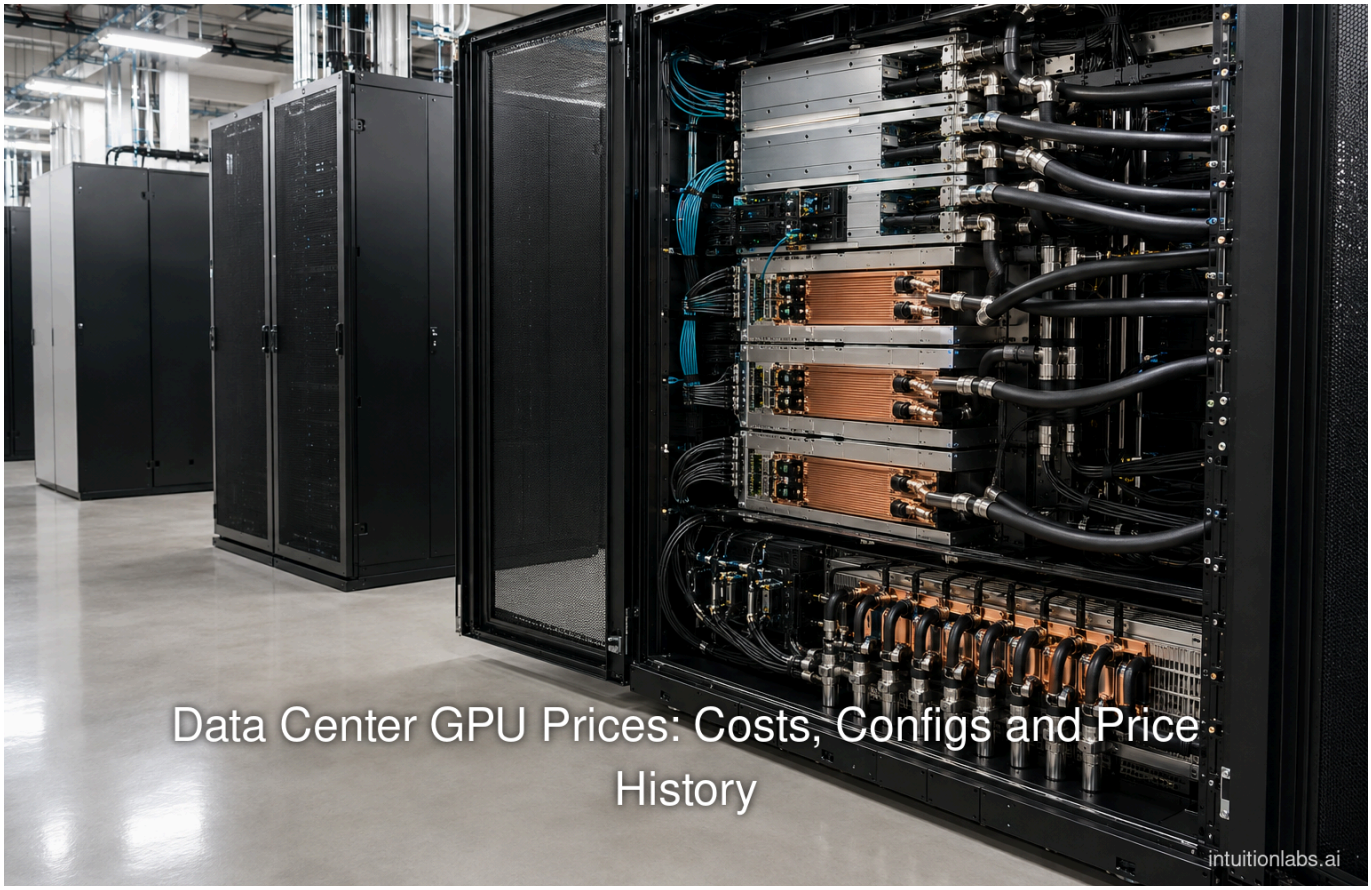
# Data Center GPU Prices: Costs, Configs and Price History

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · data center gpu prices · gpu pricing · h100 price · nvidia gpu cost · gpu cluster pricing · ai gpu price history · blackwell gpu price · enterprise gpu pricing




## Executive Summary

Data center GPU prices in 2026 span a wide range that depends on form factor, form of purchase, and how recently a source was updated, and no single number answers "how much does a data center GPU cost." Cited third-party pricing guides estimated a single NVIDIA H100 (80GB HBM3) at roughly $25,000 to $31,000 as of August 2026 (jarvislabs.ai) (cloudzero.com), with an 8-GPU HGX server running $250,000 to $320,000. The cited third-party guide estimated the newer H200 (141GB HBM3e) at $30,000 to $55,000 per unit and $320,000 to $420,000 per 8-GPU server (mercatus-ai.com), while NVIDIA's current Blackwell generation pushed prices higher still: a cited third-party guide estimated a B200 at $30,000 to $50,000 per unit and $400,000 to $500,000 per 8-GPU server (mercatus-ai.com), and a full GB300 NVL72 rack of 72 GPUs was documented anywhere from $3 million to $6.5 million depending on the measurement method used (gpusmith.com) (datagravity.dev).

Cloud rental rates, gathered directly from provider pricing pages and APIs in September 2026, show an even wider structural spread: hyperscaler on-demand pricing ran roughly $6.88 to $14.24 per GPU-hour equivalent on AWS, Google Cloud, and Microsoft Azure (pricing.us-east-1.amazonaws.com) (prices.azure.com), against $1.99 to $6.79 per GPU-hour on specialist neoclouds and marketplaces such as Voltage Park, RunPod, and Vast.ai (voltagepark.com) (runpod.io). Independent tracking from SemiAnalysis shows H100 spot-contract pricing falling from $6.62 per hour in late 2023 to $2.83 per hour by February 2026 (gpu-index.semianalysis.com), even as some listings simultaneously showed capacity sold out.

The forces behind these numbers are documented in public filings: NVIDIA's Data Center segment generated $89.0 billion in revenue in its fiscal quarter ended July 26, 2026, up 117 percent year over year (sec.gov), while high bandwidth memory now accounts for 30 to 40 percent of an AI accelerator's build cost, up from under 20 percent two generations earlier (fool.com). Alternatives are gaining ground on price: the cheapest genuinely on-demand AMD Instinct MI300X capacity rented for $2.59 per GPU-hour as of mid-2026 (spheron.network), and TrendForce projects ASIC-based servers will reach 27.8 percent of 2026 AI server shipments (trendforce.com).

This report's central conclusion is methodological as much as numerical: every price above is a dated observation, not a fixed fact, and comparisons that ignore whether a figure reflects an advertised rate, a spot price, or a completed multi-year contract will systematically mislead. Readers should treat the tables and sourced ranges below as a reproducible starting point to verify against the same primary sources on any future date.

## Introduction and Background

A **data center GPU** is a graphics processing unit engineered for parallel computation at scale: training and running large AI models, high performance computing, and virtualized workloads inside data centers, as distinct from consumer or gaming cards. Unlike consumer hardware, NVIDIA has never published an official list price for its data center GPUs. The company sells almost exclusively through original equipment manufacturers (OEMs), cloud partners, and system integrators, so "the price of an H100" is really a range of observed market transactions rather than a single number (cloudzero.com). The absence of a public list price is the central reason pricing guides for this category diverge so widely, and it is why this report treats every figure below as a dated observation rather than a fixed fact.

This article updates and extends IntuitionLabs' prior GPU pricing coverage. An earlier analysis on this site, last revised in April 2026, catalogued NVIDIA's Hopper-era pricing in detail (intuitionlabs.ai). This report does not restate that ground. Instead it adds five months of newer market data through September 2026, extends coverage to the Blackwell generation's rack-scale systems, brings in AMD and Intel accelerators as competitive reference points, and publishes a reproducible price history spanning the NVIDIA A100 (2020) through the B300 "Blackwell Ultra" (2026).

**Methodology note:** every price below is labeled by what it actually measures. Three layers are kept distinct throughout: (1) **hardware purchase prices**, the cost to buy a bare GPU module or a multi-GPU server outright; (2) **rack-scale system prices**, for integrated products such as the GB200 and GB300 NVL72; and (3) **cloud rental rates**, the dollar-per-GPU-hour cost of renting capacity rather than owning it. Within each layer, this report further distinguishes **advertised or list-adjacent prices** from **completed transactions** where evidence of an actual deal exists (for example, contract renewal data), and states the source's own "as of" date wherever one was published.

Readers wishing to reproduce this comparison can pull the same cloud-provider pricing pages and APIs cited below on any date; given the pace of change in this market, expect the specific numbers to have moved by the time you check them.

IntuitionLabs is a life sciences and AI consultancy founded in 2023 (intuitionlabs.ai); it does not sell GPUs, cloud capacity, or competing infrastructure, and nothing below should be read as a product recommendation from an interested vendor.

## Hopper Generation Pricing: H100 and H200 Configurations and Costs

The **NVIDIA H100** (Hopper architecture, 80GB HBM3 memory) has been the reference AI training and inference chip since its 2022 launch, and it remains the most heavily documented data center GPU on the market. As of August 2026, a new 80GB H100 card was priced at roughly $31,000 (cloudzero.com), while another 2026 guide put a single unit closer to $25,000 (jarvislabs.ai); the spread itself illustrates the absence of a fixed list price. At the system level, a new 8-GPU HGX H100 board runs $250,000 to $320,000 (cloudzero.com), a figure consistent with a separate build-up of $200,000 in GPUs plus roughly $50,000 of supporting infrastructure (jarvislabs.ai). A complete **DGX H100** appliance, NVIDIA's own integrated 8-GPU system, sells for approximately $300,000 to $400,000 (thundercompute.com). A PCIe variant, the **H100 NVL**, pairs two cards over an NVLink bridge into a pooled 188GB HBM3 memory space aimed at large language model inference, per NVIDIA's own product page (nvidia.com); the SXM form factor draws up to 700W, versus 350 to 400W (configurable) for the PCIe and NVL variants (nvidia.com).

The **H200** (141GB HBM3e) commands a premium for its larger, faster memory pool. Single-unit prices in 2026 span $30,000 to $40,000 at street price, or $40,000 to $55,000 at list (jarvislabs.ai), while a separate guide narrows SXM pricing to $32,000 to $40,000 and quotes the PCIe card at $28,000 to $34,000 (mercatus-ai.com). At the system level, an 8-GPU HGX H200 server runs $320,000 to $420,000, roughly $370,000 typical (mercatus-ai.com), roughly 30 percent above the comparable 8-GPU H100 build (mercatus-ai.com). One vendor's cost breakdown of that $370,000 system attributes $256,000 to $320,000 to the eight GPUs themselves, with the remainder split across chassis, NVSwitch fabric, networking, storage, and integration (mercatus-ai.com). A 4-GPU HGX H200 board lists at $180,000 to $220,000 (jarvislabs.ai). At its Q4 2024 launch, H200 OEM pricing ran $40,000 to $48,000 per GPU with cloud rental at $5.00 to $8.00 per hour; by 2026 those figures had compressed to $32,000 to $40,000 and a wider $2.50 to $7.00 per hour band as more suppliers entered the market (mercatus-ai.com). On a like-for-like basis, one guide puts H100 hardware at $25,000 to $30,000 against H200's $30,000 to $40,000 (jarvislabs.ai), a 15 to 20 percent premium; H200's cloud rental rate carries a comparable 48 percent premium for 76 percent more memory and 60 percent more bandwidth (jarvislabs.ai).

Hopper's residual value has held up unusually well given the chip's age. Secondary-market H100 SXM units have traded as low as $6,000 to $15,000 as Blackwell supply expanded (thundercompute.com), while a 2026 secondary-marketplace survey put new H100 cards at $25,000 to $40,000, refurbished units at $21,000 to $34,000, and used units at $15,000 to $28,000 (compute.exchange). Cloud pricing for H100 fell sharply as Blackwell reached the market: AWS cut its H100 on-demand pricing by roughly 44 percent in June 2025 (cloudzero.com), a trend examined further in the cloud pricing section below.

## Blackwell Generation Pricing: B200, B300, and Rack-Scale Systems

NVIDIA's **Blackwell** architecture, which NVIDIA announced, introduced the **B200** (192GB HBM3e) and the higher-memory **B300** "Blackwell Ultra" (288GB HBM3e). Standalone B200 cards list at $30,000 to $40,000 in 8-GPU volumes, though street quotes ran $45,000 to $50,000 under 2026 supply constraints (mercatus-ai.com). Cost-modeling firm Silicon Analysts estimates the B200's manufacturing cost (bill of materials, not sale price) at approximately $6,400, nearly double the H100's estimated $3,320, driven by HBM3e now representing about 45 percent of the total bill of materials (gpusmith.com), a data point that helps explain why finished-system prices scale up so quickly. An 8-GPU HGX B200 server costs $400,000 to $500,000 in 2026, roughly $450,000 as a typical integrated quote, about $56,000 per GPU deployed (mercatus-ai.com), and NVIDIA's own **DGX B200** appliance lists near $515,000 (mercatus-ai.com); NVIDIA's DGX B200 datasheet states the system's maximum power draw at approximately 14.3 kilowatts (nvidia.com).

The B300 commands a further step up. An 8-GPU **DGX B300** system is anchored at $300,000 to $350,000, roughly $37,500 to $43,750 per GPU at the system level, a range that held from the first quarter of 2026 through September 2026 (tech-insider.org). A single B300 GPU purchased outright ran about $53,000 as of July 2026 (tech-insider.org), and a single-GPU DGX Station, pairing one B300 with a Grace CPU in a desktop tower, was priced at $80,000 to $125,000 as of September 2026 (tech-insider.org).

Rack-scale systems are where Blackwell pricing becomes genuinely difficult to pin down, because the same product is described using at least three different measurement methods. An early teardown of the **GB200 NVL72** rack (72 GPUs plus 36 Grace CPUs in one liquid-cooled cabinet) pegged the smaller NVL36 variant at $1.8 million and the full NVL72 rack at $3 million, implying a GB200 superchip average selling price of $60,000 to $70,000 (gpusmith.com). For its successor, one analyst estimate places a full **GB300 NVL72** rack at $3 million to $4 million as of mid-2026 (gpusmith.com), while an August 2026 analysis of actual purchase orders found a materially higher $5.0 million to buy a rack outright and $5.7 million to fully deploy it, plus $240,000 to $410,000 per year to operate, working out to $2.60 to $2.90 per utilized GPU-hour at 80 to 90 percent utilization (datagravity.dev). Media reporting has cited a still-higher $6 million to $6.5 million per rack (datagravity.dev). The rack integrates 72 Blackwell Ultra GPUs and 36 Grace CPUs in a liquid-cooled, rack-scale architecture (nvidia.com); at roughly $5 million for 72 GPU slots, the implied per-slot price works out to about $69,000, well above the $40,000 to $50,000 reported for a standalone GPU module, a premium attributable to the rest of the machine (Grace CPUs, NVLink switching, liquid cooling, and integration) (datagravity.dev). Cloud rental rates for Blackwell parts reflect this same uncertainty: B200 hourly rates span $3.50 to $27 depending on provider, the widest cross-provider spread tracked for any data center GPU (mercatus-ai.com), and B300 cloud rates were still emerging in a $3.02 to $30 per GPU-hour range as of mid-2026 (gpusmith.com).

NVIDIA says its export and import classifications are provided for informational purposes only, are not a warranty of proper classification, and are subject to change (NVIDIA).

## Cloud and API-Based GPU Pricing: Dollars per GPU-Hour

Renting GPU capacity by the hour, rather than buying hardware outright, is how most organizations actually consume data center GPUs, and it is the segment with the most directly verifiable pricing, since every major provider publishes machine-readable rate cards. *Table 1* below summarizes on-demand rates gathered directly from provider pricing pages and APIs in September 2026.

**Table 1. On-Demand Cloud GPU Rental Rates by Provider, by Quoted Billing Unit (observed September 2026, USD)**

| Provider and billing unit | H100 quoted rate | H200 quoted rate | B200 quoted rate | Source |
|---|---|---|---|---|
| AWS EC2 (8-GPU instance, total/hr) | $55.04 (p5.48xlarge) | $63.30 (p5en.48xlarge) | $113.93 (p6-b200.48xlarge) | pricing.us-east-1.amazonaws.com (pricing.us-east-1.amazonaws.com) |
| Google Cloud (8-GPU instance, total/hr) | $88.49 (a3-highgpu-8g) | $84.81 (a3-ultragpu-8g) | $64.44 flex-start (a4-highgpu-8g) | cloud.google.com (cloud.google.com) |
| Microsoft Azure (8-GPU instance, total/hr) | $98.32 (ND96isr H100 v5) | $110.24 (ND96isr H200 v5) | not published for a mainstream US region at time of writing | prices.azure.com |
| CoreWeave (8-GPU instance, total/hr) | $49.24 | $50.44 | $68.80 | coreweave.com (coreweave.com) |
| Lambda (per GPU) | $3.99 | not listed separately | $6.69 (SXM6) | lambda.ai (lambda.ai) |
| RunPod Secure Cloud (per GPU) | $3.29 | $4.59 | $6.79 | runpod.io (runpod.io) |
| Vast.ai (marketplace median, per GPU) | not separately confirmed | $4.23 | not separately confirmed | vast.ai |
| Voltage Park (per GPU, from) | $1.99 | not published | not published | voltagepark.com |
| Crusoe Cloud (per GPU) | not separately confirmed | $4.29 | contact sales | crusoe.ai |
| Together AI (per GPU, standard/promo through 9/30/26) | $5.49 / $3.99 | not separately confirmed | not separately confirmed | together.ai |

*Table 1* retains providers' quoted billing units: hyperscaler prices are shown per 8-GPU instance, while GPU-cloud specialists quote per single GPU; the provider labels and column headers make that distinction explicit. Dividing the hyperscaler figures by eight for comparison yields roughly $6.88 to $14.24 per GPU-hour for on-demand hyperscaler capacity, well above the $1.99 to $6.79 per GPU-hour charged by neocloud and marketplace providers for the same silicon, a spread that has persisted through 2026 despite repeated price cuts. Discount routes exist within each tier: Azure's H100 spot price runs $18.17 to $18.99 per hour for the full 8-GPU node, over 80 percent below its $98.32 on-demand rate (prices.azure.com), and Google Cloud's spot rate for its Blackwell instance runs roughly $39.63 per hour against a $64.44 flex-start rate. CoreWeave additionally lists an on-demand rate of $42.00 per hour for a GB200 NVL72 allocation (coreweave.com), though the GPU count included in that instance was not specified on the pricing page, underscoring how difficult like-for-like rack-scale comparisons remain.

Independent tracking corroborates the downward trend in blended pricing. SemiAnalysis' GPU Spot-Contract Composite Index shows H100 pricing falling from $6.62 per hour in the second half of 2023 to $2.83 per hour by February 2026 (gpu-index.semianalysis.com), even as the same tracker flagged H100 capacity as "Sold Out" in some listings that same month, evidence that low headline pricing has coexisted with renewed tightness in physical availability (gpu-index.semianalysis.com); the same index put its B200 composite at $3.68 per hour in April 2026 (gpu-index.semianalysis.com). A separate analysis groups the market into three price tiers by late 2025 and 2026: roughly $2 per hour on spot marketplaces, $3.30 on neocloud specialists, and $6.30 on hyperscalers (silicondata.com), a structure this report's own provider-by-provider figures above independently reproduce.

Advertised rate cards are also not the same thing as prices actually paid on multi-year commitments. One secondary-market analysis reports that CoreWeave's 2022-vintage H100 capacity contracts rebooked at 95 percent of their original pricing upon expiry (compute.exchange), meaning a buyer locked into a long-term contract may have seen essentially none of the on-demand price declines documented above. Any comparison of "GPU cloud prices" should specify whether it describes an advertised on-demand rate, a spot price, or a completed multi-year contract, because the three can diverge by an order of magnitude.

## Comparative Context and Market Positioning

NVIDIA remains dominant, but it is not the only option, and the price gap to alternatives is a genuine part of the buying decision. AMD's **Instinct MI300X** (192GB HBM3) is available on a genuinely on-demand basis from DigitalOcean at $2.59 per GPU-hour, the cheapest verified on-demand rate identified for the chip as of mid-2026 (spheron.network); its higher-memory successor, the **MI355X**, had one verified on-demand rate at the time, Oracle OCI's eight-GPU instance at $68.80 per hour, equivalent to $8.60 per GPU-hour (spheron.network). The intermediate **MI325X** carried a median on-demand price of $3.10 per GPU-hour across three tracked providers as of September 5, 2026 (getdeploying.com). These rates sit below most H100 hyperscaler pricing and are broadly competitive with neocloud H100 rates, though buyers should weigh this against AMD's smaller software ecosystem, a qualitative factor outside the scope of a pricing comparison, and against the same billing-mode confusion documented across this market, where preemptible and spot rates are frequently misquoted as on-demand. AMD's own specification page lists the MI300X's typical board power at 750W peak (amd.com), modestly above the H100 SXM's 700W maximum cited earlier.

**Intel's Gaudi 3** occupies a lower price tier still. Intel published list prices for its 8-chip OAM baseboard assembly, inclusive of networking, at $65,000 for Gaudi 2 and $125,000 for Gaudi 3 (servethehome.com), implying roughly $15,625 per accelerator. At its 2024 disclosure, Intel priced the Gaudi 3 processor at approximately $15,650, roughly half the price of an equivalent H100 by Intel's own account (tomshardware.com). Neither figure has been updated by a more recent Intel disclosure at the time of writing, so it should be read as the most recent list-adjacent price rather than a current 2026 market rate.

Custom silicon designed by cloud providers themselves is a third alternative that increasingly substitutes for merchant GPUs. Google Cloud lists its **TPU v6e (Trillium)** at $2.70 per chip-hour in US regions (cloud.google.com), and AWS states that its **Trainium2 (Trn2)** instances deliver 30 to 40 percent better price-performance than its own GPU-based P5e and P5en instances (aws.amazon.com); the latter is a vendor claim rather than an independently verified benchmark and should be treated accordingly. TrendForce projects that ASIC-based AI servers, which include custom accelerators like these, will represent 27.8 percent of all AI server shipments in 2026 (trendforce.com), a meaningful erosion of the merchant-GPU share NVIDIA and AMD have

historically held. Geographic fragmentation compounds this: TrendForce forecasts the combined share of NVIDIA and AMD chips in China's AI server market will fall to 21 percent in 2026, down from 34 percent in 2025, as domestic suppliers Huawei and Cambricon rise to a combined 56 percent share (scmp.com).

*Table 2* below assembles the cross-generation, cross-vendor comparison this report is designed to make reproducible: pull the cited vendor and tracker pages on any date to check whether these figures still hold.

**Table 2. Data Center GPU and AI Accelerator Price Comparison (third-party reported purchase-price estimates where shown; observed 2026, USD)**

| Accelerator | Vendor / Architecture | Memory | Single-Unit Price (approx.) | 8-GPU System Price (approx.) |
|---|---|---|---|---|
| A100 (2020 launch reference) | NVIDIA, Ampere | 40 to 80GB HBM2 | approx. $12,500 at 2020 launch (pcgamer.com); approx. $10,000 by 2023 (cnbc.com) | not separately tracked |
| H100 | NVIDIA, Hopper | 80GB HBM3 | $25,000 to $31,000 (2026) (jarvislabs.ai) | $250,000 to $320,000 |
| H200 | NVIDIA, Hopper (enhanced) | 141GB HBM3e | $30,000 to $55,000 | $320,000 to $420,000 |
| B200 | NVIDIA, Blackwell | 192GB HBM3e | $30,000 to $50,000 | $400,000 to $500,000 |
| B300 (Blackwell Ultra) | NVIDIA, Blackwell Ultra | 288GB HBM3e | approx. $53,000 | $300,000 to $350,000 (DGX B300) |
| Instinct MI300X | AMD, CDNA 3 | 192GB HBM3 | list price undisclosed; cloud on-demand from $2.59/hr | not disclosed by AMD |
| Instinct MI325X | AMD, CDNA 3 | 256GB HBM3e | list price undisclosed; cloud median $3.10/hr | not disclosed by AMD |
| Gaudi 3 | Intel | 128GB HBM2e | approx. $15,625 to $15,650 (2024 list) | $125,000 (8-chip OAM baseboard) |

A recurring theme across every row of *Table 2* is that generation-over-generation prices have not fallen the way consumer GPU prices historically have. Epoch AI's analysis of 470 GPU models released between 2006 and 2021 found that floating-point operations per dollar doubled roughly every 2.5 years across the industry as a whole (epoch.ai), yet the memory-bound economics of the most recent AI-specific generations have partly offset that trend at the sticker-price level, a dynamic examined further in the data section below.

## Data Analysis and Evidence

The financial scale behind these prices is documented in public filings. NVIDIA reported **Data Center segment revenue of $89.0 billion** for its fiscal second quarter of 2027 (the quarter ended July 26, 2026), up 117 percent year over year, within total company revenue of $96.2 billion, up 106 percent (sec.gov) (nvidianews.nvidia.com). NVIDIA guided its subsequent quarter's GAAP and non-GAAP gross margin to 74.0 percent, plus or minus 50 basis points (sec.gov), and one earnings analysis reported the company guiding its following quarter to a margin trough of

71 to 72 percent (fool.com). The stated driver is memory: high bandwidth memory (HBM) now represents roughly 30 to 40 percent of an AI accelerator's build cost, up from under 20 percent two generations earlier (fool.com), and memory suppliers Samsung and SK hynix were reported in December 2025 to be planning a nearly 20 percent HBM3E price increase for 2026, citing NVIDIA H200 and custom-ASIC demand (trendforce.com). Consistent with this, a separate analysis calculated NVIDIA's gross profit on a single DGX H100 server at almost $190,000 (newsletter.semianalysis.com), a margin that helps explain why finished-system prices have stayed high even as underlying silicon costs are debated.

*Table 3* summarizes the market-size and spending figures most relevant to a purchasing decision.

**Table 3. Selected 2026 Market Size and Capital Expenditure Figures**

| Metric | Figure | Period | Source |
|---|---|---|---|
| NVIDIA Data Center segment revenue | $89.0 billion (+117% YoY) | Q2 FY2027 (ended Jul 26, 2026) | sec.gov |
| Global data center GPU market size | $27.98 billion (2026), forecast $226.87 billion by 2035 | 2026 to 2035 | precedenceresearch.com |
| Combined capex, top 9 cloud service providers | over $886.7 billion | 2026 | trendforce.com |
| Combined capex, top 9 cloud service providers (forecast) | approx. $1.3 trillion (+50% YoY) | 2027 | trendforce.com |
| Microsoft capital expenditures | $35.8 billion (quarter); $115.9 billion (full year) | FQ4 2026 / FY2026 | sec.gov |
| Meta capital expenditure guidance | $130 to $145 billion | full-year 2026 | prnewswire.com |

The capex figures in *Table 3* matter directly to GPU buyers because they represent the demand pressure sitting behind every price above: TrendForce separately raised its 2026 global AI server shipment growth forecast to nearly 31 percent year over year, up from an earlier 28 percent estimate, attributing the revision to roughly 90 percent growth in the capital spending of the largest cloud providers (trendforce.com) and Microsoft reported fiscal fourth-quarter 2026 revenue of $90.0 billion, up 18 percent, alongside its capital spending (sec.gov). Power is a related, underappreciated cost driver at the system level: an H100 SXM draws up to 700W (nvidia.com), an MI300X draws 750W at peak (amd.com), and a full 8-GPU DGX B200 system draws approximately 14.3 kilowatts (nvidia.com); at typical US commercial electricity rates, that single system's power draw alone can add tens of thousands of dollars per year to its total cost of ownership, a figure separate from and additive to every purchase or rental price cited above. The Export Administration Regulations provide rules for determining whether items and activities are subject to the EAR (BIS); consult qualified trade-compliance professionals before making procurement or export decisions.

## Case Studies and Real-World Examples

### GB300 NVL72 rack economics: three measurement methods, three numbers.

The rack-scale examples above are worth revisiting as a single worked case, because they show concretely why "the price of a GB300 rack" cannot be answered with one number. An analyst estimate places a full rack at $3 million to $4 million (gpusmith.com); an analysis built from actual purchase orders puts the same product at $5.0 million to buy and $5.7 million to deploy, with an operating cost of $240,000 to $410,000 per year, yielding a computed $2.60 to $2.90 per utilized GPU-hour at realistic utilization (datagravity.dev); and media reporting has cited a still-higher $6 million to $6.5 million (datagravity.dev). The lesson for a buyer or forecaster is procedural: ask whether a quoted rack price reflects an analyst's cost model, a documented transaction, or unverified press reporting, since the three methods here produced a more than 60 percent spread on the same product in the same year.

### Contract-based residual value versus hardware resale value.

A second, related example concerns what happens to a GPU's price after its initial sale. One report found that CoreWeave's 2022-vintage H100 capacity contracts rebooked at 95 percent of their original pricing upon expiry, indicating that customers holding multi-year capacity contracts saw almost no benefit from the broader market's price declines (compute.exchange). By contrast, the same 2026 survey found used, non-refurbished H100 hardware trading 40 percent or more below new pricing, and refurbished units roughly 15 to 20 percent below new (compute.exchange), and a separate source described secondary-market SXM units trading as low as $6,000 to $15,000 (thundercompute.com). Capacity contracts and physical hardware can therefore retain value on completely different trajectories for the same underlying chip.

## Implications and Future Directions

Several structural forces point toward continued price complexity rather than convergence on a single number. Memory remains the tightest constraint: with HBM now consuming 30 to 40 percent of an AI accelerator's bill of materials (fool.com) and suppliers signaling further HBM3E price increases into 2026 (trendforce.com), buyers should expect hardware purchase prices to remain sticky or rise even as cloud rental rates compress under competitive pressure. Competitive dynamics are shifting on two fronts simultaneously: TrendForce's forecast that ASIC-based servers will reach 27.8 percent of 2026 AI server shipments (trendforce.com) suggests merchant GPU pricing power will face growing pressure from cloud providers' own silicon, while the projected drop in NVIDIA and AMD's combined China market share, from 34 percent in 2025 to 21 percent in 2026 (scmp.com), points toward an increasingly bifurcated global market with distinct China and rest-of-world price curves.

For organizations evaluating a purchase or a multi-year cloud commitment, the practical decision rarely reduces to "buy versus rent" in the abstract. It depends on utilization: the GB300 NVL72 case study above shows a computed cost of $2.60 to $2.90 per utilized GPU-hour at 80 to 90 percent utilization, a rate that only beats on-demand cloud pricing if that utilization level is actually sustained. Enterprises outside the AI infrastructure industry, including regulated sectors such as life sciences, are increasingly running this kind of buy-versus-rent, on-demand-versus-reserved analysis as part of broader technology investment decisions rather than treating GPU acquisition as a purely technical procurement question. IntuitionLabs, a life sciences and AI consultancy, structures exactly this kind

of decision as a technology assessment exercise: evaluating an organization's current technology stack and making recommendations for optimization before capital is committed (intuitionlabs.ai), a discipline that applies to compute infrastructure decisions as much as to any other enterprise technology investment. Firms in this position should weight the case-study lesson above heavily: a quoted price is only informative once its measurement method, its utilization assumption, and its commitment length are all known.

## Frequently Asked Questions (FAQs)

### How much does a data center GPU cost in 2026?

It depends heavily on the generation and form factor. Cited third-party pricing guides estimate a single NVIDIA H100 at roughly $25,000 to $31,000, an H200 at $30,000 to $55,000, and a B200 at $30,000 to $50,000 (jarvislabs.ai) (mercatus-ai.com). Full 8-GPU systems run from roughly $250,000 (H100) to $500,000 (B200) (cloudzero.com).

### What was the reported H100 GPU price in August 2026?

As of August 2026, a new 80GB H100 card was priced around $31,000, with an 8-GPU HGX server running $250,000 to $320,000 (cloudzero.com). Cloud rental of the same chip ranges from $1.99 per hour on the cheapest specialist provider to over $12 per GPU-hour equivalent on a hyperscaler (voltagepark.com).

### How is enterprise GPU pricing structured in 2026?

Enterprises typically encounter three distinct price layers: outright hardware purchase, integrated rack-scale systems such as the GB200 and GB300 NVL72, and hourly cloud rental, each with its own discount structure (volume discounts on hardware, spot and reserved discounts in the cloud). See the cloud pricing table above for a provider-by-provider breakdown.

### How has AI GPU pricing changed historically?

An A100 cost approximately $12,500 at its 2020 launch (pcgamer.com) and was described as a roughly $10,000 chip by 2023 (cnbc.com). The H100 that succeeded it resold on eBay for $39,995 to nearly $46,000 in April 2023 amid acute shortages (cnbc.com), before settling into the $25,000 to $40,000 range by 2026 (compute.exchange).

### How does a GPU cluster get priced?

A GB300 NVL72 rack, containing 72 GPUs and 36 Grace CPUs, has been priced anywhere from $3 million to $6.5 million depending on whether the figure comes from an analyst cost model, an actual purchase order, or press reporting (gpusmith.com) (datagravity.dev); annual operating cost on top of the hardware runs $240,000 to $410,000.

### Are AMD and Intel accelerators cheaper than NVIDIA?

On a cloud dollar-per-hour basis, yes: AMD's MI300X has been available on-demand from DigitalOcean at $2.59 per GPU-hour (spheron.network), and Intel's Gaudi 3 was priced at roughly half of an equivalent H100 at its 2024 disclosure (tomshardware.com). Neither vendor publicly discloses standardized list prices for its data center accelerators, so these figures come from cloud marketplaces and OEM disclosures rather than official price sheets.

## Conclusion

Data center GPU prices in 2026 resist a single headline number, and that is the central, reproducible finding of this report rather than a caveat to it. A buyer researching "the price of an H100" will find defensible figures anywhere from under $2 per GPU-hour on a spot marketplace to over $14 per GPU-hour on a hyperscaler, and anywhere from $25,000 to over $40,000 to buy the same chip outright, depending entirely on form factor, volume, contract length, and how recently the source was updated. The same holds at every layer above: single GPUs, multi-GPU servers, and full racks each carry a documented spread that widens as systems get larger and more integrated, culminating in the GB300 NVL72 rack's more than 60 percent spread across three legitimate measurement methods.

What is comparatively stable is the direction of the underlying forces. Memory, not logic, is now the dominant cost driver and the most likely source of further price pressure, with HBM's share of accelerator bill of materials having roughly doubled in two generations. Competitive pressure is real but uneven: cloud rental rates have fallen sharply and repeatedly, hardware purchase prices have proven far stickier, and alternative accelerators from AMD, Intel, and cloud providers' own silicon are gaining share without yet displacing NVIDIA's position at the top of the market. Buyers, whether purchasing outright or committing to cloud capacity, are best served by treating every quoted price as a dated observation tied to a specific measurement method, exactly the discipline this report has tried to model throughout, and by revisiting the primary sources cited here directly before making a purchasing decision, since a figure that is accurate in September 2026 should not be assumed to hold a year later.

---