NVIDIA Data Center GPU Pricing: H100 to GB200 Cost Guide

Published July 10, 2026 33 min read

Executive Summary

NVIDIA's data center GPU lineup spans roughly an order of magnitude in price, and the gap keeps widening as each new architecture adds more silicon, more high bandwidth memory (HBM), and more interconnect. As of July 2026, a single **NVIDIA H100** (Hopper architecture, 80GB HBM3) sells for approximately $25,000 to $40,000 depending on the PCIe or SXM5 variant and the reseller (Source: mercatus-ai.com) (Source: cloudzero.com). The newer **H200** (141GB HBM3e) occupies a similar $30,000 to $40,000 purchase band but commands materially higher cloud rental rates because of its larger memory pool (Source: jarvislabs.ai). The Blackwell-generation **B200** carries an estimated street price near $40,000 to $45,000 per GPU, with an 8-GPU HGX B200 server system quoted around $390,000, and a complete DGX B200 appliance listed at roughly $500,000 (Source: itctshop.com) (Source: blogs.novita.ai). At the top of the stack, the rack-scale **GB200 NVL72**, which links 72 Blackwell GPUs and 36 Grace CPUs into a single NVLink domain, carries a system price of roughly $2 million to $3 million per rack, with some quotes running as high as $3.4 million (Source: gpuaas.com) (Source: aiappstacker.com), while the newer **GB300 NVL72** ("Blackwell Ultra") is estimated at $3.7 million to $4 million per rack (Source: glennklockwood.com).

Cloud rental is the more accessible entry point for most buyers, and rates vary enormously by provider and commitment length. On-demand **H100** GPU-hours range from roughly $1.38 to $11.06 across tracked providers, with a market median near $2.29 per GPU-hour, while specialist clouds such as Lambda price 256-GPU H100 reservations at $5.54 per hour and single on-demand instances near $3.99 to $4.29 per hour (Source: aitooldiscovery.com) (Source: lambda.ai). **H200** rentals run from about $1.45 to $13.78 per hour, with AWS pricing its p5e H200 instance at $7.91 per GPU-hour (Source: gmicloud.ai) (Source: thundercompute.com). **B200** capacity, still supply-constrained as of mid-2026, rents for approximately $3.75 to $9.86 per GPU-hour depending on provider and reservation length (Source: gpucost.org) (Source: lambda.ai), and full **GB200 NVL72** racks rent at an estimated $10.50 to $27.00 per GPU-hour, or roughly $756 to $1,944 per hour for the entire 72-GPU rack (Source: aiappstacker.com).

These prices sit inside a fast-growing revenue base. NVIDIA reported record **Data Center revenue of $75.2 billion** for its first fiscal quarter of 2027 (the three months ended April 26, 2026), up 92% year over year, within total company revenue of $81.6 billion (Source: nvidianews.nvidia.com). Independent analysts estimate NVIDIA holds roughly 75% to 90% of the AI accelerator market by revenue, though that share is projected to compress

somewhat as AMD's MI300X and MI325X, and hyperscaler-designed custom silicon, scale up (Source: siliconanalysts.com). Pricing power stems partly from a structural bottleneck: TSMC's CoWoS advanced-packaging capacity, which bonds GPU logic dies to HBM stacks, remains sold out into 2026 and 2027, with NVIDIA alone estimated to hold around 60% of that capacity (Source: siliconanalysts.com).

This report walks through purchase pricing, cloud rental pricing, and total-cost-of-ownership economics for the H100, H200, B200, GB200, GB300, and related SKUs; compares them against AMD's competing accelerators; reviews the used-GPU secondary market; and examines real-world deployments from xAI's Colossus cluster to Meta's Llama training infrastructure and the Stargate data center program, to give buyers a grounded answer to what NVIDIA data center GPU pricing actually looks like today.

Introduction and Background

NVIDIA does not publish an official price list for its data center GPUs. Unlike consumer GeForce cards, which carry a manufacturer's suggested retail price, chips such as the **H100**, **H200**, **B200**, and rack-scale systems like the **GB200 NVL72** are sold through original equipment manufacturers (OEMs), system integrators, and cloud partners under negotiated terms that vary by volume, region, and bundled services (Source: mercatus-ai.com). That opacity is why every buyer researching this topic runs into a wide range rather than a single number, and why this report treats price as a range anchored to verifiable market data rather than a single sticker figure.

Segment matters as much as architecture generation when interpreting any single price quote. A hyperscaler negotiating tens of thousands of GPUs pays a fundamentally different price than a startup renting eight GPUs for a weekend fine-tuning run, and neither figure is "the" NVIDIA GPU price; both are legitimate data points inside a wide, opaque distribution. This report therefore presents ranges throughout, sourced from multiple independent trackers, OEM price sheets, and cloud provider pricing pages, rather than collapsing the market into a single misleading average.

The pricing story is inseparable from the underlying architecture generations. **Hopper**, launched in March 2023, produced the H100 and its memory-upgraded successor, the H200. **Blackwell**, which began shipping at volume through 2025 and into 2026, produced the B200, the GB200 Grace Blackwell Superchip, and the rack-scale GB200 NVL72. **Blackwell Ultra** followed with the B300 and the GB300 NVL72, aimed at inference-heavy, test-time-scaling workloads, detailed further below. Each generation adds substantially more HBM capacity and bandwidth: the H100 SXM5 ships with 80GB of HBM3 at 3.35 terabytes per second (TB/s), the H200 doubles usable memory to 141GB of HBM3e, and the B200 pushes to 180GB to 192GB of HBM3e at up to 8 TB/s (Source: jarvislabs.ai) (Source: runpod.io).

Table 2 below lays out the core specifications and pricing anchors for the four GPU families most buyers evaluate today, so that performance and cost can be compared side by side rather than in isolation.

SPEC	H100 SXM5	H200 SXM	B200 SXM6	GB200 NVL72 (PER GPU, RACK BASIS)
Architecture	Hopper	Hopper	Blackwell	Blackwell (rack-scale)
HBM memory	80GB HBM3 (Source: jarvislabs.ai)	141GB HBM3e (Source: aws.amazon.com)	180 to 192GB HBM3e (Source: runpod.io)	13.4TB aggregate across 72 GPUs (see text below)
Memory bandwidth	3.35 TB/s(Source: jarvislabs.ai)	4.8 TB/s (Source: aws.amazon.com)	8 TB/s (see row above)	576 TB/s aggregate (Source: nvidia.com)
TDP	700W (Source: jarvislabs.ai)	Not separately disclosed on AWS spec sheet	1,000W (Source: gpucost.org)	120 to 130 kW per 72-GPU rack (Source: gpuaas.com)
New purchase price	$25,000 to $40,000 (Source: cloudzero.com)	$30,000 to $40,000 (Source: cerebrium.ai)	$40,000 to $45,000 (Source: siliconanalysts.com)	$2M to $3M for full rack, roughly $28,000 to $42,000 per GPU equivalent (Source: gpuaas.com)
Cloud on-demand rate	$1.38 to $11.06/hr, median $2.29 (Source: aitooldiscovery.com)	$1.45 to $13.78/hr, median $3.95 (Source: gmicloud.ai)	$3.75 to $9.86/hr (Source: gpucost.org)	$10.50 to $27.00 per GPU/hr (Source: aiappstacker.com)

Table 2 makes clear that the jump from H100/H200 to B200 is a moderate step in both memory and price, while the jump to a full GB200 NVL72 rack is a categorical change in unit of purchase: buyers are no longer pricing a chip, but an integrated exascale-class system with its own power and cooling requirements. The doubling of memory bandwidth from Hopper to Blackwell (3.35 TB/s to 8 TB/s at the single-GPU level) is one of the primary reasons Blackwell commands a price premium even though its per-GPU dollar price is not dramatically higher than the H200's, since memory bandwidth, not raw compute, is often the binding constraint on large language model inference throughput.

Three acquisition paths exist for any of these chips: outright purchase (capital expenditure, typically through an OEM such as Dell, Supermicro, or HPE, or direct from a cloud provider's hardware partner), specialist GPU cloud rental (billed hourly or monthly, from providers such as CoreWeave, Lambda, and dozens of smaller "neoclouds"), and hyperscaler cloud rental (AWS, Microsoft Azure, and Google Cloud, typically at a premium over specialist clouds but with broader ecosystem integration) (Source: cloudpricecalculators.com). Which path makes financial sense depends heavily on utilization, a trade-off examined in detail later in this report.

Demand has stayed intense enough that pricing has not followed the steep, predictable decline typical of most computer hardware. Analysts at Silicon Data, tracking thousands of retail and resale listings, found that H100 sticker prices held remarkably flat in the $25,000 to $40,000 band from mid-2024 into early 2026 even as newer architectures shipped, a stability that masked much sharper swings in the secondary market underneath (Source: silicondata.com). The remainder of this report breaks pricing down by product family, then examines the market forces, real-world deployments, and forward outlook that explain why NVIDIA GPU economics look the way they do in mid-2026.

H100 and H200 (Hopper Generation) Pricing

The **H100** remains the most widely deployed data center GPU in production AI infrastructure as of July 2026, even with two newer architectures on the market, because so much existing training and inference capacity was built around it. New H100 SXM5 units list for approximately **$25,000 to $30,000** from major OEMs, though broader market surveys that include PCIe variants and various resellers put the full range at **$25,000 to $40,000** (Source: mercatus-ai.com) (Source: cloudzero.com). A fully populated 8-GPU HGX H100 server, the standard building block for training clusters, runs approximately **$250,000 to $320,000**, and some system integrators price optimized 8U configurations near $300,000 (Source: mercatus-ai.com) (Source: itctshop.com). A single NVIDIA H100 NVL card, an alternative packaging aimed at inference workloads, has been priced by resellers at roughly $30,500 to $33,000 (Source: itctshop.com).

Cloud rental gives most buyers a lower barrier to entry. Aggregated market data across 49 tracked H100 configurations puts the on-demand median at approximately **$2.29 per GPU-hour**, with the full observed range spanning **$1.38 to $11.06** depending on provider tier, region, and instance configuration (Source: aitooldiscovery.com). Specialist clouds tend to undercut hyperscalers substantially: **Lambda** prices single on-demand H100 SXM instances at $3.99 per GPU-hour and reserved 256-GPU clusters (2 weeks to 1 year commitments) as low as $5.54 per GPU-hour, while its smaller on-demand configurations range from $3.29 to $4.29 depending on GPU interconnect (Source: lambda.ai). By contrast, **AWS** bills its p5.48xlarge instance, containing 8 H100 GPUs with 640GB of aggregate HBM3, at a rate that industry trackers have historically placed around $3.93 to $4 per GPU-hour on-demand, and AWS documents the underlying hardware as supporting EC2 UltraClusters that scale to 20,000 H100 or H200 GPUs interconnected at up to 3,200 gigabits per second via Elastic Fabric Adapter (EFA) networking (Source: aws.amazon.com). Microsoft's Azure ND H100 v5 instance, an 8-GPU node, lists at approximately $98.32 per hour on-demand, or roughly $12.29 per GPU-hour, well above specialist-cloud rates, reflecting the premium hyperscalers typically charge for integrated enterprise tooling and compliance certifications (Source: instances.vantage.sh).

The **H200**, which NVIDIA differentiates from the H100 primarily through memory (141GB of HBM3e versus 80GB of HBM3, and 4.8 TB/s of bandwidth versus 3.35 TB/s), occupies a similar purchase price band of **$30,000 to $40,000** per GPU, but its cloud rental rates run consistently higher than the H100's because inference workloads serving today's largest models benefit disproportionately from the extra memory (Source: jarvislabs.ai) (Source: cerebrium.ai). Market-wide H200 on-demand pricing spans **$1.45 to $13.78 per GPU-hour**, with a median near $3.95 (Source: gmicloud.ai). AWS documents its p5e and p5en instance families as pairing 8 H200 GPUs with up to 1,128GB of aggregate HBM3e memory, and independent trackers put AWS's per-GPU H200 rate at approximately **$7.91 per hour**, or **$63.28 per hour** for the full 8-GPU node (Source: aws.amazon.com) (Source: thundercompute.com). Specialist providers again undercut hyperscalers considerably: RunPod advertises H200 rentals starting from $4.39 per hour, and Vast.ai's marketplace lists rates as low as $3.41 per hour (Source: runpod.io) (Source: vast.ai). A complete DGX system with eight H200 GPUs has been quoted at roughly $400,000 to $500,000, reflecting NVIDIA's premium fully-integrated appliance positioning versus commodity OEM builds (Source: intuitionlabs.ai).

The persistent price gap between the H100 and H200 on cloud platforms, despite similar purchase prices, is one of the clearer signals in the entire market: buyers are effectively paying a premium of roughly 50% to 70% per hour for the H200's extra memory headroom, which matters most for serving large mixture-of-experts models and long-context inference workloads where the model's key-value cache no longer fits comfortably in 80GB.

For training workloads that are more compute-bound than memory-bound, the price premium is harder to justify, which is part of why large H100 fleets, such as those run by Meta and xAI, remain in heavy production use well after the H200 and Blackwell generations became available.

B200, GB200, and Blackwell-Generation Pricing

NVIDIA's **Blackwell** architecture, the successor to Hopper, brought a step-change in both performance and price. The flagship **B200** is a dual-die design carrying 208 billion transistors and up to 192GB of HBM3e at roughly 8 TB/s of bandwidth, and industry estimates put its manufacturing cost (bill of materials plus packaging) at approximately **$6,400**, nearly double the H100's estimated **$3,320** cost, driven mainly by HBM, which now represents an estimated 45% of total cost of goods sold (Source: siliconanalysts.com). At an estimated street price near **$40,000 to $45,000** per GPU, that manufacturing cost implies an estimated gross margin of roughly 84%, only slightly below the H100's estimated 88% (Source: siliconanalysts.com).

System-level pricing scales accordingly. An 8-GPU HGX B200 platform has been listed by system integrators at approximately **$390,000**, and a complete DGX B200 appliance, NVIDIA's fully integrated turnkey system, carries a list price of roughly **$500,000** (Source: itctshop.com) (Source: blogs.novita.ai). Cloud rental of B200 capacity remains comparatively expensive, both because supply is still constrained relative to demand and because the hardware is newer: Lambda's reserved B200 clusters run **$8.87 to $9.86 per GPU-hour** depending on scale and commitment, its on-demand rate reaches **$6.69 to $6.99 per GPU-hour**, and RunPod's on-demand B200 instances start at **$4.99 per hour** (Source: lambda.ai) (Source: runpod.io). On Google Cloud, the A4 instance family (a4-highgpu-8g), built around 8 B200 GPUs, lists at approximately $34.24 per hour on-demand in the us-central1 region, which works out to roughly **$4.28 per GPU-hour**, about 3.3 times cheaper than AWS's comparable p6.48xlarge B200 instance at approximately $14.24 per GPU-hour (Source: spheron.network) (Source: spheron.network).

The **GB300**, marketed as "Blackwell Ultra," raises memory further to 288GB of HBM3e per GPU and is aimed specifically at inference-heavy, test-time-scaling reasoning workloads. An 8-GPU DGX B300 system is anchored at an estimated **$300,000 to $350,000**, working out to roughly $37,500 to $43,750 per GPU at the system level, while cloud rental of standalone B300 GPUs spans a wide **$1.88 to $18.00 per GPU-hour**, with Vast.ai's marketplace listing rates around $6.25 per hour and Spheron advertising rates from $3.50 per hour (Source: tech-insider.org) (Source: vast.ai).

The most consequential pricing shift in the Blackwell generation, however, is that NVIDIA increasingly sells its top-tier compute as a **rack-scale system rather than a discrete chip**. The **GB200 NVL72** connects 36 Grace CPUs and 72 Blackwell GPUs into a single NVLink domain delivering 130 TB/s of GPU-to-GPU bandwidth, 13.4 terabytes of aggregate HBM3e (see *Table 2* above), and up to 1,440 petaflops of FP4 tensor performance (Source: nvidia.com). NVIDIA states the system "acts as a single, massive GPU" and delivers claimed performance of 30 times faster real-time trillion-parameter LLM inference and 4 times faster training compared with an equivalent-scale H100 cluster, alongside 25 times better energy efficiency (Source: nvidia.com) (Source: nvidia.com). A full rack costs an estimated **$2 million to $3 million**, with some enterprise-buyer guides citing a range as wide as $2.8 million to $3.4 million once networking, cooling infrastructure, and integration are included (Source: gpuaas.com) (Source: aiappstacker.com). Cloud access to GB200 capacity is priced per GPU at an estimated **$10.50 to $27.00 per hour**, translating to roughly $756 to $1,944 per hour for a fully rented 72-GPU rack, though reserved, longer-term commitments push that figure down substantially (Source: aiappstacker.com) (Source: spheron.network). Power draw is a major operational constraint: the rack requires liquid cooling and consumes an estimated 120 to 130 kilowatts, and the assembled chassis weighs approximately 1.36 metric tons, heavy enough that it does not fit through standard data center doors (Source: gpuaas.com).

The successor **GB300 NVL72** integrates 72 Blackwell Ultra GPUs and 36 Grace CPUs, is purpose-built for test-time-scaling inference and reasoning workloads, and NVIDIA claims it delivers up to a 50 times increase in overall AI factory output performance compared with Hopper-based platforms (Source: nvidia.com). A single GB300 NVL72 rack is estimated at **$3.7 million to $4 million**, roughly 25% to 45% above the GB200 NVL72's price, reflecting the added Blackwell Ultra silicon and HBM3e capacity (Source: glennklockwood.com). Looking one generation further ahead, Tom's Hardware reports that pricing for NVIDIA's next platform, Vera Rubin NVL72, has been estimated by industry sources at up to **$8.8 million per rack**, illustrating how quickly rack-scale system prices are climbing even as per-GPU prices rise more modestly (Source: tomshardware.com).

Cloud and Hyperscaler Pricing Across Providers

Because NVIDIA does not sell most enterprise buyers hardware directly, cloud pricing is where the majority of GPU spending actually happens, and rates diverge sharply between the three main channels: hyperscalers (AWS, Azure, Google Cloud), specialist GPU clouds (CoreWeave, Lambda, RunPod), and GPU marketplaces that aggregate smaller data center operators (Vast.ai, Spheron).

Table 1 below summarizes representative on-demand hourly rates for the H100, H200, and B200 across several provider types, drawn from pricing pages and market trackers accessed in July 2026.

GPU	HYPERSCALER RATE (PER GPU-HR)	SPECIALIST CLOUD RATE (PER GPU-HR)	MARKETPLACE/LOW END (PER GPU-HR)	MARKET MEDIAN
H100	Azure ND H100 v5: ~$12.29 ($98.32 for 8-GPU node) (Source: instances.vantage.sh)	Lambda: $3.99 to $4.29 on-demand (see text above)	$1.38 low end across tracked providers (Source: aitooldiscovery.com)	~$2.29/hr (Source: aitooldiscovery.com)
H200	AWS p5e: $7.91 per GPU ($63.28 for 8-GPU node) (Source: thundercompute.com)	RunPod: from $4.39 (Source: runpod.io)	Vast.ai: $3.41 (Source: vast.ai)	~$3.95/hr (Source: gmicloud.ai)
B200	Google Cloud A4: ~$4.28 per GPU (Source: spheron.network)	Lambda: $6.69 to $6.99 on-demand, $8.87 to $9.86 reserved (Source: lambda.ai)	RunPod: from $4.99 (Source: runpod.io)	$3.75 to $9.86 range across tracked providers (Source: gpucost.org)

The spread visible in *Table 1* illustrates a consistent pattern across all three GPU generations: hyperscaler pricing runs 1.5 to 3 times higher than specialist-cloud pricing for equivalent hardware, largely because hyperscalers bundle in broader compliance certifications, integrated storage and networking ecosystems, and enterprise support that many specialist clouds do not offer. AWS documents that its P5 family is deployed inside EC2 UltraClusters using petabit-scale, nonblocking networking, and it markets the instances as reducing model-training costs by up to 40% versus previous-generation hardware, which is a claim about generational efficiency gains rather than about matching specialist-cloud sticker prices (Source: aws.amazon.com). Reserved and longer-term commitments narrow the gap considerably: Lambda's pricing tiers show H100 rates falling from $6.16 per hour at 16 GPUs to $5.54 per hour at 256-plus GPUs on a multi-week-to-annual commitment, and B200 rates falling similarly from $9.86 to $8.87 across the same scale tiers, as detailed in *Table 1* above.

NVIDIA's own **DGX Cloud** offering, delivered through partners including Oracle, Azure, and Google Cloud, bills by instance-month rather than by the hour; an 8-GPU H100 node is commonly quoted around $36,000 to $40,000 per month on an annual commitment, which annualizes to roughly $5.10 to $5.71 per GPU-hour, positioned between specialist-cloud and hyperscaler on-demand rates (Source: cloudpricecalculators.com).

Comparative Context: Buy vs. Rent, AMD Alternatives, and Market Positioning

Whether purchasing or renting NVIDIA GPUs makes better financial sense depends on the utilization rate a buyer can sustain, and hardware sellers publish breakeven models to make that trade-off concrete. One buy-vs-rent model for the B200 puts the breakeven point at approximately **12,000 hours** of usage, or about 500 days (16.7 months) of continuous 24/7 operation, before purchasing the $45,000 hardware outperforms paying $3.75 per hour to rent equivalent capacity (Source: gpucost.org). A separate H100 total-cost-of-ownership analysis observes that the effective per-hour cost of an owned H100, once electricity, cooling, networking, and depreciation are included, ranges from **$0.80 to $4.20 per hour depending on the acquisition path chosen**, a fivefold spread for what is physically the same chip (Source: mercatus-ai.com).

AMD is the most credible merchant alternative to NVIDIA in the data center GPU market, primarily through its MI300X and MI325X accelerators. Independent analysis places MI300X and MI325X systems at **roughly 30% to 50% cheaper** than equivalent H100 configurations, with cloud pricing spanning $1.50 to $6.98 per hour for MI300X versus $1.99 to $12.29 per hour for comparable H100 instances (Source: siliconanalysts.com). Independent inference benchmarking from SemiAnalysis similarly found that AMD's MI300X and MI325X "generally present lower total hourly costs compared to NVIDIA's H100 and H200 GPUs" across most latency and model configurations tested (Source: newsletter.semianalysis.com). Despite the price advantage, AMD's estimated share of the AI accelerator market remains in the mid-single digits to roughly 8% by revenue, reflecting the durability of NVIDIA's CUDA software ecosystem and NVIDIA's outsized share of TSMC's CoWoS advanced-packaging capacity, an estimated 60% versus roughly 11% for AMD (Source: siliconanalysts.com).

Power and colocation costs are a frequently underestimated component of NVIDIA GPU total cost of ownership. A 100-GPU H100 cluster requires roughly **145 kilowatts (kW)** of contracted power capacity, and colocation pricing for that capacity ranges from approximately $80 per kW per month at wholesale rates to $300 per kW per month in premium metro facilities such as New York, a 3.75-times variance for functionally identical infrastructure that flows directly into the effective per-GPU-hour cost of compute (Source: mercatus-ai.com). At the full cluster level, one detailed capital-allocation model puts the three-year, all-in total cost of owning 100 H100 GPUs, including hardware, power, colocation, networking, and operations, at **$5 million to $7 million**, with the $2 million spread driven mainly by achievable utilization, regional power costs, and how quickly Blackwell availability accelerates H100 depreciation (Source: mercatus-ai.com). At data-center scale, independent research group Epoch AI estimates a typical one-

gigawatt AI data center requires **$38 billion** in upfront capital expenditure and **$0.9 billion** in annual operating expenses, with servers, meaning GPUs and associated compute hardware, accounting for roughly 60% of total annualized cost and energy representing the largest single operating-expense category at an estimated $0.6 billion per year (Source: epoch.ai) (Source: epoch.ai). These figures reinforce that the price of the GPU itself, while the largest single line item, typically represents well under two-thirds of what a buyer actually spends to put NVIDIA compute into production.

Independent analysts estimate NVIDIA's overall share of the AI accelerator market by revenue at approximately **80% to 90%** as of 2024 to 2025, with share exceeding 90% in training workloads specifically and running lower, roughly 60% to 75%, in inference, where custom silicon from hyperscalers competes more directly (Source: siliconanalysts.com). That share is projected to compress modestly, to around 75%, by the end of 2026 as AMD and custom silicon from Google, Amazon, and Microsoft scale, even as NVIDIA's absolute dollar revenue keeps growing because the overall addressable market is expanding faster than any single competitor can capture share from it. One prominent example of that custom-silicon competition: Anthropic and Amazon announced an expanded agreement in April 2026 under which Anthropic committed more than $100 billion over ten years to AWS technologies, securing up to 5 gigawatts of capacity built substantially on Amazon's own Trainium2 through Trainium4 chips rather than NVIDIA GPUs, with Anthropic stating it already uses "over one million Trainium2 chips" to train and serve its Claude models (Source: anthropic.com).

Data Analysis and Evidence

NVIDIA's financial results provide the clearest quantitative anchor for how large the data center GPU market has become and how fast it is still growing. For the first quarter of fiscal year 2027 (ended April 26, 2026), NVIDIA reported total company revenue of **$81.6 billion**, up 85% from a year earlier, of which **Data Center revenue reached $75.2 billion**, up 92% year over year and up 21% sequentially (Source: nvidianews.nvidia.com). Within that figure, Data Center compute revenue (the GPU silicon itself) was $60.4 billion, up 77% year over year, while Data Center networking revenue, covering NVLink, InfiniBand, and Spectrum-X Ethernet products that connect GPUs together, reached $14.8 billion, up 199% year over year, indicating that interconnect is becoming an increasingly large share of what buyers actually pay for (Source: nvidianews.nvidia.com). GAAP gross margin for the quarter was **74.9%**, down from historical highs in the high-70s to 80s reported in prior periods, a compression analysts attribute partly to a mix shift toward lower-margin networking and system-level products alongside the core GPUs (Source: nvidianews.nvidia.com). NVIDIA's guidance for the following quarter targeted revenue of $91.0 billion, plus or minus 2%, explicitly assuming zero Data Center compute revenue from China, underscoring how completely export restrictions have removed that market from near-term forecasts, a topic examined further in the Implications section below.

Chip-level manufacturing cost estimates help explain why NVIDIA can sustain such high gross margins even as component costs rise. One cost-modeling analysis estimates the H100 SXM costs approximately **$3,320** to manufacture and sells for around $28,000, an implied 88% gross margin, while the B200 costs an estimated **$6,400** to manufacture, nearly double, and sells for around $40,000, an implied 84% margin (Source: siliconanalysts.com) (Source: siliconanalysts.com). HBM memory is the single largest driver of that cost increase: it made up an estimated 41% of the H100's bill of materials and rose to roughly 45% on the B200, at an estimated $2,900 per unit, confirming that memory economics, not logic-die cost, increasingly set the floor for AI accelerator pricing (Source: siliconanalysts.com).

Supply-side constraints reinforce the pricing floor. TSMC's advanced chip-on-wafer-on-substrate (CoWoS) packaging lines, the process step that bonds GPU logic dies to HBM stacks, are described by industry analysts as **fully booked** at 52-to-78-week lead times as of mid-2026, against total 2026 demand estimated near **1.0 million wafers**, up from roughly 370,000 wafers in 2024 and 670,000 in 2025 (Source: siliconanalysts.com). NVIDIA alone is estimated to hold roughly 60% of that capacity, approximately 595,000 wafers, and has reportedly pre-booked more than half of TSMC's planned 2026-to-2027 CoWoS expansion, meaning newly added packaging capacity is largely spoken for before it comes online (Source: siliconanalysts.com). Behind CoWoS, HBM supply from SK Hynix, Samsung, and Micron is described by one supply-chain analysis as the true binding constraint on Blackwell shipments, since every B200 requires 8 stacks of HBM3e, every GB200 superchip requires 16 stacks, and every B300 requires 12 stacks of HBM3e or HBM4 (Source: quantabundancia.com).

On the demand side, CoreWeave, a specialist GPU cloud provider and one of the largest buyers of NVIDIA hardware, reported first-quarter 2026 revenue of **$2.078 billion**, more than double the $982 million reported a year earlier, alongside a revenue backlog reaching nearly **$100 billion** ($99.4 billion) as of March 31, 2026, and closed an $8.5 billion delayed-draw term loan facility to help fund further GPU purchases, alongside a $2 billion direct equity investment from NVIDIA itself (Source: sec.gov) (Source: sec.gov). China-related export restrictions have also materially affected NVIDIA's realized revenue: the company recorded a **$5.5 billion** charge in its fiscal 2026 first quarter tied to H20 chips it could no longer ship to China after new U.S. rules, and NVIDIA's chief financial officer told investors as recently as February 2026 that despite the Trump administration later approving limited H200 sales to China subject to a 25% revenue cut to the U.S. government, the company had "yet to generate any revenue" from those approved shipments (Source: cnbc.com) (Source: cnbc.com).

The used and refurbished GPU secondary market has become large enough to shape pricing at the margin. H100 units that sold for as much as $50,000 in mid-2024 during peak scarcity now trade used for roughly **$12,000 to $22,000**, with refurbished units, typically enterprise-grade cards processed through certified IT asset disposition (ITAD) vendors, priced around **$21,000 to $34,000** (Source: servnetuk.com). That represents roughly an 85% decline in used pricing from the 2023 secondary-market peak, according to CloudZero's independent cost tracking (Source: cloudzero.com). Depreciation modeling suggests H100 SXM5 units retain, on average, roughly 50% to 60% of their original value at 36 months of use, though the range spans from a conservative 30% to an optimistic 70% depending on how quickly Blackwell adoption displaces existing Hopper fleets (Source: mercatus-ai.com).

Case Studies and Real-World Examples

xAI's Colossus Supercomputer

Elon Musk's AI company xAI built what it describes as "the world's largest AI supercomputer" in Memphis, Tennessee, initially deploying **100,000 NVIDIA H100 GPUs** in a single interconnected cluster in a build xAI states took just 122 days from groundbreaking, before doubling the cluster to **200,000 H100 GPUs** within a further 92 days (Source: x.ai). NVIDIA's own newsroom confirms the cluster relies on NVIDIA's Spectrum-X Ethernet networking platform to achieve that scale, describing the buildout as using 100,000 NVIDIA Hopper GPUs and noting it took only 19 days from the first rack rolling onto the data center floor until training began, with the fabric experiencing zero application latency degradation or packet loss due to flow collisions during operation (Source: nvidianews.nvidia.com). At the low end of publicly cited H100 system pricing, a 200,000-GPU deployment implies a hardware outlay in the multiple-billions-of-dollars range even before power, cooling, and networking infrastructure are added, illustrating the capital intensity behind frontier-scale AI training.

Meta's Llama Training Clusters

Meta disclosed in March 2024 that it operates two flagship AI training clusters, each built around **24,576 NVIDIA H100 Tensor Core GPUs**, used to train its Llama 3 family of large language models, and stated its broader infrastructure roadmap for the end of 2024 called for a build-out including **350,000 NVIDIA H100 GPUs** as part of a portfolio Meta said would feature compute power equivalent to nearly **600,000 H100s** once other accelerator types were included (Source: engineering.fb.com). Meta's engineering team specifically credited the H100's Transformer Engine, and the new FP8 (8-bit floating point) data type it enables, with unlocking training efficiency gains, while noting that fully utilizing clusters of this scale required substantial software investment: unoptimized large clusters saw network utilization as low as 10%, versus 90%-plus after optimization work (Source: engineering.fb.com).

The Stargate Project

Announced in January 2025 by OpenAI, SoftBank, Oracle, and investment firm MGX, the **Stargate Project** committed to investing **$500 billion over four years** to build new U.S. AI data center infrastructure, with $100 billion earmarked for immediate deployment; NVIDIA, Microsoft, Arm, and Oracle were named as key initial technology partners alongside OpenAI (Source: openai.com). By mid-2026, the project's scope had grown to target up to 10 gigawatts of compute capacity, though execution has proven uneven: Microsoft took over a Stargate site in Narvik, Norway in April 2026 that had originally been earmarked for OpenAI, adding capacity for 30,000 NVIDIA Vera Rubin chips and extending a prior $6.2 billion commitment to that site, one of several instances of OpenAI pulling back from specific Stargate locations as it manages spending ahead of a planned initial public offering (Source: datacenterknowledge.com) (Source: techrepublic.com).

Amazon and Anthropic's Trainium-Based Alternative to NVIDIA

Not every large AI buildout runs on NVIDIA silicon, and the Amazon-Anthropic relationship illustrates the competitive pressure NVIDIA's pricing faces at the very top of the market. In April 2026, Amazon and Anthropic announced an expanded collaboration in which Anthropic committed to spend **more than $100 billion over the next ten years** on AWS technologies, securing up to **5 gigawatts** of compute capacity built around Amazon's own Trainium2, Trainium3, and future Trainium4 custom silicon, with Anthropic stating it already uses over one million Trainium2 chips to train and serve its Claude models through the jointly built Project Rainier cluster (Source: anthropic.com). Amazon separately confirmed it would invest up to $25 billion in Anthropic as part of the deal, on top of a prior $8 billion, underscoring how much capital hyperscalers are willing to deploy to build alternatives to NVIDIA-dependent infrastructure (Source: cnbc.com).

CoreWeave as an NVIDIA-Aligned Neocloud

CoreWeave illustrates the opposite dynamic: a cloud provider built almost entirely around reselling NVIDIA GPU capacity at scale. In its first quarter of 2026, CoreWeave signed a new $21 billion commitment with Meta and a multi-year agreement with Anthropic to support Claude model development and deployment, while NVIDIA itself invested $2 billion directly into CoreWeave's Class A common stock, a rare instance of NVIDIA taking an equity stake in one of its own largest infrastructure customers (Source: sec.gov).

Implications and Future Directions

Several forces are likely to shape NVIDIA data center GPU pricing over the next 12 to 24 months. First, the shift toward rack-scale systems as the primary unit of sale, from the GB200 NVL72 at an estimated $2 million to $3 million, to the GB300 NVL72 at $3.7 million to $4 million, to early estimates for Vera Rubin NVL72 racks as high as $8.8 million, suggests that per-GPU price comparisons will become progressively less meaningful for buyers evaluating frontier-scale training capacity (Source: tomshardware.com). Buyers increasingly need to evaluate total system cost, including networking, cooling, and power infrastructure, rather than the price of an individual GPU in isolation.

Second, packaging and memory supply, not GPU logic fabrication, are likely to remain the binding constraint on how quickly prices can fall. With CoWoS lead times still running 52 to 78 weeks and HBM allocated through 2026 across all three major suppliers, meaningful price relief is unlikely before late 2026 or 2027 even as TSMC ramps monthly CoWoS capacity toward a targeted 120,000 to 130,000 wafers per month (Source: siliconanalysts.com).

Third, competitive pressure from AMD's MI-series accelerators and from hyperscaler-designed custom silicon, such as Amazon's Trainium line and Google's TPUs, is real but has so far translated into modest share erosion rather than a fundamental repricing of NVIDIA's top-line products; NVIDIA's own projected data center revenue growth from roughly $100 billion to $150 billion-plus between 2024 and 2026 suggests the total addressable market is expanding faster than competitors can capture share (Source: siliconanalysts.com). Buyers evaluating AMD as a cost-saving alternative should weigh the software ecosystem gap against the 30% to 50% price advantage AMD currently offers on comparable hardware.

Fourth, geopolitical export policy toward China remains a wildcard that can add or remove tens of billions of dollars in addressable revenue on short notice, as demonstrated by the $5.5 billion charge NVIDIA took in 2025 and the fact that its most recent quarterly guidance assumed zero China Data Center compute revenue despite partial policy easing (Source: nvidianews.nvidia.com). Finally, the growing secondary market for used H100 and H200 units gives budget-constrained buyers a genuine lower-cost entry point, and its continued expansion as Blackwell fleets displace Hopper hardware is likely to push used pricing down a further estimated 10% to 20% over the coming year (Source: servnetuk.com).

Power availability, not chip supply alone, is emerging as the practical ceiling on how much NVIDIA compute any single buyer can actually deploy. Epoch AI's model of a one-gigawatt AI data center, requiring an estimated $38 billion in upfront capital spending, illustrates why hyperscalers increasingly describe their growth constraint in gigawatts rather than GPU counts: CoreWeave, for instance, cited surpassing 1 gigawatt of active power in its first-quarter 2026 results as a headline operational milestone alongside its revenue figures, and said it believes it is "well on our way to more than 8 GW by 2030" (Source: epoch.ai) (Source: sec.gov). For buyers evaluating where to site new GPU capacity, colocation power pricing that ranges nearly fourfold, from $80 to $300 per kilowatt per month depending on market, means the choice of physical location can move total cost of ownership by a wider margin than the choice between generations of GPU (Source: mercatus-ai.com). Over the medium term, the combination of a maturing secondary market, a widening set of merchant and custom-silicon alternatives, and a still-tight packaging and memory supply chain suggests NVIDIA data center GPU pricing will stay elevated on new hardware through the rest of 2026, even as the total cost of running AI infrastructure becomes an increasingly multidimensional problem spanning chips, power, networking, and physical siting rather than a single per-GPU number.

Frequently Asked Questions (FAQs)

How much does an NVIDIA H100 cost? As detailed above, new H100 SXM5 GPUs sell for approximately $25,000 to $40,000 per unit depending on OEM and variant, while cloud rental runs $1.38 to $11.06 per GPU-hour with a market median near $2.29.

How much does an NVIDIA H200 cost? The H200 sells new for roughly $30,000 to $40,000 per GPU, similar to the H100, but its cloud rental rates run higher, from $1.45 to $13.78 per hour with a median around $3.95, because of its larger 141GB HBM3e memory pool (Source: cerebrium.ai) (Source: gmicloud.ai).

How much does an NVIDIA B200 cost? Estimated street price is $40,000 to $45,000 per GPU, an 8-GPU HGX B200 server costs around $390,000 (Source: itctshop.com), and a complete DGX B200 system is priced around $500,000 (Source: blogs.novita.ai).

How much does a GB200 NVL72 rack cost? As detailed above, industry estimates put a complete GB200 NVL72 rack, with 72 Blackwell GPUs and 36 Grace CPUs, at roughly $2 million to $3 million, with some quotes reaching $3.4 million once full integration is included.

What is the price difference between H100 vs. H200 vs. B200? Purchase prices for all three sit in a broadly overlapping $25,000 to $45,000 band, but performance and memory differ sharply: the H100 has 80GB HBM3, the H200 has 141GB HBM3e, and the B200 has up to 192GB HBM3e with roughly 4 times the training throughput of an H100 in NVIDIA's own benchmarks (Source: jarvislabs.ai) (Source: runpod.io).

Is AI chip pricing expected to change in 2026? Analysts do not expect a sharp price decline in 2026 given that CoWoS packaging and HBM memory supply remain sold out into 2026 and 2027; the more likely trend is continued high pricing on new hardware alongside a growing, cheaper secondary market for older-generation GPUs (Source: siliconanalysts.com) (Source: servnetuk.com).

Should I buy or rent NVIDIA GPUs? As detailed above, published breakeven models put the purchase-versus-rental crossover at roughly 12,000 hours of usage for a B200-class GPU; workloads below that utilization threshold generally favor renting, while sustained, high-utilization workloads favor purchase.

How does AMD's pricing compare to NVIDIA's? AMD's MI300X and MI325X accelerators are estimated to cost roughly 30% to 50% less than comparable NVIDIA H100 configurations, with cloud rates spanning $1.50 to $6.98 per hour versus $1.99 to $12.29 per hour for the H100, though AMD's market share remains in the mid-single digits by revenue because of NVIDIA's CUDA software ecosystem advantage (Source: siliconanalysts.com).

What does it cost to power a cluster of NVIDIA GPUs, beyond the hardware itself? A 100-GPU H100 cluster requires roughly 145 kW of contracted power capacity, which costs between $11,600 and $43,500 per month depending on whether the facility is priced at wholesale or premium metro colocation rates, on top of the GPU hardware itself (Source: mercatus-ai.com).

Why has NVIDIA GPU pricing stayed high instead of falling like typical computer hardware? Two structural constraints keep prices elevated: TSMC's CoWoS advanced-packaging lines, which bond GPU dies to HBM stacks, remain fully booked at 52-to-78-week lead times, and NVIDIA alone has reserved an estimated 60% of that global capacity, leaving competitors to bid over the remainder (Source: siliconanalysts.com).

Conclusion

NVIDIA data center GPU pricing in mid-2026 reflects a market still defined by scarcity rather than commoditization. Purchase prices for the H100, H200, and B200 all cluster in a broad $25,000 to $45,000 band per GPU, but the real economics increasingly live at the system and rack level, where an 8-GPU server runs $250,000 to $500,000 and a full GB200 or GB300 NVL72 rack runs $2 million to $4 million. Cloud rental remains the more accessible option for most buyers, with rates ranging from roughly $1.50 to $14 per GPU-hour depending on generation, provider tier, and commitment length, and specialist clouds consistently underpricing hyperscalers by a factor of two or more. That pricing power is underpinned by NVIDIA's estimated 75% to 90% share of the AI accelerator market, sustained gross margins in the mid-70s to mid-80s percent range, and a structural packaging and memory bottleneck at TSMC and its HBM suppliers that shows no sign of easing before 2027. Buyers weighing NVIDIA hardware against AMD's lower-priced alternatives, hyperscaler custom silicon, or the fast-maturing used-GPU secondary market now have more genuine options than at any point since the H100 launched, but for training and inference workloads that depend on NVIDIA's CUDA software ecosystem, the price of admission in July 2026 remains among the highest in the history of computing hardware.

Tags: nvidia gpu pricing, h100 price, h200 cost, b200 price, gb200 cost, blackwell gpu price, ai chip pricing, data center gpu, nvidia data center gpu pricing

---