NVIDIA HGX B300 Price, Specs, and Lead Times Explained

Published July 15, 2026 35 min read

<current_article_content>

Executive Summary

The **NVIDIA HGX B300** is an 8-GPU Blackwell Ultra baseboard that original equipment manufacturers (OEMs) such as Dell, Supermicro, Lenovo, and HPE integrate into their own server chassis, and its pricing spans a wide range depending on how it is purchased. As a bare OEM baseboard it starts around **$485,000** (Source: petronellatech.com, a fully built 8-GPU OEM server such as the Supermicro SYS-822GS-NB3RT lists at roughly **$420,000** (Source: turbomaxgpu.com), and NVIDIA's own turnkey **DGX B300** system (which uses the same 8x B300 GPU configuration in an NVIDIA-integrated chassis) sells for approximately **$400,000** in the United States according to Financial Times reporting cited by PCMag (Source: pcmag.com), with European reseller list prices ranging from **€535,000 to €714,000** depending on configuration (Source: aiserver.eu). Cloud rental of B300 capacity is far more accessible: as of **July 2026**, on-demand hourly rates for a single B300 GPU range from about **$2.63 per hour** on the low end (Source: computeprices.com) to **$18.00 per hour** for premium configurations, with an average on-demand rate near **$9.53 per hour** across tracked providers (Source: getdeploying.com).

The HGX B300 is built on NVIDIA's Blackwell Ultra architecture, unveiled at GTC 2025 in March 2025 alongside the rack-scale GB300 NVL72 (Source: nvidianews.nvidia.com), with partner products slated for the second half of 2025 (Source: crn.com) and cloud availability continuing to ramp into early 2026, with one marketplace dating individual B300 GPU shipments to January 2026 (Source: spheron.network). Each B300 GPU packs up to **288 GB of HBM3e memory** and **8 TB/s of memory bandwidth**, a 50% memory increase over the prior-generation HGX B200's 180 GB per GPU (Source: docs.nvidia.com), and the HGX B300 baseboard delivers **144 petaFLOPS of sparse FP4 inference performance** (Source: voltagepark.com) versus 72 petaFLOPS of dense FP8 training throughput (Source: nvidia.com). Compared with the rack-scale GB300 NVL72, which unifies 72 Blackwell Ultra GPUs and 36 Grace CPUs into a single liquid-cooled cabinet, the HGX B300 remains an air-cooled, 8-GPU-per-node building block intended for enterprises and OEM server integrators rather than hyperscale AI factories (Source: nvidia.com).

Lead times for physical HGX B300 systems have compressed since launch but remain measured in weeks rather than days: Dell-based OEM configurations list an estimated delivery of roughly **4 weeks** when ordered through channel partners (Source: marketplace.uvation.com), while NVIDIA and Dell both targeted the air-cooled Dell PowerEdge XE9785 with HGX B300 GPUs for availability starting in the second half of 2025 (Source: dell.com). Demand has been strong enough that NVIDIA's Data Center segment posted record quarterly revenue of **$62.3 billion** in fiscal Q4 2026, up 75% year over year, with Blackwell Ultra cited as a key growth driver (Source: nvidianews.nvidia.com). Real-world deployments illustrate the scale of adoption: Eli Lilly's "LillyPod" DGX SuperPOD uses more than 1,000 B300 GPUs for drug discovery (Source: blogs.nvidia.com), and CoreWeave became the first cloud provider to stand up GB300 NVL72 systems in July 2025 (Source: coreweave.com). Buyers should also weigh a persistent discrepancy in advertised memory: NVIDIA's marketing materials cite up to 288 GB per GPU, but its own HGX comparison table and Dell's shipping product specifications list roughly 270 GB to 279 GB of usable memory per GPU in production systems, a gap this report examines in detail (Source: nvidia.com) (Source: dell.com).

Introduction and Background

Enterprises evaluating large-scale AI infrastructure in mid-2026 face a crowded menu of NVIDIA hardware options, and the **NVIDIA HGX B300** sits near the top of that menu for organizations that want Blackwell Ultra-class compute without committing to a full rack-scale deployment. This report answers the question "what does the NVIDIA HGX B300 cost" from every angle a buyer is likely to encounter: the bare baseboard price paid by OEM integrators, the price of complete OEM and NVIDIA-branded servers, and the hourly cloud rental rate charged by GPU marketplaces and hyperscalers. It also addresses the technical and procurement questions that inevitably accompany a pricing decision, including how the HGX B300 compares with its predecessor the HGX B200, how it differs from the rack-scale GB300 NVL72, what lead times buyers should expect, and when the platform became generally available.

NVIDIA announced Blackwell Ultra, the architecture underpinning the B300 GPU, at its GTC 2025 developer conference in March 2025. CEO **Jensen Huang** described the platform as designed for an era in which "the amount of computation we need at this point as a result of agentic AI, as a result of reasoning, is easily 100 times more than we thought we needed this time last year" (Source: crn.com). Blackwell Ultra comprises two rack- and node-level platforms: the liquid-cooled **GB300 NVL72**, which links 72 Blackwell Ultra GPUs and 36 Grace CPUs in a single cabinet, and the air-cooled **HGX B300 NVL16**, an 8-GPU baseboard design (sold in pairs for a 16-GPU NVLink domain in some OEM chassis) intended for traditional enterprise data centers (Source: nvidianews.nvidia.com). NVIDIA and its OEM partners, including **Dell Technologies**, **Cisco**, **Hewlett Packard Enterprise**, **Lenovo**, and **Supermicro**, along with major cloud providers, targeted the second half of 2025 for Blackwell Ultra product availability (Source: nvidianews.nvidia.com).

The HGX product line itself predates Blackwell Ultra by several generations; it is NVIDIA's standardized baseboard platform that lets OEMs mix and match CPUs, chassis, and cooling around a common 8-GPU NVIDIA compute module connected by NVLink and NVSwitch </current_article_content> (Source: docs.nvidia.com). This flexibility is the primary reason enterprises choose the HGX B300 over NVIDIA's fully integrated DGX B300: it allows data center operators to standardize on a preferred CPU vendor (Intel Xeon or AMD EPYC), select their own storage and networking mix, and negotiate pricing directly with OEMs such as **Dell**, **Supermicro**, **Lenovo**, and **HPE**, all of whom now ship certified HGX B300 systems (Source: petronellatech.com). By July 2026, when this report was compiled, NVIDIA listed both the HGX B300 and the prior-generation HGX B200 as "shipping now" on its official platform comparison page, indicating the two generations are being sold in parallel rather than the newer platform having fully displaced the older one (Source: nvidia.com).

NVIDIA HGX B300 Baseboard and OEM System Pricing

The HGX B300 is sold to the market in three broadly distinct commercial forms, and understanding which one a quoted price refers to is the single most important step in evaluating any "HGX B300 price" figure. The first and cheapest entry point is the **bare 8-GPU baseboard** itself, the component OEMs buy from NVIDIA's supply chain and integrate into their own chassis designs. Specialty AI infrastructure reseller **Petronella** lists this configuration starting **from $485,000**, covering the HGX baseboard alone with 2.3 TB of total HBM3e GPU memory and 72 petaFLOPS of FP8 or 36 petaFLOPS FP4 (their listed figures), noting that the buyer must supply their own CPU, chassis, cooling, and storage (Source: petronellatech.com. This baseboard-only price point is aimed almost exclusively at OEM server manufacturers and systems integrators rather than end-user enterprises, which typically buy a complete server.

The second commercial form is a **complete OEM server** built around the HGX B300 baseboard, which is what most enterprise buyers actually purchase. Pricing here varies with CPU choice, memory, storage, networking, and cooling configuration. **Turbo Max GPU**, a hardware reseller, lists a Supermicro SYS-822GS-NB3RT system with 8x HGX B300 GPUs (288 GB HBM3e each), dual Intel Xeon 6767P 64-core processors, 3 TB of DDR5 ECC memory, and 8x ConnectX-8 800 Gb/s NICs for networking (Source: turbomaxgpu.com), priced at **$420,000.00** and backed by six redundant 6,600W Titanium power supplies (Source: turbomaxgpu.com). Dell's PowerEdge XE9785, the air-cooled successor to the popular XE9680, ships with

8x NVIDIA HGX B300 NVL8 GPUs at 270 GB each and 1,100 watts per GPU, connected via SXM6 sockets and NVLink (Source: dell.com) and cooled by 15 hot-swappable GPU fans (Source: dell.com); channel partner Uvation quotes this configuration with an estimated delivery window of **4 weeks** from order and directs bulk buyers to contact sales for negotiated lead times (Source: marketplace.uvation.com).

The third and most expensive form, if purchased as a fully managed NVIDIA-branded product rather than an OEM server, is the **DGX B300**, discussed in detail in the following section, which uses the same underlying 8x B300 GPU configuration but is assembled, tested, and supported directly by NVIDIA. Table 1 below summarizes the acquisition-price landscape across these commercial forms as documented by vendors and resellers as of **July 2026**.

Table 1 below summarizes list and street pricing for HGX B300 acquisition paths documented during this research.

ACQUISITION PATH	CONFIGURATION	PRICE (LIST OR STREET)	SOURCE
Bare HGX B300 baseboard	8x B300 SXM5, 2.3 TB total HBM3e, OEM-selected CPU/chassis	**From $485,000**	Petronella reseller listing (Source: petronellatech.com
Supermicro SYS-822GS-NB3RT (complete server)	8x HGX B300 288 GB, dual Xeon 6767P, 3 TB DDR5, 8U air-cooled	**$420,000.00**	Turbo Max GPU listing (Source: turbomaxgpu.com)
Dell PowerEdge XE9785 (complete server)	8x HGX B300 NVL8 270 GB SXM6	Delivery quoted; price via sales inquiry, **~4-week** lead time	Uvation marketplace (Source: marketplace.uvation.com)
DGX B300 (NVIDIA turnkey, US)	8x B300 SXM, Intel Xeon 6776P, 2.1 TB GPU memory	**~$400,000** (US retail, per FT sourcing)	PCMag citing Financial Times (Source: pcmag.com)
DGX B300 (EU reseller)	8x B300 SXM, 2.1 TB memory, 3-5 year support tiers	**€535,000 to €714,000**	AIserver.eu (Source: aiserver.eu)
Cloud, on-demand, cheapest tracked	1x B300 GPU, per-hour billing	**$2.63/hr**	ComputePrices.com aggregator (Source: computeprices.com)
Cloud, on-demand, average tracked	1x B300 GPU, blended across providers	**$9.53/hr** average	GetDeploying.com aggregator (Source: getdeploying.com)

The spread in Table 1, from a $420,000 street price for a complete OEM server to nearly $714,000 for a fully configured European DGX B300, reflects real differences in support tiers, warranty length, software bundles (NVIDIA AI Enterprise licensing can be purchased in 3-, 4-, or 5-year terms), and regional taxes and logistics costs, rather than simple price gouging. Buyers comparing quotes should always confirm whether a quoted number includes NVIDIA AI Enterprise software, onsite support, and networking gear, since these line items commonly add **10% to 20%** to the base hardware price in reseller quotes such as AIserver's DGX B300 configurator, which separately prices 3-, 4-, and 5-year NVIDIA AI Enterprise terms alongside the base hardware configuration (Source: aiserver.eu).

Export-control dynamics add a further wrinkle to HGX B300 and DGX B300 pricing that is unique among enterprise IT hardware categories. Because Blackwell-generation GPUs remain banned from export to China, a gray-market premium has emerged: the Financial Times, as reported by PCMag, found that DGX B300 systems selling for roughly $589,000 (4 million yuan) in China's AI server black market earlier in 2026 were fetching **$1.1 million** (8 million yuan) by mid-2026, more than double the roughly $400,000 US retail price (Source: pcmag.com). NVIDIA has stated it does not support or service systems operating in unauthorized territories, and China itself has separately stepped up customs scrutiny of imported NVIDIA AI chips (Source: pcmag.com).

NVIDIA DGX B300 Turnkey System Pricing

While the HGX B300 baseboard is the component OEMs build around, the **DGX B300** is NVIDIA's own fully integrated, factory-tested, single-SKU system built on the same 8x B300 GPU configuration. PCMag describes it as packing "eight Blackwell Ultra GPUs, Intel Xeon 6776P CPUs, and over 2TB of GPU memory," promising twice the performance of the prior DGX B200 (Source: pcmag.com). NVIDIA markets DGX B300 as "the powerhouse for AI innovators," claiming it boosts dense FP4 performance by **1.5x** and attention performance by **2x** over the prior-generation DGX B200, with a specification sheet listing **8x NVIDIA Blackwell Ultra SXM GPUs**, dual Intel Xeon 6776P processors, **2.1 TB of total GPU memory**, 144 petaFLOPS of sparse FP4 Tensor Core performance (108 petaFLOPS dense), 72 petaFLOPS of FP8 Tensor Core performance, 14.4 TB/s of aggregate NVLink bandwidth, and roughly **14 kW** of power consumption per system (Source: nvidia.com).

Pricing for the DGX B300 varies meaningfully by region and configuration. NVIDIA does not publish a headline price on its own product page and instead routes buyers to "Get DGX" and partner channels; independent reporting fills that gap. PCMag, citing Financial Times sourcing, states the DGX B300 "retails for around $400,000 in the US" (Source: pcmag.com). European reseller AIserver.eu, which sells directly to enterprise buyers, lists a base configuration (2x Intel Xeon 6776P, 8x NVIDIA B300 SXM, 2 TB RAM, 30 TB NVMe) starting at **€535,000** with 3-year support and reaching **€714,000** for a fully specified system (Source: aiserver.eu). For context, the same reseller prices the prior-generation DGX B200 at **€455,000 to €628,000**, indicating the B300 generation carries roughly a **15% to 20% price premium** at like-for-like support tiers (Source: aiserver.eu).

The DGX B300's per-GPU memory specification is worth flagging directly, since it is a recurring source of confusion in the market: NVIDIA's own DGX B300 datasheet lists **"Total GPU Memory: 2.1 TB"** for the 8-GPU system, which implies roughly 262 GB usable per GPU rather than the oft-quoted 288 GB ceiling (Source: nvidia.com). This is consistent with independent technical analysis noting that Blackwell Ultra's rated 288 GB HBM3e capacity is a chip-level maximum, and that only **279 GB is "usable" in GB300-class systems** once reserved capacity for reliability and error correction is subtracted (Source: glennklockwood.com). Dell's own product specification sheet for the HGX B300-based PowerEdge XE9785 similarly lists **270 GB per GPU**, not 288 GB, in its shipping configuration, with memory DIMM speeds up to 6400 MT/s across 24 DDR5 slots and a fully populated B300 chassis weighing **360.79 lb (163.65 kg)** (Source: dell.com). Buyers building total cost of ownership models around advertised 288 GB-per-GPU figures should therefore expect roughly **6% to 8% less usable memory** in production hardware than headline marketing numbers suggest.

Beyond hardware, DGX B300 systems are typically sold as part of the broader NVIDIA DGX SuperPOD or DGX BasePOD reference architectures, which bundle networking, storage, and orchestration software; this bundling is one reason enterprise buyers frequently see DGX quotes that run substantially higher than the base hardware price alone, since multi-year support and software licensing are commonly quoted as recurring add-ons rather than one-time costs.

Cloud and Rental Pricing: HGX B300 by the GPU-Hour

For organizations that do not want to commit six or seven figures to owned hardware, HGX B300 capacity is available on an hourly, monthly, or multi-year reserved basis from dozens of GPU cloud providers. This is by far the most accessible entry point to Blackwell Ultra compute, and pricing here moves far more quickly than hardware list prices, so all figures below should be treated as a snapshot as of **mid-July 2026**.

Aggregator **ComputePrices.com**, which tracks pricing across 12 providers, found HGX B300 capacity starting **from $2.63 per hour**, with the cheapest listed provider (Verda) sitting roughly 60% below the tracked average (Source: computeprices.com). At the other end of its tracked range, Oracle Cloud listed on-demand B300 capacity at **$15.00 per hour** per GPU as of July 16, 2026 (Source: computeprices.com). A second aggregator, **GetDeploying.com**, which tracks 91 listings across 16 providers, reports a wider band: on-demand B300 pricing ranges from **$3.27 per hour** (a 36-month reserved rate from Impossible Cloud) up to **$18.00 per hour** for an Oracle Cloud bare-metal instance, with billing-type averages of **$9.53/hr on-demand**, **$6.66/hr reserved**, **$4.09/hr spot**, and **$4.49/hr custom contract** (Source: getdeploying.com). At a standard 720-hour month, GetDeploying calculates that renting a single B300 GPU costs between **$2,354.40 and $12,960.00 per month** depending on provider and commitment level (Source: getdeploying.com).

GPU marketplace **Spheron** reports a somewhat higher on-demand figure of **$9.16 per hour** for B300 SXM6 instances as of early July 2026, noting this undercuts fully managed DGX B300 cloud offerings, which it prices at **$12 to $18 per hour**, by roughly **30% to 50%** while still offering bare-metal access (Source: spheron.network).

Hyperscale-adjacent cloud provider **CoreWeave** publishes its own HGX B300 pricing directly. As of this research, CoreWeave's public pricing page lists an 8-GPU HGX B300 instance (270 GB per GPU, 2.2 TB total VRAM) available via a "Contact sales" custom quote rather than a fixed on-demand rate, while the prior-generation HGX B200 instance carries a published on-demand rate of **$68.80 per hour** for the full 8-GPU node, equivalent to roughly $8.60 per GPU-hour (Source: coreweave.com). For comparison, CoreWeave's GB200 NVL72 rack-scale instances carry a published on-demand rate of **$42.00 per hour** per GPU-equivalent unit, illustrating that rack-scale liquid-cooled systems do not necessarily command a per-GPU premium over air-cooled HGX nodes once instance-level pricing is normalized (Source: coreweave.com). CoreWeave's rack-scale GB200

and GB300 instances are each built from paired Superchips, with the provider noting "each instance is comprised of 2 GB200 or GB300 Superchips, where each Superchip has 1 Grace CPU and 2 Blackwell GPUs," and clarifying that per-GPU pricing on its inference platform is billed separately from its standard rack-instance pricing (Source: coreweave.com) (Source: coreweave.com).

For workload-planning purposes, it is useful to benchmark B300 rental costs against older Hopper-generation and prior Blackwell-generation GPUs. Spheron quotes H200 SXM instances at roughly **$3.70 per hour** on-demand and H100 instances between **$2.01 per hour** (PCIe) and **$3.92 per hour** (SXM5) on its own platform, and separately estimates the B300 delivers roughly **11 to 15 times more inference throughput per GPU** than the H100 for large language model serving, which can make the B300 the cheaper option on a per-token basis despite its higher hourly rate, particularly for models above roughly 70 billion parameters that would otherwise require multi-GPU sharding on smaller-memory cards (Source: spheron.network).

HGX B300 Specifications and Comparative Context

Understanding what the HGX B300 actually delivers for its price requires separating three distinct levels of specification: the individual **B300 GPU die**, the **8-GPU HGX B300 baseboard**, and the rack-scale **GB300 NVL72** system that shares the same GPU silicon. NVIDIA's Blackwell Ultra GPU is a dual-reticle design combining two reticle-limited dies connected by a 10 TB/s NV-HBI (High-Bandwidth Interface) link, manufactured on TSMC's 4NP process with **208 billion transistors** (a 2.6x increase over the single-die Hopper GPU's 80 billion transistors), and it introduces **NVFP4**, a new 4-bit floating point precision format that delivers 15 petaFLOPS of dense compute per GPU while claiming near-FP8 accuracy at roughly 1.8x lower memory footprint than FP8 (Source: developer.nvidia.com). Independent processor documentation corroborates the dual-die design, describing the B300 as containing "2 reticle-sized GPUs" with "20,480 CUDA cores (128 per SM)" and "640 tensor cores" across the package (Source: glennklockwood.com). Memory-wise, each GPU carries up to **288 GB of HBM3e** across eight 12-Hi memory stacks, delivering up to **8 TB/s of bandwidth**, a 3.6x capacity increase and 2.4x bandwidth increase versus the original H100 (Source: developer.nvidia.com), a figure Voltage Park's own product listing echoes, describing HGX B300 as delivering "up to 11× faster inference and 4× faster training than the previous generation" thanks to Blackwell Ultra GPUs, ConnectX-8 networking, and Mission Control software (Source: voltagepark.com).

At the baseboard level, the 8-GPU HGX B300 aggregates these individual GPUs with fifth-generation NVLink and NVSwitch, providing 1.8 TB/s of GPU-to-GPU bandwidth and 14.4 TB/s of total NVLink bandwidth across the node, an identical topology to the HGX B200. Where the two platforms diverge sharply is memory capacity, networking bandwidth, and attention-layer inference throughput, as summarized in Table 2.

Table 2 below reproduces NVIDIA's own published side-by-side specifications for the HGX B300 and HGX B200 platforms.

SPECIFICATION	HGX B300 (BLACKWELL ULTRA)	HGX B200 (BLACKWELL)
Form Factor	8x NVIDIA Blackwell Ultra SXM	8x NVIDIA Blackwell SXM
FP4 Tensor Core (sparse | dense)	144 PFLOPS | 108 PFLOPS	144 PFLOPS | 72 PFLOPS
FP8/FP6 Tensor Core	72 PFLOPS	72 PFLOPS
INT8 Tensor Core	3 POPS	72 POPS
FP16/BF16 Tensor Core	36 PFLOPS	36 PFLOPS
FP64/FP64 Tensor Core	10 TFLOPS	296 TFLOPS
Total Memory	2.1 TB (up to 2.3 TB per reference-architecture docs)	1.4 TB
NVIDIA NVLink	Fifth generation	Fifth generation
NVLink GPU-to-GPU Bandwidth	1.8 TB/s	1.8 TB/s
Total NVLink Bandwidth	14.4 TB/s	14.4 TB/s
Networking Bandwidth	1.6 TB/s	0.8 TB/s
Attention Performance (vs. Blackwell)	2x	1x

*Source: NVIDIA HGX platform specifications page (Source: nvidia.com), cross-checked against AceCloud's independent republication of the same table, which also notes the HGX B200's per-GPU memory as **180 GB HBM3e** with **14.4 TB/s** of total NVLink bandwidth (Source: acecloud.ai).*

Two things stand out in Table 2. First, the HGX B300's **FP64 (double-precision) performance actually falls dramatically**, from 296 TFLOPS on the B200 to just 10 TFLOPS on the B300, a nearly 30x reduction, because Blackwell Ultra's die area and power budget were reallocated toward FP4 and attention-layer inference acceleration rather than traditional scientific double-precision compute (Source: nvidia.com). Organizations running classic HPC or FP64-dependent simulation workloads alongside AI training should treat the B300 as an inference- and reasoning-optimized platform, not a general-purpose HPC upgrade over B200 (Source: acecloud.ai). Second, total memory again shows the discrepancy noted earlier: NVIDIA's own HGX specifications page lists **2.1 TB** total memory for the B300 baseboard in its headline comparison table, while its separate HGX AI Factory enterprise reference architecture documentation lists **"Memory per Node: 2.30TB"** for the same 8-GPU configuration (Source: docs.nvidia.com). This report presents both figures rather than resolving the discrepancy in either direction, since it appears to reflect a difference between theoretical maximum (288 GB x 8 = 2,304 GB, rounded to 2.3 TB) and NVIDIA's own conservative "shipping" specification of 2.1 TB, itself consistent with Dell's 270 GB-per-GPU shipping spec (270 x 8 = 2,160 GB, close to 2.1 TB).

The HGX B300 also differs meaningfully from the **GB300 NVL72**, the other Blackwell Ultra platform NVIDIA introduced alongside it. Where the HGX B300 is an air-cooled, 8-GPU building block designed to slot into conventional enterprise racks, the GB300 NVL72 is a fully liquid-cooled, rack-scale system that fuses **72 Blackwell Ultra GPUs and 36 Grace CPUs** into what NVIDIA describes as acting "as a single massive GPU built for test-time scaling" (Source: nvidianews.nvidia.com). The Grace CPUs anchoring the rack are, per NVIDIA, "designed for modern data center workloads" and provide "2x the energy efficiency of today's leading server processors" (Source: nvidia.com). NVIDIA's GB300 NVL72 specification sheet lists **20 TB of GPU memory** and up to 576 TB/s of aggregate GPU memory bandwidth at the rack level, and 130 TB/s of NVLink bandwidth across 2,592 Arm Neoverse V2 CPU cores (Source: nvidia.com) (Source: nvidia.com). CRN's own reporting on the GB300 NVL72 corroborates the rack-level scale, noting the platform "achieve[s] 1.1 exaflops of FP4 dense computation, and it comes with 20 TB of high-bandwidth memory as well as 40 TB of fast memory" (Source: crn.com), while independent technical documentation separately places the full 72-GPU, 36-CPU rack at roughly **130 trillion transistors**, **37 TB of "Fast Memory,"** and **18 NVLink Switches at 130 TB/s** in aggregate (Source: glennklockwood.com) (Source: glennklockwood.com), a minor variance from CRN's "40 TB" figure likely reflecting rounding or configuration differences between sources. Dell, which builds the GB300 NVL72 into its PowerEdge XE9712 platform, states the system "offers efficiency at rack scale for training, 50 times more AI

reasoning inference output and 5x improvement in throughput" versus Hopper (Source: dell.com), a claim Voltage Park echoes in its own marketing copy citing "up to 50x higher inference output for reasoning models compared to Hopper" (Source: voltagepark.com). Table 3 compares the three Blackwell Ultra system classes side by side.

Table 3 below compares the HGX B300, DGX B300, and GB300 NVL72 at the system level to clarify which product family a given price quote refers to.

SYSTEM	GPU COUNT	COOLING	TOTAL GPU MEMORY	TYPICAL BUYER	INDICATIVE PRICE
HGX B300 baseboard	8x B300 SXM	Air (OEM chassis dependent)	~2.1 to 2.3 TB	OEMs, systems integrators	From $485,000 (baseboard) (Source: petronellatech.com
DGX B300	8x B300 SXM	Air	2.1 TB	Enterprises wanting a turnkey NVIDIA appliance	~$400,000 to €714,000 depending on region/config (Source: pcmag.com)
GB300 NVL72	72x B300 SXM, 36x Grace CPU	Liquid, rack-scale	20 TB	Hyperscalers, frontier AI labs	Not publicly listed; sold via cloud partners on contract (Source: coreweave.com)

The practical decision for most enterprise buyers, per Table 3, comes down to whether they need rack-scale, liquid-cooled density (GB300 NVL72, generally reserved for hyperscalers and frontier labs), a fully managed NVIDIA appliance (DGX B300), or a customizable OEM server built on the same silicon at a somewhat lower price point (HGX B300 via Dell, Supermicro, Lenovo, or HPE). Independent analysis from GPU marketplace Spheron summarizes this trade-off concisely: HGX B300 is "the server-grade baseboard... Same GPU performance as DGX, but you choose the CPU, cooling, and chassis configuration. Generally more flexible and often cheaper than DGX" (Source: spheron.network).

Comparative Context and Market Positioning

Positioning the HGX B300 against the broader NVIDIA data-center GPU lineup requires looking both backward, at the Hopper-generation H100 and H200, and forward, at the Rubin-generation hardware NVIDIA has already begun previewing. Compared with the H100, which NVIDIA's own developer documentation uses as the baseline for most Blackwell Ultra performance claims, the B300 offers **3.6x more on-package memory** (288 GB versus 80 GB) and **2.4x more memory bandwidth** (8 TB/s versus 3.35 TB/s) (Source: developer.nvidia.com).

Compared with the H200, the immediate Hopper-refresh predecessor with 141 GB of HBM3e per GPU, the B300's 288 GB advertised capacity (or roughly 270 to 279 GB usable, per the discrepancy noted above) more than doubles per-GPU memory. AceCloud's decision framework for the platform captures why this matters most for reasoning workloads: "more HBM per GPU keeps KV cache, activations and shards resident, which reduces eviction and recompute," particularly for long-context, high-concurrency serving scenarios (Source: acecloud.ai). By the same token, AceCloud notes that the prior-generation B200 baseboard already offers "1.44 TB of HBM3e across an 8-GPU baseboard," meaning only workloads genuinely constrained by that ceiling need the B300's larger pool (Source: acecloud.ai). That memory advantage comes at a real hourly cost premium, however: at prevailing July 2026 cloud rates, the roughly $9 to $18 per hour B300 range documented above costs meaningfully more than the H200's approximately $3.70-per-hour on-demand rate on the same marketplace, a premium buyers should weigh against their actual per-GPU memory requirements before defaulting to the newest SKU, per the hourly figures presented in the Cloud and Rental Pricing section above.

Looking forward, NVIDIA has already signaled the HGX B300's successor lineage. At its fourth-quarter fiscal 2026 earnings call, NVIDIA unveiled the **Rubin platform**, describing six new chips that promise "up to a 10x reduction in inference token cost, compared with the NVIDIA Blackwell platform," with AWS, Google Cloud, Microsoft Azure, and Oracle Cloud Infrastructure named among the first cloud providers set to deploy Vera Rubin-based instances (Source: nvidianews.nvidia.com). NVIDIA's own HGX platform page already lists the next-generation **HGX Rubin NVL8**, claiming up to **10x more token factory throughput** than HGX B200 using architectural upgrades including 400 petaFLOPS of NVFP4 compute (versus the B300's 144 petaFLOPS) and new HBM4 memory (Source: nvidia.com). At GTC 2025, NVIDIA also disclosed a broader roadmap: the Vera Rubin NVL144 platform is slated for the second half of 2026, Rubin Ultra NVL576 for the second half of 2027, and a next-generation Feynman-based platform for 2028, confirming NVIDIA's now-annual data-center GPU release cadence that began with the H200 launch in 2024 (Source: crn.com).

Within the current Blackwell Ultra generation, the HGX B300 also competes for enterprise budget against a same-generation AMD alternative on shared server platforms; Dell's own PowerEdge XE9785 configurator, for example, lists 8x AMD Instinct MI355X (288 GB) as an alternative GPU option in the identical chassis alongside the 8x NVIDIA HGX B300 NVL8 configuration, underscoring that server OEMs are increasingly designing shared, GPU-agnostic chassis rather than NVIDIA-exclusive platforms (Source: dell.com). Independent GPU-cloud comparison site Voltage Park similarly lists HGX B200, HGX GB200, HGX B300, and HGX GB300 side by side as concurrently reservable platforms, underscoring that the four-way Blackwell and Blackwell Ultra product family, not just the B300 in isolation, is what most buyers are actually choosing among in 2026 (Source: voltagepark.com).

Data Analysis and Evidence

The quantitative record around the HGX B300's launch, adoption, and financial impact is unusually well documented because NVIDIA is a publicly traded company subject to SEC disclosure and earnings-call scrutiny. NVIDIA's Data Center segment, which houses all HGX and DGX B300 revenue, posted **record quarterly revenue of $51.2 billion** in the third quarter of fiscal 2026 (ended late October 2025), up **66% year over year**, with CEO Jensen Huang telling analysts "Blackwell sales are off the charts, and cloud GPUs are sold out" (Source: futurumgroup.com) (Source: futurumgroup.com). The same analyst coverage noted that "Blackwell GB300 shipments have crossed GB200 and now account for roughly two-thirds of Blackwell revenue," and that "Blackwell Ultra delivered 5x faster time-to-train versus Hopper on MLPerf, and on DeepSeek-R1 mixture-of-experts, Blackwell achieved 10x performance per watt and 10x lower cost per token versus H200" (Source: futurumgroup.com) (Source: futurumgroup.com). That growth accelerated further in the fourth quarter: NVIDIA reported **record quarterly Data Center revenue of $62.3 billion**, up **22% from Q3 and 75% year over year**, contributing to full-year fiscal 2026 Data Center revenue of **$193.7 billion**, up 68%, on total company revenue of **$215.9 billion** (Source: nvidianews.nvidia.com). CEO **Jensen Huang** attributed part of this growth directly to the Blackwell Ultra generation, stating "Grace Blackwell with NVLink is the king of inference today, delivering an order-of-magnitude lower cost per token" (Source: nvidianews.nvidia.com).

Third-party benchmarking corroborates NVIDIA's performance claims with independent data. According to **SemiAnalysis InferenceX** benchmarks referenced in NVIDIA's own fourth-quarter earnings materials, "NVIDIA Blackwell Ultra delivers up to 50x better performance and 35x lower cost for agentic AI compared with the NVIDIA Hopper platform" (Source: nvidianews.nvidia.com). NVIDIA's own DGX B300 product page cites the same benchmark family for a companion pair of claims: "up to 50x higher throughput per megawatt and up to 35x lower cost per token than NVIDIA Hopper," and on industry-standard MLPerf Inference v6.0 (April 2026), Blackwell Ultra-powered systems delivered **2.5 million tokens per second** on the DeepSeek-R1 671-billion-parameter reasoning model, up to 2.7x higher than Blackwell Ultra's debut MLPerf submissions roughly six months earlier (Source: nvidia.com).

On the hardware architecture side, NVIDIA's own technical comparison table across three GPU generations quantifies the generational leap precisely: transistor count rose from 80 billion (Hopper, single die) to 208 billion (Blackwell and Blackwell Ultra, dual die), and maximum HBM capacity rose from 141 GB (H200) to 192 GB (Blackwell) to 288 GB (Blackwell Ultra), with per-GPU bandwidth confirmed by independent technical documentation at up to **8 TB/s** (Source: developer.nvidia.com) (Source: glennklockwood.com). At the cloud-market level, GetDeploying's tracked dataset of 91 B300 listings across 16 providers shows the market segmenting cleanly by commitment level, with reserved-capacity listings averaging **$6.66/hr** against **$9.53/hr** for on-demand listings, a roughly **30% discount** for buyers willing to commit to multi-month or multi-year terms, consistent with the billing-type breakdown presented in the Cloud and Rental Pricing section above. Separately, contemporaneous analyst coverage of NVIDIA's Q3 fiscal 2026 earnings call noted that "Rubin remains on track to ramp in the second half of FY 2027," reinforcing the multi-year window before Blackwell Ultra pricing is likely to ease (Source: futurumgroup.com).

Case Studies and Real-World Examples

Eli Lilly's LillyPod: The Largest Pharmaceutical-Owned AI Factory. Pharmaceutical giant **Eli Lilly** announced in October 2025 that it was deploying what NVIDIA and Lilly both describe as the largest, most powerful AI factory wholly owned and operated by a pharmaceutical company: the world's first NVIDIA DGX SuperPOD built with **DGX B300 systems** (Source: blogs.nvidia.com). The Lilly investor relations announcement is subtitled "New AI capabilities will help scientists identify, optimize and validate new molecules," with additional applications spanning "manufacturing, medical imaging and enterprise AI agents" (Source: investor.lilly.com). Lilly states that with the new AI factory, "scientists will be able to train AI models on millions of experiments to test potential medicines, dramatically expanding the scope and sophistication of drug discovery efforts" (Source: investor.lilly.com). The deployment, built with **1,016 NVIDIA Blackwell Ultra GPUs**, delivers over **9,000 petaflops of AI performance**, which NVIDIA's own materials describe as the equivalent of "over 9 quintillion math problems every second" (Source: blogs.nvidia.com). NVIDIA notes that its Mission Control software "allows Lilly to manage its DGX SuperPOD, orchestrate workloads, monitor performance and automate AI operations securely and efficiently across more than 1,000 GPUs" (Source: blogs.nvidia.com). NVIDIA's vice president of health care, Kimberly Powell, framed the deployment's significance for medicine broadly: "the AI industrial revolution will have its most profound impact on medicine, transforming how we understand biology," adding that "modern AI factories are becoming the new instrument of science" (Source: investor.lilly.com). Eli Lilly's own investor

relations release describes the system as "powered by more than 1,000 B300 GPUs on a unified networking fabric, which means communication across GPUs, storage and related systems runs on just one high-speed network" (Source: investor.lilly.com). The deployment sits within Lilly's broader **$50 billion** commitment to expanding US manufacturing and research and development infrastructure (Source: blogs.nvidia.com).

CoreWeave's First-to-Market GB300 NVL72 Deployment. On **July 3, 2025**, cloud provider CoreWeave announced it was the first AI cloud provider to deploy NVIDIA's GB300 NVL72 systems for customers, working with **Dell, Switch, and Vertiv** to build the initial deployment (Source: coreweave.com). CoreWeave co-founder and CTO Peter Salanki said, "We're proud to be the first to stand up this transformative platform and help innovators prepare for the next exciting wave of AI" (Source: coreweave.com). This initial deployment expanded CoreWeave's existing Blackwell fleet, which already included NVIDIA HGX B200 and GB200 NVL72 instances, illustrating how cloud providers typically run multiple Blackwell generations in parallel rather than fully cutting over to each new SKU (Source: coreweave.com). CoreWeave later built on that GB300 footprint for benchmarking purposes as well, reporting record results across "the largest NVIDIA GB300 NVL72 cluster in the benchmark" for MLPerf Training v6.0 (Source: coreweave.com).

Dell and NVIDIA's Coordinated OEM Rollout. In May 2025, Dell Technologies and NVIDIA jointly unveiled the next generation of Dell's enterprise AI server lineup built specifically around HGX B300 and GB300 NVL72, with Dell stating its new PowerEdge XE9780 and XE9785 servers "can deliver up to four times faster large language model (LLM) training with the 8-way NVIDIA HGX B300" compared with the prior-generation PowerEdge XE9680 (Source: dell.com). Dell's public availability commitment specified that air-cooled PowerEdge XE9780 and XE9785 servers with HGX B300 GPUs "will be available in 2H 2025," with direct liquid-cooled XE9780L and XE9785L variants following later (Source: dell.com). This case illustrates the practical, months-long lag that is typical between an NVIDIA architecture announcement (Blackwell Ultra, March 2025) and volume OEM shelf availability (second half of 2025), a gap buyers should factor into any procurement timeline.

China's Gray Market for Restricted DGX B300 Hardware. As detailed in the pricing section above, Financial Times reporting relayed by PCMag documents a functioning gray market for DGX B300 systems in China despite an active US export ban on Blackwell-class GPUs to the country. Sources told the FT that DGX B300 units "until recently... was available for 4 million yuan, or around $589,000, but they are now being sold for 8 million yuan, or around $1.1 million," a premium the article attributes to escalating smuggling risk and enforcement rather than simple demand (Source: pcmag.com). This case demonstrates that HGX and DGX B300 pricing cannot be evaluated purely on a hardware-cost basis; regulatory and export-control risk materially affects real-world transaction prices in specific geographies.

Implications and Future Directions

Several forward-looking implications emerge from the pricing, specification, and adoption data gathered in this report. First, the memory-capacity discrepancy documented throughout this report, NVIDIA's marketed 288 GB per GPU ceiling versus the 262 GB to 279 GB commonly shipped in production HGX and DGX B300 systems, is likely to recur with each future Blackwell Ultra and Rubin-generation product, and procurement teams should build total cost of ownership models around vendor-confirmed shipping specifications (as published directly by Dell or NVIDIA's own datasheets) rather than headline marketing maximums.

Second, cloud pricing for HGX B300 capacity is likely to remain elevated through at least the remainder of 2026. GPU marketplace Spheron does not expect meaningful relief until Vera Rubin volume shipments begin pulling frontier training and inference workloads away from Blackwell Ultra capacity, an event it estimates is roughly **6 to 12 months** out from mid-2026 (Source: spheron.network). Buyers planning multi-year infrastructure budgets should therefore weight reserved or multi-year cloud contracts, which carry a roughly 30% discount to on-demand rates per the data presented above, more heavily than they might for a more mature, price-stable GPU generation.

Third, the coexistence of HGX B200 and HGX B300 as simultaneously "shipping now" products signals that price-sensitive buyers with workloads that do not require the B300's extra memory headroom retain a legitimate, still-current lower-cost Blackwell option rather than being forced onto the newest and most expensive SKU. This dual-track availability, paired with NVIDIA's now-established annual cadence (Hopper, Blackwell, Blackwell Ultra, and upcoming Rubin), suggests enterprises should plan hardware refresh cycles around workload-specific memory and throughput requirements rather than simply chasing each new architecture as it launches.

Finally, the sustained pace of NVIDIA's Data Center revenue growth, record quarters in both Q3 and Q4 of fiscal 2026, alongside Blackwell Ultra's outsized role in that growth per Jensen Huang's own earnings commentary, indicates continued tight allocation and multi-week lead times for HGX B300 hardware are likely to persist through 2026 even as OEM manufacturing capacity ramps. Enterprises with time-sensitive AI infrastructure needs should engage OEM sales channels early and confirm current lead times directly, since the roughly 4-week estimate documented by one reseller in this report is best treated as a floor rather than a guarantee across all configurations and order volumes (Source: marketplace.uvation.com).

Frequently Asked Questions (FAQs)

What is the price of the NVIDIA HGX B300? Pricing depends entirely on the acquisition path. A complete OEM server such as Supermicro's SYS-822GS-NB3RT lists near **$420,000** (Source: turbomaxgpu.com), and cloud rental of a single B300 GPU ranges from roughly **$2.63 to $18.00 per hour** depending on provider and commitment (Source: computeprices.com) (Source: getdeploying.com).

What are the NVIDIA HGX B300 specifications? The HGX B300 pairs 8x Blackwell Ultra SXM GPUs, each with up to 288 GB of HBM3e memory, on a baseboard delivering up to **144 petaFLOPS for inference and 72 petaFLOPS for training** (Source: voltagepark.com), connected by fifth-generation NVLink at 1.8 TB/s GPU-to-GPU and 14.4 TB/s aggregate bandwidth (Source: docs.nvidia.com).

How does the HGX B300 compare to the HGX B200? The B300 offers roughly 50% more memory per GPU (288 GB versus 180 GB), doubled networking bandwidth (1.6 TB/s versus 0.8 TB/s), and 2x the attention-layer inference performance, but a dramatically lower FP64 rating (10 TFLOPS versus 296 TFLOPS), since Blackwell Ultra's design trades scientific double-precision throughput for inference and reasoning acceleration (Source: acecloud.ai).

What is the NVIDIA HGX B300 baseboard? It is NVIDIA's standardized 8-GPU hardware module, internally connected by NVLink and NVSwitch and externally connected via 8x ConnectX-8 SuperNICs at up to 800 Gb/s per GPU, which OEMs integrate into their own server chassis with a CPU and cooling solution of their choice (Source: docs.nvidia.com).

What is the HGX B300's lead time? Documented reseller lead times run around **4 weeks** for standard Dell-based configurations ordered through channel partners, though buyers requiring bulk orders are directed to contact sales for confirmed schedules, and OEM general availability windows opened in the second half of 2025 (Source: marketplace.uvation.com) (Source: dell.com).

What are the individual NVIDIA B300 GPU specifications? A single Blackwell Ultra B300 GPU is a dual-reticle design with 208 billion transistors, 160 Streaming Multiprocessors, 640 Tensor Cores, up to 288 GB of HBM3e memory across eight 12-Hi stacks, up to 8 TB/s of memory bandwidth, and 15 petaFLOPS of dense NVFP4 compute (Source: developer.nvidia.com).

When was the HGX B300 released? NVIDIA announced Blackwell Ultra and the HGX B300 NVL16 platform at GTC 2025 in March 2025, and targeted partner and OEM availability for the second half of 2025 (Source: nvidianews.nvidia.com). By July 2026 the platform remained listed as "shipping now" alongside its predecessor, per the Introduction above.

How does the HGX B300 compare to the GB300 NVL72? The HGX B300 is an air-cooled, 8-GPU building block for conventional enterprise racks, while the GB300 NVL72 is a fully liquid-cooled, 72-GPU, 36-Grace-CPU rack-scale system delivering roughly 9x the GPU memory (20 TB versus ~2.1 TB) of a single HGX B300 node (Source: nvidia.com) and roughly 9x its sparse FP4 compute (1,440 versus 144 petaFLOPS) (Source: nvidia.com).

Conclusion

The NVIDIA HGX B300 does not have a single price; it has a price for every stage of the buying journey, from a roughly $485,000 bare OEM baseboard, to $400,000 to $714,000 complete DGX B300 systems depending on region and support tier, down to a few dollars per GPU-hour for cloud rental. What unifies these numbers is the underlying Blackwell Ultra silicon: a dual-reticle, 208-billion-transistor GPU offering up to 288 GB of HBM3e memory and 15 petaFLOPS of dense FP4 compute, assembled eight-wide on the HGX baseboard that OEMs from Dell to Supermicro to Lenovo now ship in volume. Buyers evaluating the platform in the second half of 2026 should anchor their budgets to vendor-confirmed shipping specifications rather than marketing maximums, given the documented gap between advertised 288 GB and shipped 262 to 279 GB per-GPU memory figures, and should expect cloud pricing to remain elevated until Vera Rubin capacity begins easing demand on Blackwell Ultra hardware later in 2026 or into 2027. For workloads that genuinely need the B300's extra memory headroom, such as long-context reasoning inference or training models above roughly 70 billion parameters without aggressive sharding, the premium over the H200 and HGX B200 is well documented and, per independent cost-per-token analysis, often justified. For workloads that fit comfortably within 141 GB to 180 GB of per-GPU memory, the still-shipping HGX B200 and H200 remain lower-cost alternatives built on the same broader NVIDIA software stack.

Tags: nvidia hgx b300, hgx b300 price, nvidia b300, blackwell ultra, hgx b300 vs hgx b200, gb300 nvl72, dgx b300, gpu pricing, ai infrastructure, nvidia blackwell

---