year) list at \$5.54 to \$6.16 per GPU-hour ([Lambda pricing](#)). The 8-GPU rate is **well below AWS on-demand** (\$6.88).

CoreWeave also offers 8×H100 HGX nodes; their pricing is around \$49.24/hr per node (www.thundercompute.com), which normalizes to **\$6.16/GPU-hr**; CoreWeave's pricing page still lists that \$49.24 on-demand rate in October 2026 ([CoreWeave pricing](#)). CoreWeave thus is pricier than Lambda, reflecting its niche HPC focus and InfiniBand support.

Paperspace's "dedicated" H100 VM (1×80GB) is **\$5.95/GPU-hr** (www.thundercompute.com). This is a dedicated single GPU instance, competitive with Azure. Paperspace also provides discounts for multi-GPU bookings, but on-demand single price stays near \$6.

Runpod, Vast.ai, TensorDock, HPC-AI, and Cudo

Runpod publishes separate Pods prices by GPU SKU and service option. Its displayed Secure Cloud Pods rates are \$2.89 per hour for H100 PCIe, \$3.99 per hour for H100 SXM, and \$3.19 per hour for H100 NVL; Community Cloud rates are \$1.99, \$2.69, and \$2.59 for the same SKUs. Prices and availability vary by SKU and service, so use the provider page for the current selection ([Runpod pricing](#)).

Vast.ai is a GPU marketplace where independent hosts list idle GPUs. Vast's lowest H100 listing has been as low as **\$1.87/GPU-hr** ([www.thundercompute.com](#)) (used NVIDIA 80GB PCIe GPUs). In July 2025, Vast.ai ran promotions cutting SXM H100 to \$1.49 ([gpucompare.com](#)). Vast.ai prices bounce with supply; it often presents the *floor* for what's possible on the open market. A query of Vast.ai's marketplace on October 10, 2026 returned single-GPU H100 SXM offers from about \$1.68 per hour (from an unverified host) and H100 PCIe offers from about \$2.37 per hour ([Vast.ai pricing](#)).

HPC-AI (a smaller cloud) advertises H100 at \$1.99 ([gpucompare.com](#)). **TensorDock** (Europe/US cloud) similarly offers H100 at \$2.25 ([gpucompare.com](#)) (for clients like FloydHub). **Cudo Compute** lists H100 at \$2.45 in many clusters, and was reported to cut to **\$1.80/GPU-hr** in June 2025 ([gpucompare.com](#)). Other entrants like **Hyperstack** and **Nebius** had H100 pricing around \$2.00–2.49 after cuts in June 2025 ([gpucompare.com](#)). Nebius's own price list now shows HGX H100 at \$4.50 per GPU-hour on demand from October 1, 2026 (previously \$3.85), with preemptible capacity from \$0.79 ([Nebius pricing](#)).

Specialist-cloud, marketplace, spot, and committed-use offerings may have materially different prices and service characteristics. This limited comparison does not establish a market-wide discount or ranking. Listings below \$2 per GPU-hour exist (Runpod Community Cloud H100 PCIe at \$1.99 and some Vast.ai marketplace offers), but they come from community or marketplace hosts with different service characteristics; compare only like-for-like region, SKU, service tier, and purchase terms.

Spot and Term Discounts

Spot and committed-use products have different availability, interruption, and commitment terms from on-demand and prepaid-capacity products. Check the provider's current pricing page for the exact configuration, region, and terms before comparing options ([Google Cloud Spot VM pricing](#)).

Comparison Method

The comparison uses only the published product pages for the entries in the table:

- **Normalized Comparisons:** When comparing per-GPU costs, larger instances distort price-per-GPU. Thus, most analyses (e.g. Thunder Compute) *normalize* to \$/GPU-hr. Table 1 and 2 use this metric.
- **Temporal Trending:** GPUCompare's daily updates document **month-to-month declines**. For example, in June–July 2025 multiple providers slashed H100 pricing by ~20–25% in short order ([gpucompare.com](#)) ([gpucompare.com](#)). By September, ThunderCompute notes general H100 availability around \$2.85–3.50 ([www.thundercompute.com](#)). A Go4Hosting analysis reported AWS/Azure at \$3.9/\$3.0 in 2025 ([go4hosting.in](#)), down from \$7.57 (AWS) and \$6.98 (Azure) in the Thunder Compute snapshot ([www.thundercompute.com](#)); AWS's and Azure's own list prices in October 2026 (\$6.88 and

\$6.98 per GPU-hour on demand) do not reflect those lower figures. This chronological perspective (Jun 2025: cuts ([gpucompare.com](https://www.gpucompare.com)) → Sep 2025: ~\$3–4 (www.thundercompute.com) → Nov 2025: similar) shows sustained downward pressure through 2025; first-party list prices checked in October 2026 show increases at several specialist providers.

- **Regional Variance:** Anecdotal tracking indicates AWS West Coast regions ~15% higher than East (www.linkedin.com). These differences arise from data center costs and local supply. Users can often save by choosing an east/midwest zone. Azure also has up to ~30% spread between cheapest and priciest region (cloudprice.net).
- **Cost vs. Purchase:** For context, the **cost to buy** an H100 is cited at \$25–40K (cyfuture.cloud). Renting at \$3/GPU-hr full-time (24/7) costs ~\$2,160/month, which would “pay off” the GPU in ~1–2 years. But owning incurs capital, power, facility costs (estimated \$1000–2000/month for electricity alone (www.cudocompute.com)). The rental model avoids these and adds maintenance support, which explains why enterprises often prefer cloud rental for short-to-medium-term workloads.
- **Cost-Performance:** Analysts emphasize cost-per-computation, not just sticker price (www.nextplatform.com). For instance, AWS’s price/performance improved over prior generations despite higher absolute cost (a theme in The NextPlatform’s GPU pricing articles (www.nextplatform.com)). Our focus here, however, is raw rental price; detailed cost-efficiency (FLOPS per dollar) is beyond this report’s scope.
- **Citations and Method:** Table values are drawn from official pricing pages or trusted compilations: e.g. AWS/Azure public calculators, Thunder Compute blog (Sep 2025 data) (www.thundercompute.com), Go4Hosting analysis (go4hosting.in) (go4hosting.in), GPUCompare daily logs (gpucompare.com) (gpucompare.com), and vendor docs. Where providers only list multi-GPU rates, we divided by GPU count (noting that licensed per-instance often prices by node). We avoided any hidden fees or long-term commitments in these figures.

Case Studies and Examples

Startup GPU procurement: A CloudCombinator analysis of GenAI startups highlights the complexities of acquiring H100 capacity on AWS (cloudcombinator.ai) (cloudcombinator.ai). It notes that “popular GPU types (H100, P6-B200) may be out of stock” in busy regions (cloudcombinator.ai), forcing planners to use alternatives like GPU Capacity Reservations or AWS UltraClusters for guaranteed access (cloudcombinator.ai). This underscores that price alone isn’t enough; availability is also a factor. The blog recommends strategies combining spot buys, regional flexibility, and multi-cloud to manage costs, echoing our observation of price variation.

Cloud vs. On-Premises TCO: A Cudo Compute blog breaks down the total cost of owning an 8-H100 server. They estimate each H100 costs ~\$30,971 part-price (www.cudocompute.com), so an 8-card system ~\$247,766 plus CPU (\$25K) and extras (www.cudocompute.com). They estimate a roughly \$325,000–\$425,000+ on-premises configuration ([Cudo Compute](https://www.cudocompute.com)). In Cudo’s illustrated configuration, \$19.60 per hour is the GPU component only; adding the listed CPU, memory, and storage brings the total to \$22.68 per hour. At continuous use for a 30-day month, that total is \$16,329.60 ($22.68 \times 24 \times 30$). Relative to the cited \$325,000–

\$425,000+ estimate, the simple hardware-cost crossover is about 20–26 months before electricity, cooling, maintenance, financing, utilization, and configuration differences. It is an illustrative comparison, not a purchase recommendation. For many, annual or monthly GPU rental provides budget flexibility and eliminates maintenance overhead.

Regulatory procurement: Government and academic HPC centers often publish utilization cases. While specific H100 rental deals are proprietary, general budget reports (e.g. Argonne/Lawrence projects) note that hardware budgets hammered by rising chip prices make cloud an attractive option. For example, The Next Platform reported that the price of Argonne National Laboratory's Aurora supercomputer was lowered by a \$300 million write-off that Intel took on the system (www.nextplatform.com). This contrasts with cloud's retail rates and hints why high-demand research might qualify for subsidized purchases instead.

Economic and Adoption Implications

The steep fall in H100 rental costs through 2025 had several implications. For enterprises and researchers, AI training became more accessible: fine-tuning large models or running vision pipelines on H100 became far cheaper than in 2023. This democratization fuels innovation but also squeezes margins for small GPU vendors. Market analysts observe “**GPU rental market exploded**” demand but concurrently pressure to “drive down costs” (www.thundercompute.com). In effect, the anticipated GPU shortage of 2023–24 eased somewhat, as vendors responded by lowering prices and increasing supply chains (purchase from multiple chip fabs, interconnect tech improvements, etc).

For cloud providers, these price cuts reflect economies of scale and competitive necessity. AWS's 44% P5 cut may partly reflect customer pushback over scarcity pricing. Google and Microsoft risk losing GPU-hungry customers to smaller clouds, which offer more attractive hourly rates (albeit usually with less enterprise support). Conversely, providers expect to retain workloads by offering discounts for committed usage (1–3 years), and third-party analyses cite savings-plan rates near ~\$1.90/GPU-hr (go4hosting.in). Potentially, if H100 prices drop further, providers may shift emphasis to newer chips (like H200) as the premium offering.

Technological Currency and Inventory

Another aspect is inventory management. Some smaller clouds (e.g. Lambda Labs, Runpod) offering H100 at \$2–4/GPU-hr risk hardware resale losses if H100 market prices fall below purchase cost. The Next Platform calculated that the FP32 price/performance of AWS's P-class GPU instances improved 16.8 times from the P2 (2016) to the P5 (2023) (www.nextplatform.com). If H100 commodity price falls, companies may offload old inventory cheap. Conversely, manufacturer reactions (e.g. Nvidia's own pricing guidance) could support higher prices via limiting supply of discounts. But as of Nov 2025, H100 chips were no longer “limited edition”, and refurbished boards had appeared.

Future GPUs (H200 and beyond)

H200 is a memory-enhanced **Hopper** GPU, not a Blackwell product. NVIDIA said H200 systems would be available from global system manufacturers and cloud service providers beginning in the second quarter of 2024. NVIDIA introduced the Blackwell platform separately in March 2024; B200 is a Blackwell GPU. These

product releases do not establish a universal future H100 rental price, so this report does not forecast a specific H100 rate. By October 2026, Blackwell capacity is publicly priced: Lambda lists on-demand B200 instances from \$6.69 per GPU-hour ([Lambda pricing](#)), and Nebius lists HGX B200 at \$8.50 and HGX B300 at \$9.50 per GPU-hour from October 1, 2026 ([Nebius pricing](#)).

Broader Impact

Falling rent may hasten AI adoption in SMBs and academia. For startups and labs previously deterred by cost, H100 compute is reaching commodity price points. This could contribute to an AI boom or democratize research. However, some caution is warranted: extremely low prices often come with caveats (limited slots, older GPUs sold as H100, or multi-tenancy on slower networking). Additionally, as H100 rent flattens, differences between providers may hinge on value-adds (support, specialized interconnects, compliance). It's also possible that in later 2026, if GPU supply outstrips demand, providers will deliberately throttle lower-end rates to protect profit margins, e.g. by promoting multi-year commitments or preemptible offerings over on-demand.

Conclusion

In summary, this article provides selected H100 pricing examples, not a definitive market-wide snapshot. Current comparisons must identify the provider, product, region, service tier, and purchase model; listed rates such as Google Cloud on-demand A3 High, AWS Capacity Blocks, and Runpod Secure Cloud are not interchangeable.

Key drivers of the 2025 declines were broad market trends (surging supply, price wars, and alternative chips), all documented here with citations ([www.datacenterdynamics.com](#)) ([gpucompare.com](#)) ([www.thundercompute.com](#)) ([go4hosting.in](#)). Users planning ML workloads should obtain a current quote or pricing-tool result for the exact GPU SKU, region, capacity type, and commitment term. The selected examples in this article do not support a universal H100 rate or a predicted spot or reserved rate.

As of October 10, 2026, published on-demand list prices per H100-hour in this article ranged from \$1.99 (Runpod Community Cloud, H100 PCIe) to \$12.29 (Azure ND96isr H100 v5), with AWS at \$6.88 on demand and \$5.970 through Capacity Blocks, and several specialist providers above their 2025 levels. These are selected examples, not coverage of the full provider market.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://aws.amazon.com/ec2/capacityblocks/pricing/>
2. <https://aws.amazon.com/ec2/pricing/on-demand/>
3. <https://cloud.google.com/products/compute/pricing/accelerator-optimized>
4. <https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/>
5. <https://www.oracle.com/cloud/price-list/>
6. <https://www.runpod.io/pricing>

7. <https://lambda.ai/pricing>
8. <https://www.coreweave.com/pricing>
9. <https://nebius.com/prices>
10. <https://www.thundercompute.com/blog/ai-gpu-rental-market-trends>
11. <https://go4hosting.in/knowledgebase/hosting/cost-of-renting-an-nvidia-h100-on-aws-or-google-cloud>
12. <https://cyfuture.cloud/blog/how-to-find-the-best-gpu-deals-in-2025/>
13. <https://intuitionlabs.ai/articles/openai-stargate-datacenter-details>
14. <https://intuitionlabs.ai/articles/ai-in-healthcare-profiling-5-key-technology-companies>
15. <https://cloud.google.com/spot-vms/pricing>
16. <https://aws.amazon.com/blogs/aws/announcing-up-to-45-price-reduction-for-amazon-ec2-nvidia-gpu-accelerated-instances/>
17. <https://nvidianews.nvidia.com/news/nvidia-announces-hopper-architecture-the-next-generation-of-accelerated-computing>
18. <https://aws.amazon.com/blogs/aws/new-amazon-ec2-p5-instances-powered-by-nvidia-h100-tensor-core-gpus-for-accelerating-generative-ai-and-hpc-applications/>
19. <https://www.thundercompute.com/blog/nvidia-h100-pricing>
20. <https://www.datacenterdynamics.com/en/news/aws-cuts-costs-for-h100-h200-and-a100-instances-by-up-to-45/>
21. <https://gpucompare.com/blog/26-06-2025-daily-gpu-cloud-pricing-update>
22. <https://gpucompare.com/blog/21-07-2025-daily-gpu-cloud-pricing-update>
23. <https://www.linkedin.com/pulse/gpu-rental-prices-vary-more-than-you-think-even-across-carmen-li-qiaze>
24. <https://vast.ai/pricing>
25. <https://gpucompare.com/blog/13-06-2025-daily-gpu-cloud-pricing-update>
26. https://cloudprice.net/vm/Standard_NC40ads_H100_v5
27. <https://www.cudocompute.com/blog/what-does-it-cost-to-rent-cloud-gpus>
28. <https://www.nextplatform.com/2023/07/27/h100-gpu-instance-pricing-on-aws-grin-and-bear-it/>
29. <https://cloudcombinator.ai/case-studies/blog-post-securing-affordable-nvidia-gpu-capacity-on-aws-for-startups?hsLang=en>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.