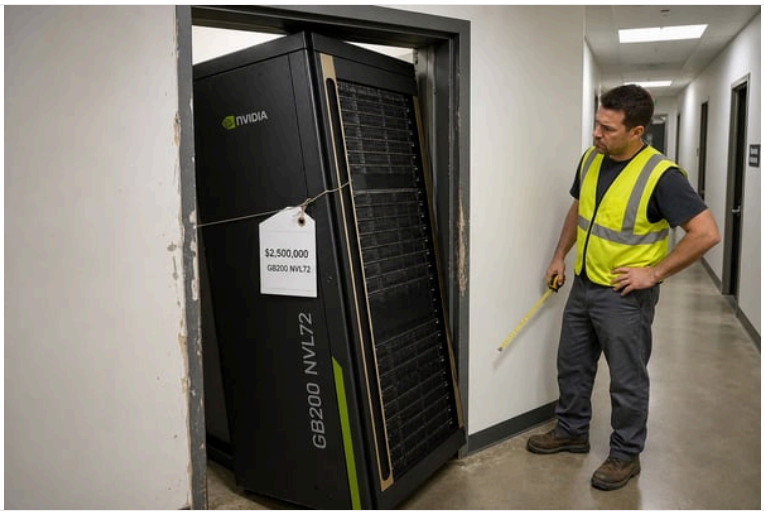
GPU Price Index 2026: H100, H200, B200, GB200 Street Prices

Published July 10, 2026 34 min read



Executive Summary

As of July 2026, the market for NVIDIA's data center graphics processing units (GPUs) is characterized by extreme price dispersion, tight memory supply, and a widening gap between list price and street price. The **NVIDIA H100** (Hopper architecture, 80GB HBM3) now rents for as little as $1.33 per GPU-hour on marketplace platforms and as much as $12.29 per GPU-hour on premium hyperscaler instances (Source: gpufinder.dev) (Source: emma.ms), a spread the market itself has taken to calling the "GPU price index." New H100 units sell for $25,000 to $40,000, while used cards trade for as little as $8,200 to $25,000 depending on age and channel (Source: cloudzero.com) (Source: craftrigs.com).

The **NVIDIA H200**, a memory-upgraded Hopper variant with 141GB of HBM3e, carries a modest premium over the H100 in most channels, ranging from $1.32 to roughly $13.78 per GPU-hour, with a market median near $3.82 per GPU-hour across tracked providers (Source: gpufinder.dev) (Source: gpus.io). The **NVIDIA B200** (Blackwell architecture, 192GB HBM3e) is priced from roughly $2.71 per GPU-hour on marketplaces up to $27.04 per GPU-hour on constrained hyperscaler allocations, while its one-time purchase price sits near $30,000 to $40,000 per GPU, with an estimated manufacturing cost of only about $6,400 (Source: spheron.network) (Source: siliconanalysts.com). Rack-scale **GB200 NVL72** systems, which bind 72 Blackwell GPUs and 36 Grace CPUs into a single NVLink domain, list at roughly $2 million to $3 million per rack, with cloud rental running $10.50 to $27 per GPU-hour (Source: aiappstacker.com). The newer **GB300 NVL72** (Blackwell Ultra), which NVIDIA says delivers 1.5 times the AI performance of GB200 NVL72, commands a further premium, with racks estimated at $3 million to $4 million (Source: nvidianews.nvidia.com) (Source: aiappstacker.com).

These prices sit against a backdrop of extraordinary demand: NVIDIA reported record Data Center revenue of $75.2 billion for its fiscal first quarter of 2027 (ended April 26, 2026), up 92% year over year, following $62.3 billion in Data Center revenue in the prior quarter (Source: nvidianews.nvidia.com) (Source: sec.gov) (Source: nvidianews.nvidia.com). Independent researchers put the global data center GPU market at $119.97 billion in 2025, growing to $228.04 billion by 2030 (Source: marketsandmarkets.com), though estimates vary widely by scope and methodology. A structural shortage of high-bandwidth memory (HBM), the specialized DRAM stacked onto every modern AI accelerator, is now the single biggest upward pressure on GPU prices: DRAM contract prices were expected to rise 50% to 55% quarter over quarter in early 2026, and analysts at Bernstein project HBM contract prices could climb 2 to 2.5 times further by 2027 (Source: cnbc.com) (Source: stockwirex.com).

Simultaneously, U.S. export controls have created a parallel, higher-priced market in China, where a restricted NVIDIA DGX B300 server now sells for more than 8 million yuan (roughly $1.1 million) on the black market, over double its price six months earlier and well above its U.S. retail value (Source: otontechnology.com). Researchers at Epoch AI estimate that between 290,000 and 1.6 million H100-equivalent GPUs were smuggled into China through the end of 2025 (Source: epoch.ai). This report synthesizes cloud on-demand rates, purchase and resale prices, manufacturing-cost estimates, and market-size data from more than 40 independent sources to answer, as precisely as the fragmented public record allows, what an H100, H200, B200, or GB200/GB300 actually costs in mid-2026, and why the answer keeps changing.

Introduction and Background

Pricing an NVIDIA data center GPU in 2026 is unlike pricing almost any other piece of enterprise hardware. There is no single manufacturer's suggested retail price, no standardized SKU, and no central exchange. Instead, the effective price of an **H100**, **H200**, **B200**, or rack-scale **GB200/GB300 NVL72** system depends on the provider, the region, the contract length, the form factor (PCIe versus SXM), and whether the buyer is renting by the hour, reserving capacity for a year, or purchasing hardware outright. A single generation of silicon can carry an eight-fold to twelve-fold spread in hourly rental price depending on where it is rented (Source: getdeploying.com) (Source: emma.ms).

This "GPU price index" matters because compute cost has become one of the largest line items in AI infrastructure budgets, alongside power and networking. NVIDIA's own financial results illustrate the scale of the market these prices are transacted in: the company reported record quarterly revenue of $81.6 billion for its fiscal first quarter of 2027, up 85% year over year, with Data Center revenue alone reaching $75.2 billion (Source: nvidianews.nvidia.com). For the prior quarter, Data Center revenue was $62.3 billion, and the Data Center segment now accounts for more than 91% of total company sales (Source: cnbc.com).

Four generations of NVIDIA silicon are simultaneously in active commercial deployment as of mid-2026. The **H100**, launched in 2022 on the Hopper architecture with 80GB of HBM3 memory and a dedicated Transformer Engine, remains the most widely available accelerator, tracked by pricing aggregators across 45-plus cloud providers (Source: getdeploying.com).The **H200**, a memory-upgraded Hopper variant released in late 2023 with 141GB of HBM3e (nearly double the H100's capacity) at 4.8 terabytes per second of bandwidth, targets memory-bound inference on large models (Source: nvidia.com). The **B200**, NVIDIA's first Blackwell-architecture GPU, packs 192GB of HBM3e, a second-generation Transformer Engine, and native FP4 arithmetic support, delivering up to 9,000 TFLOPS of dense FP4 compute per card (Source: spheron.network). And the rack-scale **GB200 NVL72** and its successor, the **GB300 NVL72** (Blackwell Ultra), bind dozens of GPUs and Grace CPUs into single coherent NVLink domains for trillion-parameter model training and inference (Source: nvidia.com).

This report, anchored to data collected as of July 10, 2026, walks through cloud rental and purchase pricing for each of these four product families, compares that pricing across providers and against manufacturing-cost estimates, and examines the macro forces (HBM scarcity, export controls, neocloud financing) that are pushing prices in different directions depending on the channel a buyer uses.

Readers should keep two caveats in mind throughout. First, almost none of the pricing cited here comes from NVIDIA directly. Unlike consumer GPUs, NVIDIA does not publish a manufacturer's suggested retail price for data center accelerators; it sells primarily to system integrators, hyperscalers, and large enterprise customers under negotiated terms that are rarely disclosed. Every purchase-price figure in this report is therefore an estimate, reconstructed from analyst notes, reseller listings, or cost-modeling firms, and should be read as a defensible range rather than an authoritative list price. Second, cloud rental prices for high-demand SKUs like the B200 and GB300 NVL72 can change week to week as capacity comes online, making any single snapshot a moving target; this report notes the "as of" date for every figure so readers can judge how current a given number remains.

H100 Cloud and Purchase Pricing

The H100 is the benchmark against which every newer GPU's price is measured, in part because it is the most heavily tracked SKU in the market. Aggregator gpufinder.dev, which monitors 21 cloud providers, reported an on-demand price range of $1.29 to $127.82 per hour as of July 10, 2026, with the cheapest confirmed in-stock rate at **$1.33 per hour** on the Vast.ai marketplace and spot pricing available from $0.34 per hour (Source: gpufinder.dev). The same tracker put the H100 "floor price" at $2.33 per hour, describing this as mid-range relative to a six-month band of $2.00 to $15.20 per hour (Source: gpufinder.dev).

Independent trackers converge on a similar shape even when their absolute numbers differ. ComputeTape's June 2026 report calculated a median of **$4.29 per H100-hour** across seven sourced on-demand rows, with a public list-price band of $3.29 to $12.29 across five providers: RunPod Secure at $3.29, Crusoe at $3.90, and Lambda at $4.29 (Source: computetape.com). GetDeploying, which tracks 45 providers, reported an average on-demand price of $3.46 per hour, with the lowest confirmed spot price at $0.34 per hour (Source: getdeploying.com). CloudZero, an independent cost-tracking firm that states it does not sell GPU compute itself, put the market median around $2.29 to $3.12 per GPU-hour, with a wider range of $1.38 to $8.00-plus (Source: cloudzero.com).

Hyperscaler pricing sits at the top of the range. On **AWS**, the official EC2 P5 page confirms instances are "powered by NVIDIA H100 Tensor Core GPUs," with the p5.48xlarge configuration providing 8 GPUs and up to 3,200 Gbps of Elastic Fabric Adapter networking (Source: aws.amazon.com). Third-party pricing tracker Vantage lists that instance at $55.04 per hour, which normalizes to roughly $6.88 per GPU-hour (Source: instances.vantage.sh). On **Google Cloud**, the official accelerator-optimized VM pricing page lists the a3-highgpu-8g instance (8x H100) at $88.49 per hour on-demand and $38.32 per hour for the corresponding Spot price, working out to roughly $11.06 and $4.79 per GPU-hour respectively (Source: cloud.google.com). On **Microsoft Azure**, third-party tracker Vantage lists the ND96isr H100 v5 instance (8 GPUs) starting at $98.32 per hour on-demand, or $18.17 per hour with a spot commitment, normalizing to roughly $12.29 and $2.27 per GPU-hour (Source: instances.vantage.sh).

Purchase pricing for the physical card has been comparatively stable. CloudZero reports a new H100 costs $25,000 to $40,000 depending on the PCIe versus SXM5 variant (Source: cloudzero.com), a figure corroborated independently by DeployBase, which also notes buying outright runs "$25,000 to $40,000 depending on configuration" (Source: deploybase.ai) and by Compute Exchange, which pegs new H100 pricing at $25,000 to $40,000, refurbished at $21,000 to $34,000, and used at $15,000 to $28,000 (Source: compute.exchange.

H200 Cloud and Purchase Pricing

The H200 shares its Hopper compute die with the H100 but nearly doubles its memory capacity to 141GB of HBM3e, delivered at 4.8 TB/s of bandwidth versus the H100's 3.35 TB/s (Source: nvidia.com). NVIDIA's own benchmarks claim the H200 delivers 1.9 times faster Llama2 70B inference and 1.6 times faster GPT-3 175B inference than the H100 (Source: nvidia.com). Cloud pricing reflects a modest but real premium over the H100 rather than a dramatic one.

GPU aggregator gpufinder.dev listed H200 on-demand pricing across 14 providers from $1.32 to $87.35 per hour as of July 9, 2026, with the cheapest confirmed in-stock rate at **$1.32 per hour** on Vast, and a six-month range of $2.93 to $15.97 per hour (Source: gpufinder.dev) (Source: gpufinder.dev). Aggregator gpus.io reported H200 availability from 8 providers starting at $2.49 per GPU-hour on Theta EdgeCloud, with a market median of $3.82 per GPU-hour across 19 tracked configurations and a 90-day upward price trend of 6.4% (Source: gpus.io) (Source: gpus.io). GetDeploying, tracking 32-plus providers, found a 92% spread between the highest and lowest H200 listings, with rates reaching $13.78 per hour on-demand and a spot floor of $1.00 per hour (Source: getdeploying.com).

Specialist providers consistently undercut hyperscalers on H200. Thunder Compute's July 2026 comparison listed Hyperbolic at $2.40 per GPU-hour, Vast.ai at $3.71, Crusoe Cloud at $4.29, RunPod at $4.39, and Nebius at $4.50, and noted that "even after AWS's June 2025 price cut, the cheapest hyperscaler H200 hour costs 3 times more than Thunder Compute's H100" (Source: thundercompute.com. The same comparison put AWS at $7.91 per GPU-hour, Oracle Cloud at $10.00, and Azure at $10.60 (Source: thundercompute.com). GPU Cost Compare, updated June 17, 2026, showed Nebius as the cheapest on-demand provider at $4.50 per GPU-hour ("63% below the priciest listing"), with CoreWeave at $6.31, AWS at $7.91, Oracle at $10.00, and Google Cloud's a3-ultragpu-8g at $12.27 (Source: gpucostcompare.com). Google Cloud's own DevZero-sourced instance listing confirms the a3-ultragpu-8g configuration reaches $12.879 per GPU on-demand, with a spot rate of $4.094 (Source: devzero.io).

B200 and Blackwell Pricing

The B200 is NVIDIA's first mainstream Blackwell-architecture accelerator, built on TSMC's 4NP process with 208 billion transistors across a dual-die design, 192GB of HBM3e, and an 8 TB/s memory bus, doubling Hopper's throughput (Source: modal.com) (Source: modal.com). Because Blackwell dropped the PCIe form factor from major data center GPUs, deployment now requires top-tier NVLink-equipped systems, which constrains the supply channel and keeps pricing volatile (Source: thundercompute.com).

At the low end, marketplace aggregator Spheron advertises B200 rentals starting at **$2.71 per GPU-hour**, which it describes as "the lowest live marketplace rate," delivered through 8-GPU HGX B200 nodes connected via NVLink 5.0 (Source: spheron.network). Thunder Compute's July 2026 provider table showed Vultr and Hyperbolic both at $3.50 per GPU-hour, RunPod at $5.89, Vast.ai at $5.64, Modal at $6.25, Lambda at $6.69, and Nebius at $7.15, climbing to $14.00 (8 GPUs) on Oracle Cloud, $14.13 (8 GPUs) on AWS's p6-b200.48xlarge instance, and $27.04 (4 GPUs) on Azure's NDsrGB200NDRv6 instance (Source: thundercompute.com) (Source: thundercompute.com. A year earlier, in July 2025, Modal's own pricing survey showed a similarly wide spread: Lambda Labs at $3.79 per hour on-demand (falling to $2.99 with a three-year reservation), RunPod at $5.99, and AWS as high as $14.24 per hour on-demand, or a discounted $8.14 with a capacity block (Source: modal.com) (Source: modal.com).

Purchase pricing for the B200 is not publicly listed by NVIDIA; the chip is sold inside boards, servers, or racks rather than as a standalone unit. Cost-modeling firm Silicon Analysts estimates the B200's manufacturing cost, or cost of goods sold, at approximately **$6,400**, nearly double the H100's estimated $3,320, with HBM3e memory now representing 45% of total bill-of-materials cost, up from 41% on the H100 (Source: siliconanalysts.com) (Source: siliconanalysts.com. Against an implied selling price of roughly $40,000 per unit, that COGS estimate implies an approximate 84% gross margin (Source: siliconanalysts.com. That range is consistent with NVIDIA CEO Jensen Huang's own public guidance that the chip would cost "between $30,000 and $40,000 per unit," while stressing NVIDIA prefers to sell complete systems rather than bare chips (Source: siliconanalysts.com. At the system level, an 8-GPU DGX B200 appliance is priced by NVIDIA partners in the $300,000 to $500,000 range (Source: aiappstacker.com. NVIDIA's own DGX B200 product page confirms the system delivers "3X the training performance and 15X the inference performance of previous-generation systems" using 8 Blackwell GPUs interconnected with fifth-generation NVLink, for a combined 1,440GB of HBM3e memory and 64 TB/s of aggregate bandwidth (Source: nvidia.com) (Source: nvidia.com).

GB200 and GB300 NVL72 Rack-Scale Pricing

The GB200 NVL72 is not sold as a discrete GPU but as an entire rack: 72 Blackwell GPUs and 36 Grace CPUs bound into a single liquid-cooled, NVLink-connected system that NVIDIA describes as acting like "one giant GPU," delivering 30 times faster real-time trillion-parameter LLM inference than the prior generation (Source: primeline-solutions.com) (Source: primeline-solutions.com. Buyer guide GPUaaS.com describes a fully configured rack as costing "roughly $2-3M per rack, 120-130 kW of power draw, and a 1.36 metric ton chassis that doesn't fit through standard datacenter doors" (Source: gpuaas.com. Early analyst estimates from HSBC, reported by Tom's Hardware in 2024, pegged the smaller GB200 NVL36 rack at $1.8 million and the full NVL72 rack at $3 million, with an implied GB200 superchip ASP (average selling price) of $60,000 to $70,000 and a B100 ASP of $30,000 to $35,000 (Source: tomshardware.com.

Cloud rental of GB200 NVL72 capacity by the GPU is available from a growing number of neoclouds. Product review site AI App Stacker cites a per-GPU cloud rental range of roughly $10.50 to $27.00 per hour, which translates to $756 to $1,944 per hour for a full 72-GPU rack (Source: aiappstacker.com. GPU marketplace Spheron advertises a starting cloud price "from ~$10.50/hr" per GB200 GPU (Source: spheron.network. CoreWeave's own live pricing page, one of the earliest providers to offer GB200 NVL72 at scale, lists an on-demand price of **$42.00 per hour** and an inference-optimized single-GPU price of $10.50 per hour for its GB200 NVL72 configuration, alongside a still-quote-only "Contact sales" listing for the newer GB300 NVL72 (Source: coreweave.com) (Source: coreweave.com.

NVIDIA's Blackwell Ultra generation, headlined by the **GB300 NVL72**, delivers what the company describes as "1.5x more AI performance than the NVIDIA GB200 NVL72," alongside what NVIDIA calls a 50-fold increase in Blackwell's addressable revenue opportunity (Source: nvidianews.nvidia.com. Early cloud rental pricing for GB300 GPUs is emerging in the $3.02 to $30 per GPU-hour range depending on provider and commitment. Provider Runcrate lists on-demand GB300 rental at $4.25 per hour, with a $4.00 to $4.50 range across configurations and a reserved rate of $2.97 per hour (a 30% discount), against comparison figures of $7.20 on AWS, $7.50 on Azure, and $7.95 on Google Cloud (Source: runcrate.ai) (Source: runcrate.ai. Aggregator ComputePrices.com found the cheapest GB300 configuration, on provider Verda, starting at $3.02 per hour, roughly 60% below the tracked average (Source: computeprices.com. Full GB300 NVL72 racks are estimated to cost $3 million to $4 million, a step up from GB200 NVL72's $2 million to $3 million range (Source: aiappstacker.com.

Table 1 below consolidates on-demand cloud rental price ranges across the four GPU families discussed in this report, drawn from the aggregator and provider data cited above.

GPU (ARCHITECTURE)	CHEAPEST ON-DEMAND $/GPU-HR	TYPICAL HYPERSCALER $/GPU-HR	HIGHEST CONFIRMED $/GPU-HR	NEW UNIT PURCHASE PRICE
H100 (Hopper, 80GB)	$1.33 (Vast) (Source: gpufinder.dev)	$6.88 (AWS p5.48xlarge) (Source: instances.vantage.sh)	$127.82 (Source: gpufinder.dev)	$25,000-$40,000 (Source: cloudzero.com)
H200 (Hopper, 141GB)	$1.32 (Vast) (Source: gpufinder.dev)	$7.91-$12.27 (AWS/GCP) (Source: gpucostcompare.com)	$87.35 (Source: gpufinder.dev)	Not separately listed; systems priced with H100-comparable premium
B200 (Blackwell, 192GB)	$2.71 (Spheron) (Source: spheron.network)	$8.60 (CoreWeave HGX B200, per-GPU) (Source: coreweave.com)	$18.53 (GCP on-demand) (Source: modal.com)	~$30,000-$40,000 (Source: siliconanalysts.com)
GB200 NVL72 (per GPU)	$10.50 (Spheron/CoreWeave inference) (Source: coreweave.com)	$42.00 (CoreWeave on-demand) (Source: coreweave.com)	$27.00 (marketplace ceiling) (Source: aiappstacker.com)	$2M-$3M per 72-GPU rack (Source: aiappstacker.com)
GB300 NVL72 (Blackwell Ultra, per GPU)	$3.02 (Verda) (Source: computeprices.com)	$7.20-$7.95 (AWS/Azure/GCP) (Source: runcrate.ai)	$30.00 (early neocloud) (Source: aiappstacker.com	$3M-$4M per 72-GPU rack (Source: aiappstacker.com)

The table underscores a consistent pattern across every product tier: marketplace and specialist-cloud pricing sits at roughly one-quarter to one-tenth of hyperscaler on-demand rates for the same silicon, a gap explained less by hardware differences than by managed-service overhead, SLA guarantees, networking bundles, and regional availability. Buyers who can tolerate marketplace-grade support and variable capacity consistently pay far less per GPU-hour than buyers who need guaranteed hyperscaler capacity with enterprise support contracts.

Comparative Context and Market Positioning

Two comparisons put the H100-to-GB300 pricing ladder in context: how NVIDIA's own generations compare to each other, and how NVIDIA compares to its primary merchant-silicon competitor, AMD.

Generation over generation, the price-per-GPU-hour increase has been far smaller than the underlying performance gain. The **H200**, despite offering nearly double the H100's memory capacity, in practice rents for only a modest premium over H100 rates on most specialist clouds (Source: thundercompute.com. The **B200**, by contrast, carries a much larger step-up in both purchase price (roughly $30,000-$40,000 versus the H100's $25,000-$40,000, but with materially higher performance) (Source: siliconanalysts.com) and cloud rental, commonly double to triple H100 hyperscaler rates for equivalent 8-GPU nodes (Source: aiappstacker.com). Thunder Compute's own comparison concluded that "for many use cases, two H100 GPUs can deliver similar performance with a similar combined VRAM capacity," making the B200 a poor value proposition "for most teams" as of mid-2026, given waitlists and enterprise-restricted access (Source: thundercompute.com.

AMD's **Instinct MI300X** is the most established non-NVIDIA alternative in merchant silicon, with 192GB of HBM3 memory (2.4 times the H100's capacity) at a price point close to the H100's own cloud rate. Aggregator CloudGPUPrice found MI300X available from $1.85 per hour versus H100 SXM from $1.57 per hour as of June 2026 (Source: cloudgpuprice.com. GridStackHub's May 2026 comparison found a similarly narrow $0.11 per hour gap between the two chips, noting that MI300X's memory advantage lets a single card replace two H100s for 70B-plus parameter models, "cutting effective cost in half" for memory-bound workloads (Source: gridstackhub.ai. However, semiconductor analysis firm SemiAnalysis, after benchmarking both platforms extensively, concluded that "for customers that are using short to medium term rentals (sub-6 month) from Neoclouds, Nvidia always wins on performance per dollar," attributing this to a scarcity of dedicated AMD neoclouds that keeps MI300X and MI325X rental rates elevated (Source: newsletter.semianalysis.com. AMD's next-generation MI325X, with 256GB of HBM3e, starts at $2.00 per hour on Vultr, against $1.75 per hour for MI300X on the same platform (Source: vultr.com.

NVIDIA's structural advantage remains the CUDA software ecosystem, which market researcher Markets NXT describes as "creating a switching cost moat that is as or more durable than its hardware performance advantage," built on "decades of developer code, tooling, and model libraries" that AMD, Intel, and custom silicon "cannot dislodge without matching the software investment" (Source: marketsnxt.com. That moat helps explain why, despite AMD offering comparable or better raw specifications at similar list prices, NVIDIA GPUs continue to command the overwhelming majority of both cloud provider inventory and enterprise procurement budgets.

Data Analysis and Evidence

The scale of demand underlying these prices is best illustrated by NVIDIA's own financial disclosures. In its fiscal first quarter of 2027 (ended April 26, 2026), NVIDIA reported record revenue of **$81.6 billion**, up 85% year over year, with record Data Center revenue of **$75.2 billion**, up 92% year over year (Source: nvidianews.nvidia.com) (Source: sec.gov). Within that figure, Data Center compute revenue alone reached $60.4 billion, and Data Center networking revenue reached $14.8 billion, up 199% year over year (Source: nvidianews.nvidia.com. The prior quarter (fiscal fourth quarter 2026) saw record quarterly revenue of $68.1 billion, up 73% year over year, and full fiscal-year 2026 revenue of $215.9 billion, up 65% (Source: nvidianews.nvidia.com). CNBC's earnings coverage noted the company "now gets over 91% of sales from its data center unit" (Source: cnbc.com.

Independent market-size estimates for the addressable GPU market vary considerably by scope and definition, which itself is a useful data point about how immature standardized market measurement remains in this category. MarketsandMarkets values the global data center GPU market at **$119.97 billion** in 2025, projected to reach $228.04 billion by 2030 at a 13.7% compound annual growth rate (Source: marketsandmarkets.com. Technavio's March 2026 report projects the market will grow by $125.43 billion between 2025 and 2030, a 15.8% CAGR, with on-premises deployment alone valued at $59.05 billion in 2024 (Source: technavio.com. Mordor Intelligence, scoping specifically to AI training GPUs, sizes that narrower market at $30.84 billion in 2026, growing to $98.65 billion by 2031 at a 26.18% CAGR, and notes hyperscale and cloud installations accounted for 70.27% of 2025 revenue (Source: mordorintelligence.com (Source: mordorintelligence.com. Precedence Research, using a still narrower "AI data center GPU" scope, put 2025 revenue at just $10.51 billion, forecasting growth to $77.15 billion by 2035 (Source: precedenceresearch.com. These figures should not be reconciled as if they measure the same thing; they differ by scope (data center GPUs broadly versus AI training GPUs specifically), time horizon, and methodology, but the direction is unanimous: double-digit compound annual growth through at least 2030.

Table 2 below lines up these four independent market-size estimates side by side, along with their scope and base year, to make the methodology differences explicit rather than implicit.

RESEARCH FIRM	MARKET SCOPE	BASE YEAR VALUE	FORECAST VALUE	CAGR / PERIOD
MarketsandMarkets	Data center GPU (broad)	$119.97B (2025)	$228.04B (2030)	13.7% (2025-2030) (Source: marketsandmarkets.com)
Technavio	Data center GPU (broad)	Not stated (2024 base)	+$125.43B by 2030	15.8% (2025-2030) (Source: technavio.com)
Mordor Intelligence	AI training GPU (narrow)	$25.28B (2025)	$98.65B (2031)	26.18% (2026-2031) (Source: mordorintelligence.com)
Precedence Research	AI data center GPU (narrowest)	$10.51B (2025)	$77.15B (2035)	22.06% (2025-2034) (Source: precedenceresearch.com)

The nearly twelve-fold gap between Precedence Research's $10.51 billion 2025 base and MarketsandMarkets' $119.97 billion figure is not an error in either report; it reflects fundamentally different market boundaries, one counting only AI-specific data center GPU revenue narrowly defined, the other counting the entire data center GPU category including non-AI workloads such as visualization and general HPC. Readers using any single market-size figure in a procurement or investment context should confirm which scope definition it uses before comparing it against a different source.

The single largest cost pressure feeding into every GPU price in this report is the high-bandwidth memory (HBM) shortage. CNBC reported in January 2026 that DRAM prices were "expected to rise more than 50%" quarter over quarter, with Taipei-based researcher TrendForce calling the increase "unprecedented" (Source: cnbc.com. Micron's business chief Sumit Sadana told CNBC the company has "seen a very sharp, significant surge in demand for memory" that "far outpaced" the entire memory industry's supply capability (Source: cnbc.com. Each GPU generation demands sharply more HBM: the H100 uses 80GB of HBM3, the H200 uses 141GB of HBM3e across six stacks, the B200 requires 192GB of HBM3e, and the B300 (Blackwell Ultra) pushes that to 288GB of 12-layer HBM3e (Source: blog.barrack.ai. Research firm IDC, quoted in the same analysis, called the shift "a potentially permanent, strategic reallocation of the world's silicon wafer capacity" toward high-margin AI memory products (Source: [blog.barrack.ai](https://blog.barrack.ai/2026-gpu-memory-crisis/#:~:text=According%20to%20IDC%2C%20it%20represents%20%22a%20potentially%20permanent%2C%20strategic%20reallocation%20of%20the%20world%27s%20silicon%20wafer%20capacity%22%20toward%20high%20margin%20AI%20memory%20products. Independent analysis from AInvest found front-end wafer allocation to HBM production is estimated to reach 480,000 to 520,000 wafers per month by the end of 2026, up from roughly 380,000 to 400,000 wafers per month at the end of 2025, a 30% increase in a single year (ainvest.com. Looking further out, analysts at Bernstein, cited by StockWireX, project HBM contract prices could rise a further 2 to 2.5 times by 2027, a shock that would amplify roughly fourfold at the finished-GPU level (Source: stockwirex.com.

Reddit sentiment, while not a substitute for verified pricing data, corroborates the volatility documented above. Threads on r/StockMarket in 2026 documented sharp short-term swings, including a reported 30% weekend decline in B200 rental prices and a 37.7% decline in H200 rental rates through the second half of a single month (Source: reddit.com. An earlier but widely discussed analysis shared on r/LocalLLaMA, titled "$2 H100s: How the GPU Rental Bubble Burst," argued that "H100s were $8/hr if you could get them" at the height of the 2023 to 2024 shortage, and that within about a year "there's 7 different resale markets selling them under $2" (Source: latent.space. These community reports should be read as anecdotal color rather than authoritative pricing, but they are directionally consistent with the aggregator data cited throughout this report: extreme, fast-moving dispersion driven by marketplace supply catching up to what was, as recently as 2023 and 2024, a severe shortage.

Case Studies and Real-World Examples

CoreWeave and Mistral AI: GB200 NVL72 at Production Scale

CoreWeave, the largest publicly traded GPU-only cloud provider, became one of the first cloud providers to bring GB200 NVL72 systems online at scale in April 2025, with Cohere, IBM, and Mistral AI as its first customers (Source: blogs.nvidia.com. Mistral AI, a CoreWeave customer since 2023 that initially signed on for H100 access before expanding to H200s and GB200 clusters, reported it was able to train its large language models **2.5 times faster** using GB200 NVL72 than on previous hardware generations (Source: datacenterdynamics.com (Source: datacenterdynamics.com. CoreWeave CEO Mike Intrator framed the launch as evidence of speed to market: "We work closely with NVIDIA to quickly deliver to customers the latest and most powerful solutions for training AI models and serving inference" (Source: blogs.nvidia.com. Reviewer AI Tool Discovery reports Mistral AI achieved a 75% cloud cost reduction after migrating workloads to CoreWeave, and cites CoreWeave's own pricing page showing H100 PCIe at $4.25 per hour in Q1 2026, versus $6.88 on AWS and $12.29 on Azure for comparable capacity (Source: aitooldiscovery.com (Source: aitooldiscovery.com. This case illustrates how contract-based neocloud pricing, rather than public on-demand rate cards, increasingly determines the effective GPU price for frontier AI labs.

Amazon and Anthropic: Hyperscaler H100/H200 Adoption

AWS's own EC2 P5 product page carries a direct customer endorsement from Anthropic co-founder Tom Brown, who stated the company was "using Amazon EC2 P4 instances extensively today" and was "excited about the launch of P5 instances," expecting them "to deliver substantial price-performance benefits over P4d instances" at "the massive scale required for building next-generation LLMs" (Source: aws.amazon.com. Cohere CEO Aidan Gomez similarly credited H100-powered P5 instances with helping "businesses create, grow, and scale faster" by combining "computing power" with "state-of-the-art LLM and generative AI capabilities" (Source: aws.amazon.com. These endorsements illustrate why hyperscaler pricing, despite sitting well above marketplace rates, retains a durable customer base among frontier labs that prioritize guaranteed capacity, integrated tooling, and enterprise support over minimizing per-GPU-hour cost.

Operation Gatekeeper: A $160 Million H100/H200 Smuggling Ring

On December 8, 2025, federal prosecutors in Texas unsealed documents describing "Operation Gatekeeper," a smuggling network accused of attempting to export at least **$160 million** worth of NVIDIA H100 and H200 GPUs to China between October 2024 and May 2025, using front companies and mislabeled shipping paperwork that described the chips as "adapters" and "contactor controllers" (Source: cnbc.com (Source: cnbc.com. SemiAnalysis analyst Ray Wang told CNBC that "more than 60% of the leading AI models in China are currently using Nvidia's hardware," underscoring the economic incentive behind such smuggling (Source: cnbc.com. Researchers at Epoch AI, examining the full pattern of diversion and resale, estimate with 90% confidence that between 290,000 and 1.6 million H100-equivalent GPUs were smuggled into China through the end of 2025, with a median estimate of 660,000, comparable to the entire compute stockpile of frontier U.S. AI lab xAI at the time (Source: epoch.ai.

China's Black Market Repricing of the DGX B300

Perhaps the starkest illustration of export controls' effect on GPU pricing comes from China's restricted-hardware black market. Reporting from otontechnology.com, citing the Financial Times' sourcing of multiple Chinese chip traders, found that NVIDIA's flagship DGX B300 server "now sells for more than 8 million yuan ($1.1 million), up from 4 million yuan six months ago and well above its US retail value" (Source: otontechnology.com. A parallel report from KeyToFinancialTrends noted that a DGX B300 "which carries a US retail price of roughly $400,000, is changing hands in China at prices above 8 million yuan," while the AMD-competing "Nvidia RTX 6000 Pro workstation chip has surged from around 50,000 yuan at the start of 2026 to as much as 130,000 yuan" (Source: keytofinancialtrends.com (Source: keytofinancialtrends.com. This is a clean natural experiment: identical hardware, roughly triple the price, purely as a function of legal accessibility.

The Secondary Market: H100 Depreciation After Blackwell's Ramp

As Blackwell-generation GPUs ramp into production, previous-generation H100 hardware is flowing onto the secondary market at increasingly steep discounts. Auction and reseller data compiled by CraftRigs found that "used H100s hit $8,200-$12,500 in March 2026 auctions," and that the March 2025 launch of NVIDIA's B100 "triggered 34% H100 SXM5 depreciation by January 2026, per Exit Technologies auction data" (Source: craftrigs.com (Source: craftrigs.com. UK reseller Servnet, by contrast, reports H100 cards "that sold for around $40,000 in late 2023 now move on secondary markets for roughly $12,000-$22,000 used, with refurbished units around $21,000-$34,000," and forecasts that Blackwell's general availability will push H100 secondary pricing down a further 10% to 20% as fleets rotate (Source: servnetuk.com (Source: servnetuk.com. GPU financing firm Mercatus separately tracks a residual-value curve showing H100 cards at 18 to 30 months of age retaining 75% to 85% of a $30,000 OEM list reference, or roughly $22,000 to $25,500, before declining to a 25% to 35% floor after 60 months (Source: mercatus-ai.com. Analytics firm Hashrate Index notes that at peak 2023-2024 scarcity, used and refurbished H100s traded "as high as $50,000 per GPU," well above the eventual new-unit list price, before dropping sharply as supply normalized (Source: hashrateindex.com.

Lambda Labs, CoreWeave, and the Neocloud Pricing Race

The rise of "neoclouds," specialist providers that rent NVIDIA GPUs without the broader cloud portfolio of AWS, Azure, or Google Cloud, has been one of the defining structural changes in GPU pricing since 2023. **Lambda Labs**, one of the earliest neoclouds to court individual researchers and startups, illustrates how quickly on-demand rate cards can move: the company raised its H100 SXM on-demand price from $2.59 to $3.49 per hour in January 2024, a 35% increase, according to a review of Lambda's own pricing history by American Compute (Source: amcompute.com. That kind of single-quarter repricing, driven by real-time supply and demand rather than an annual list-price cycle, is characteristic of the neocloud segment and helps explain why aggregator "current price" figures can shift meaningfully between report vintages just weeks apart.

CoreWeave's growth trajectory illustrates the scale these neoclouds have reached. The company priced its March 2025 initial public offering at $40 per share, implying a $35 billion valuation (Source: aitooldiscovery.com, and had already signed an $11.9 billion, five-year compute deal with OpenAI in 2023 (Source: aitooldiscovery.com. Its contracted revenue backlog, a widely watched proxy for future GPU demand, was reported at $30.1 billion in June 2025 by AI Tool Discovery, $66.8 billion at the end of 2025 by American Compute (roughly 13 times annual revenue), and over $88 billion as of the first quarter of 2026 by TechStackIPO (Source: aitooldiscovery.com (Source: amcompute.com (Source: techstackipo.com. Independent research firm Useluminix separately put the figure at $99.4 billion as of March 2026 (Source: useluminix.com. The variation across these figures is a useful reminder that "backlog" is not a standardized accounting term across providers and reporting dates, and that readers should treat single-source backlog figures as approximate rather than precise. By the same measurements, TechStackIPO put CoreWeave's GPU fleet at over 250,000 units across 32 data centers, against a private valuation of $4 billion to $5 billion and roughly $505 million in run-rate revenue for smaller rival Lambda Labs, which has raised $2.3 billion in total funding including a $1.5 billion Series D round in November 2025 (Source: techstackipo.com (Source: techstackipo.com.

The Older-Generation Secondary Market: A100 as a Pricing Anchor

Even as H100, H200, and B200 pricing dominates headlines, the still-active market for the previous-generation **A100** provides a useful pricing anchor for cost-sensitive inference and fine-tuning workloads. Hashrate Index reports used A100 pricing spans roughly $7,800 to $18,900 depending on SKU and condition, with active turnover across both formal ITAD (IT asset disposition) channels and informal resale markets (Source: hashrateindex.com. The same analysis notes that for many inference and fine-tuning workloads, "the relevant metric is cost per completed task, not cost per hour," and that used A100s "often win that math against newer hardware at 2-3x the price" (Source: hashrateindex.com. The same report flags a broader accounting controversy shaping how providers set depreciation schedules, and therefore hourly rates: investor Michael Burry has estimated roughly $176 billion in understated depreciation across the AI infrastructure industry, a claim that, if accurate, implies current cloud GPU pricing may not fully reflect the true economic cost of the underlying hardware (Source: hashrateindex.com. Real-world evidence from Azure, CoreWeave, and hyperscaler virtual-machine retirements, however, suggests GPUs retain economic value for five to seven-plus years, longer than some depreciation schedules assume (Source: hashrateindex.com.

Implications and Future Directions

Several structural forces will shape GPU pricing over the next 12 to 24 months. First, the HBM supply constraint is unlikely to resolve quickly. Total 2026 DRAM wafer starts are estimated at approximately 18 million, with HBM's share of that capacity rising steadily as manufacturers redirect fabrication resources away from commodity memory (Source: ainvest.com. If Bernstein's projected 2 to 2.5 times HBM contract price increase for 2027 materializes, buyers should expect further upward pressure on both new-unit GPU purchase prices and, with a lag, cloud rental rates, particularly for memory-heavy SKUs like the H200, B200, and future Blackwell Ultra and Rubin-generation parts (Source: stockwirex.com.

Second, the neocloud financing model that has driven marketplace rates well below hyperscaler rates carries its own risk that could eventually push prices back up. CoreWeave posted an adjusted EBITDA margin of roughly 60% in 2025 while still reporting a net loss exceeding $1 billion, a gap explained by depreciation and interest expense on GPU-backed debt (Source: zettabyte.space. The company carried approximately $21.4 billion in debt at the end of 2025 and paid roughly $1.2 billion in interest that year (Source: zettabyte.space. Depreciation schedules vary meaningfully across providers, from CoreWeave's six years to Lambda's five and Nebius's four, a choice that materially affects reported per-GPU-hour economics without changing the underlying hardware cost (Source: zettabyte.space. Research firm Useluminix frames the entire neocloud model as "a financed wager" on whether GPUs retain a five-to-six-year useful economic life rather than a two-to-three-year one, with CoreWeave's $99.4 billion contracted backlog as of March 2026 serving as the primary de-risking mechanism (Source: useluminix.com (Source: useluminix.com. If AI compute demand growth slows before that backlog converts to sustainable free cash flow, marketplace-rate compression could reverse.

Third, export-control enforcement is tightening rather than loosening. The U.S. Department of Commerce closed re-export loopholes at the end of May 2026 that had allowed high-end Blackwell and Rubin-generation chips to reach Chinese companies through foreign subsidiaries of non-Chinese entities, a move that directly triggered the latest black-market price surge documented above (Source: keytofinancialtrends.com. As long as this policy gap persists, expect the wedge between U.S. list price and Chinese black-market price for restricted SKUs to remain wide, with periodic step-changes each time enforcement tightens further.

Finally, NVIDIA's own roadmap points toward continued rapid product cycling. The company's next-generation Rubin GPU, which NVIDIA says recently entered production, ships with up to 288 gigabytes of next-generation HBM4 memory per chip as part of the NVL72 rack-scale system (Source: cnbc.com. CEO Jensen Huang has characterized Grace Blackwell with NVLink as "the king of inference today," delivering "an order-of-magnitude lower cost per token," while promising that the forthcoming Vera Rubin platform "will extend that leadership even further" (Source: nvidianews.nvidia.com. Each new architecture historically arrives at a higher absolute price point but a lower cost-per-unit-of-performance, a pattern buyers should expect to continue even as HBM scarcity pushes nominal prices upward across every existing SKU.

Frequently Asked Questions (FAQs)

What is the current price of an NVIDIA H100? Cloud rental ranges from about $1.33 per GPU-hour on marketplace platforms like Vast.ai up to $12.29 per GPU-hour on premium hyperscaler instances, with a market median commonly cited between $2.29 and $4.29 per GPU-hour (Source: gpufinder.dev) (Source: computetape.com. A new H100 card purchased outright costs $25,000 to $40,000 (Source: cloudzero.com.

How much does an NVIDIA H200 cost? H200 cloud rental spans roughly $1.32 to $13.78 per GPU-hour, with a market median around $3.82 per GPU-hour, a modest premium over H100 given the H200's 141GB of HBM3e memory (Source: gpufinder.dev) (Source: gpus.io.

What is the NVIDIA B200 price? B200 cloud rental starts around $2.71 per GPU-hour on specialist marketplaces and reaches $8.00 per GPU-hour on mainstream neocloud on-demand tiers, with hyperscaler allocations running higher still (Source: spheron.network (Source: aiappstacker.com. Purchase price is estimated at $30,000 to $40,000 per GPU, though NVIDIA does not sell it as a standalone retail unit (Source: siliconanalysts.com.

What is the GB200 NVL72 street price? A full 72-GPU rack is estimated to cost $2 million to $3 million, with cloud rental available at roughly $10.50 to $27.00 per GPU-hour (Source: aiappstacker.com.

Is H100 or H200 the better value? For most workloads under 40GB of active model weights, the H100 offers similar or better cost-per-token given its wider availability and lower cloud rental floor; the H200's extra memory becomes valuable specifically for long-context inference or models that overflow 80GB VRAM (Source: [thundercompute.com](https://www.thundercompute.com/blog/nvidia-h200-pricing#:~:text=Choose%20H200%20only%20when%20you%20must.%20Running%20massive%20models%20that%20overflow%2080GB%20VRAM%20or%20long-context%20inference.

How do AI GPU market trends look going into 2027? Independent researchers project double-digit compound annual growth through at least 2030 across every market-size definition tracked, from Mordor Intelligence's 26.18% CAGR for AI training GPUs specifically to MarketsandMarkets' 13.7% CAGR for the broader data center GPU category (mordorintelligence.com (Source: marketsandmarkets.com.

What is the difference between NVIDIA Blackwell pricing on the B200 versus the rack-scale GB200/GB300 NVL72? The B200 is sold and rented as an individual GPU or as part of an 8-GPU HGX node, with cloud rates from roughly $2.71 to $8.00 per GPU-hour on mainstream neoclouds and higher still on constrained hyperscaler capacity, as detailed above. The GB200 and GB300 NVL72 are sold only as complete 72-GPU, 36-CPU racks, priced in the millions of dollars, with per-GPU cloud rental typically running $10.50 to $30.00 per hour depending on generation and provider, reflecting both the newer Blackwell Ultra silicon in GB300 and the added value of a unified 130 TB/s NVLink domain that lets the rack behave as a single accelerator (Source: aiappstacker.com (Source: aiappstacker.com.

Why is Google Cloud's B200 pricing listed as "custom quote" rather than a fixed rate? As of mid-2026, Google Cloud's A3 Blackwell instances remain in a constrained rollout where capacity is allocated through direct sales engagement rather than posted on-demand rates, a pattern Thunder Compute's pricing survey also found at Crusoe Cloud (Source: thundercompute.com. This mirrors the broader Blackwell rollout pattern: NVIDIA's PCIe-free Blackwell design requires top-tier NVLink systems, and hyperscalers with the largest committed customers frequently prioritize direct enterprise agreements over public self-service pricing during the early ramp of a new architecture, a constraint that persists even as marketplace providers post fixed rates (Source: northflank.com.

Conclusion

The NVIDIA GPU price index in mid-2026 is not a single number but a wide and fast-moving band, shaped as much by channel, contract length, and geography as by silicon. An H100 can cost anywhere from roughly $1.33 to over $100 per hour depending on where it is rented, a spread that persists across every newer generation as well, from H200 through the rack-scale GB300 NVL72. Beneath that dispersion sit three durable forces: relentless hyperscaler demand that pushed NVIDIA's Data Center revenue past $75 billion in a single quarter, a structural HBM memory shortage that is inflating the cost of every new GPU generation faster than compute costs alone would predict, and a bifurcated global market in which export controls have made restricted hardware in China cost roughly double or more its U.S. retail price. Buyers evaluating GPU costs in the second half of 2026 should treat any single "price per GPU-hour" figure as a starting point rather than an answer, and should weight marketplace rates, hyperscaler rates, purchase prices, and secondary-market depreciation separately when building total cost of ownership models, since each of those four channels is currently moving on a different trajectory.

Tags: gpu price index, nvidia h100 price, nvidia h200 cost, nvidia b200 price, gb200 street price, ai gpu market trends, nvidia blackwell pricing, enterprise gpu cost comparison, h100 vs h200 pricing, gpu cloud pricing