NVIDIA DGX B300 Price: Specs and Lead Times Guide

Published July 17, 2026 39 min read

Executive Summary

NVIDIA's **DGX B300** is an 8-GPU, Blackwell Ultra-based AI system that shipped commercially beginning in January 2026, and as of July 2026 it does not carry a fixed manufacturer list price (Source: tech-insider.org). Instead, market pricing for a fully configured 8-GPU system clusters between **$300,000 and $350,000** at the low end, based on named integrator quotes from Q1 2026 (Source: ifactoryapp.com), while other integrators quote fully configured 8-GPU HGX-class systems as high as **$400,000** (Source: www.amcompute.com), and European partner storefronts list configured DGX B300 units anywhere from roughly €535,000 to over €1.4 million including VAT depending on RAM, support tier, and software bundle (Source: aiserver.eu) (Source: servermall.com). A single **NVIDIA B300** GPU purchased outright runs approximately **$53,000** as of early July 2026 (Source: spheron.network), and cloud access to B300 capacity spans roughly **$5.65 to $8.70 per GPU-hour** across named on-demand providers such as Runpod, Nebius, Sesterce, Verda, and Lyceum, up to **$17.80 to $18.00 per GPU-hour** on premium hyperscale listings from AWS and Oracle Cloud Infrastructure (Source: getdeploying.com).

Specification-wise, the DGX B300 pairs **eight NVIDIA Blackwell Ultra GPUs** with dual **Intel Xeon 6776P** processors, delivering **144 petaFLOPS** of sparse FP4 inference performance and **72 petaFLOPS** of FP8 training performance in a **10U** chassis drawing roughly **14 to 14.5 kilowatts** (Source: nvidia.com) (Source: docs.nvidia.com). Total onboard GPU memory is documented inconsistently across NVIDIA's own materials: the primary product page states **2.1 TB**, while the technical user guide, the official datasheet, and multiple independent system integrators state **2.3 TB** (eight 288 GB HBM3e modules) (Source: nvidia.com) (Source: docs.nvidia.com). Independent GPU cloud provider acecloud.ai corroborates the lower figure, noting that "NVIDIA lists 2.1 TB total GPU memory" on its DGX B300 spec page (Source: acecloud.ai), a gap this report attributes most plausibly to raw versus usable capacity after reserved overhead. Compared with the prior-generation **DGX B200**, the DGX B300 delivers **1.5 times** higher dense FP4 throughput and **2 times** higher attention-layer performance, at the cost of a dramatic reduction in FP64 double-precision compute (Source: nvidia.com) (Source: exxactcorp.com).

NVIDIA announced the Blackwell Ultra platform, including the DGX B300 and the rack-scale GB300 NVL72, at GTC in March 2025 (Source: nvidianews.nvidia.com), with partner shipments originally guided for the second half of 2025 (Source: nvidianews.nvidia.com) and volume shipping confirmed to have begun in January 2026 (Source: spheron.network) (Source: docs.nvidia.com). As of mid-2026, order-to-delivery **lead times** for a

standard DGX B300 run **roughly 8 to 20 weeks** depending on the source consulted, generally longer than DGX B200 lead times because of tighter Blackwell Ultra allocation (Source: spheron.network) (Source: ifactoryapp.com). This report walks through system, enterprise, and cloud pricing tiers; a full specification comparison against DGX B200; the release and lead-time record; and named deployments at Eli Lilly, CoreWeave, and Equinix that illustrate how buyers are actually putting DGX B300 capacity to work.

Introduction and Background

Enterprises evaluating on-premises AI infrastructure in mid-2026 face a genuinely confusing pricing landscape for NVIDIA's newest DGX system. The **NVIDIA DGX B300**, built around the **Blackwell Ultra** GPU architecture, was unveiled at NVIDIA's GTC conference on March 18, 2025, as part of a broader platform refresh that also introduced the rack-scale **GB300 NVL72** system (Source: nvidianews.nvidia.com). Unlike consumer GPUs or software subscriptions, DGX systems are not sold through a public storefront with a posted price; NVIDIA routes purchases through original equipment manufacturer (OEM) partners, systems integrators, and cloud providers, each of whom sets its own price, bundles its own support terms, and quotes its own lead time. An aggregator that tracks named-source pricing signals, Tech Insider, states plainly that NVIDIA **"has not officially published retail prices for individual B200 or B300 units"**, meaning every dollar figure a buyer encounters is an OEM quote, marketplace listing, or reseller estimate rather than a manufacturer list price (Source: tech-insider.org).

This creates a genuine research problem for anyone typing "nvidia dgx b300 price" into a search engine: the honest answer is a range, not a number, and the range itself depends on configuration, region, support term, and whether the buyer is purchasing hardware outright, **renting cloud capacity**, or committing to a multi-rack DGX SuperPOD. The DGX B300 is positioned by NVIDIA as **"the powerhouse for AI innovators, delivering the hyperscaler performance needed to build a modern AI factory,"** built for **large language model (LLM) inference** and training workloads that increasingly involve "reasoning" models generating long chains of intermediate tokens before a final answer (Source: nvidia.com). It succeeds the DGX B200, which shipped on the original Blackwell architecture (Source: acecloud.ai), and precedes NVIDIA's next-generation Vera Rubin platform, expected to reach cloud providers in the second half of 2026 (Source: spheron.network).

This report is written for technical buyers, IT procurement teams, and infrastructure analysts who need a grounded, source-verified answer to what a DGX B300 actually costs, what it contains, when it shipped, and how long it takes to receive one in mid-2026. It draws on NVIDIA's own product pages, technical documentation, and financial disclosures; named OEM and reseller price quotes from Exxact, PNY, aiserver.eu, servermall.com, and American Compute; independent GPU cloud analyses from acecloud.ai, server-parts.eu, and getdeploying.com; and market commentary from Spheron Network and iFactoryApp.Where sources disagree, meaning total memory capacity, cooling requirements, or exact lead-time windows, this report presents the discrepancy directly rather than picking a single number to appear more authoritative than the underlying market actually is. The remainder of the report covers base system pricing, enterprise and rack-scale pricing, cloud rental economics, a full DGX B200 comparison, the release and availability timeline, market positioning against competing accelerators, quantitative benchmark and revenue data, three named deployments, and forward-looking implications for buyers weighing a DGX B300 purchase in the second half of 2026.

Base System Pricing: The DGX B300 8-GPU Platform

The core DGX B300 configuration is a single 10U chassis housing **eight NVIDIA Blackwell Ultra SXM GPUs**, dual **Intel Xeon 6776P** processors, and NVIDIA's integrated networking and software stack (Source: nvidia.com). Because NVIDIA sells the system through partners rather than direct retail, the closest thing to an anchor price comes from named integrator commentary rather than a manufacturer price sheet. iFactoryApp, an on-premises AI integration consultancy, states that the DGX B300 8-GPU system is priced in the $300,000 to $350,000 band as of Q1 2026, implying $37,500 to $43,750 per GPU (Source: ifactoryapp.com). GPU rental marketplace Spheron Network quotes a noticeably wider band, stating that **"a full DGX B300 system (8x B300 GPUs) costs around $400,000-500,000"** (Source: spheron.network). Independent server integrator American Compute provides a useful cross-check at the broader HGX-class level (the baseboard platform underlying both DGX systems and third-party OEM servers), reporting that **"OEM pricing for a complete 8-GPU HGX server ranged from $200,000-$300,000 for the H100 generation and $250,000-$400,000 for B200/B300-class systems, depending on configuration and vendor"**, and attributing the spread to **"OEM choices around cooling, storage, NICs, and support tiers rather than the GPU baseboard itself"** (Source: www.amcompute.com) (Source: www.amcompute.com). Buyers should treat $300,000 to $350,000 as the base-configuration floor and $400,000 to $500,000 as a realistic ceiling once premium support, extended RAM, or NVIDIA AI Enterprise software bundles are added.

European system integrators provide a useful independent cross-check because their storefronts publish itemized, real-time pricing rather than aggregated estimates. Czech-based reseller aiserver.eu lists a DGX B300 configuration, **two 64-core Intel Xeon 6776P processors, eight B300 SXM GPUs, 2 TB RAM, 30 TB of NVMe storage, and three years of support**, at a starting price of **€535,000**, with a published range of **€535,000 to €714,000** depending on support length and NVIDIA AI Enterprise software add-ons (Source: aiserver.eu) (Source: aiserver.eu). Lithuania-based Servermall lists a comparably specified system, **two Xeon 6776P processors, 2,000 GB of DDR5 RAM, and eight B300 SXM GPUs**, at a much higher headline price of **€1,404,981**, discounted to **€1,161,141 plus €243,840 VAT** (Source: servermall.com) (Source: servermall.com). The roughly

two-fold spread between these two EU quotes for similarly specified hardware, on top of the wider US-market spread already documented, illustrates that DGX B300 pricing is genuinely non-standardized: buyers should always request itemized quotes from at least two or three NVIDIA partners rather than anchoring on any single published figure.

For buyers who do not need a full 8-GPU system, NVIDIA also sells the **Blackwell Ultra B300** GPU as a standalone module through OEM channels. Spheron Network states that **"purchase price for a single B300 GPU is approximately $53,000"** as of its July 5, 2026 pricing update (Source: spheron.network). These standalone figures exclude networking, liquid- or air-cooling infrastructure, and installation labor, none of which are trivial at the system's rated power draw. *Table 1* below summarizes the DGX B300's core hardware specification as documented across NVIDIA's own product page, technical user guide, and official datasheet.

SPECIFICATION	NVIDIA DGX B300 (AS DOCUMENTED)
GPUs	8x NVIDIA Blackwell Ultra SXM GPUs (Source: nvidia.com)
CPU	2x Intel Xeon 6776P processors, 128 cores total (Source: nvidia.com) (Source: exxactcorp.com)
Total GPU memory	Listed as 2.1 TB on NVIDIA's product page; listed as 2.3 TB (8 x 288 GB HBM3e) in the technical user guide and official datasheet (Source: nvidia.com) (Source: docs.nvidia.com)
Performance	144 PFLOPS FP4 sparse / 108 PFLOPS FP4 dense; 72 PFLOPS FP8 (Source: nvidia.com)
Memory bandwidth	64 TB/s total system bandwidth (Source: exxactcorp.com) (Source: acecloud.ai)
NVLink	2x NVLink Switch, 14.4 TB/s aggregate bandwidth (Source: nvidia.com)
Networking	8x OSFP ports via ConnectX-8 (up to 800 Gb/s each); 2x dual-port BlueField-3 DPU (up to 400 Gb/s each) (Source: nvidia.com)
System memory	2 TB default, configurable up to 4 TB (Source: docs.nvidia.com)
Storage	2x 1.9 TB NVMe M.2 (OS); 8x 3.84 TB NVMe E1.S (data) (Source: nvidia.com)
Power draw	Approximately 14 to 14.5 kW (Source: nvidia.com) (Source: docs.nvidia.com)
Chassis	10 rack units (10U); AC/PDU or DC/busbar power options (Source: nvidia.com) (Source: docs.nvidia.com)
Support	Three-year business-standard hardware and software support included (Source: nvidia.com)

The most important reader takeaway from Table 1 is the memory-capacity discrepancy in the second row, which recurs across NVIDIA's own web properties. NVIDIA's flagship product page lists **"Total GPU Memory | 2.1 TB"** in both its main specification block and its "Quick Specs" summary (Source: nvidia.com), and independent cloud provider acecloud.ai's own review of NVIDIA's DGX B300 spec page repeats the same figure, noting **"NVIDIA lists 2.1 TB total GPU memory and up to 14.4 TB/s NVLink bandwidth on its DGX B300 spec page"** (Source: acecloud.ai). Yet NVIDIA's own technical user guide states unambiguously that **"GPU memory | 8 x 288 GB = 2.3 TB total"** (Source: docs.nvidia.com), the official NVIDIA datasheet describes **"2.3TB of GPU memory space"** (Source: na.ingrammicro.com), and NVIDIA's own GTC press release for the DGX SuperPOD states that **"each system provides 2.3TB of HBM3e memory"** (Source: nvidianews.nvidia.com). Independent technical analyst Glenn Klockwood resolves the ambiguity by noting that of the 288 GB of HBM3e physically present on each B300 GPU, only about **"270 GB usable for B300"** after reserved capacity is excluded, which sums to roughly 2.16 TB across eight GPUs, close enough to the rounded 2.1 TB figure to explain the gap as raw versus usable capacity rather than a factual error (Source: glennklockwood.com). Buyers sizing a model deployment against the DGX B300's memory ceiling should plan against the more conservative 2.1 TB figure.

Enterprise-Scale Pricing: DGX SuperPOD, GB300 NVL72, and Volume Deployments

Enterprises scaling beyond a handful of DGX B300 nodes typically move to NVIDIA's **DGX SuperPOD** reference architecture, a turnkey, multi-rack cluster design that NVIDIA describes as **"leadership-class AI infrastructure purpose-built for the unique demands of AI"** and which can be configured with either DGX B300 or the larger rack-scale DGX GB300 system (Source: nvidia.com). NVIDIA's GTC announcement for the Blackwell Ultra DGX SuperPOD describes DGX GB300 systems as combining **"36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell Ultra GPUs"** in a liquid-cooled, rack-scale design capable of delivering **"up to 70x more AI performance than AI factories built with NVIDIA Hopper systems and 38TB of fast memory"** (Source: nvidianews.nvidia.com). The equivalent rack-scale product sold to cloud providers, the **GB300 NVL72**, packages 72 Blackwell Ultra GPUs with 36 Grace CPUs into a single NVLink domain; NVIDIA's March 2025 platform announcement states the GB300 NVL72 **"delivers 1.5x more AI performance than the NVIDIA GB200 NVL72, as well as increases Blackwell's revenue opportunity by 50x for AI factories, compared with those built with NVIDIA Hopper"** (Source: nvidianews.nvidia.com).

Because a full GB300 NVL72 rack is sold almost exclusively to hyperscalers and large GPU cloud providers rather than through general reseller channels, published pricing is sparser and less reliable than for the single-chassis DGX B300. Independent GPU architecture researcher Glenn Klockwood reports that **"a single GB300 NVL72 rack is rumored to cost between $3.7 and $4 million"**, citing a reported Apple order for an estimated $1 billion of GB300 NVL72 hardware as a reference point for scale (Source: glennklockwood.com). At roughly $3.7 million to $4 million per 72-GPU rack, the effective per-GPU price of a GB300 NVL72 (approximately $51,000 to $56,000) is broadly comparable to the standalone B300 GPU price of about $53,000 documented for individual-unit purchases, suggesting the rack-scale premium is concentrated in Grace CPU silicon, NVLink switching, and the mandatory liquid-cooling infrastructure rather than in the GPUs themselves (Source: glennklockwood.com).

For enterprises that do not want to own physical infrastructure at all but still want dedicated, managed DGX capacity, NVIDIA partners with data center operator **Equinix** to offer **NVIDIA Instant AI Factory**, described in NVIDIA's press release as **"a managed service featuring the Blackwell Ultra-powered NVIDIA DGX SuperPOD with NVIDIA Mission Control software"**, delivered inside preconfigured Equinix facilities in **45 markets around the world** (Source: nvidianews.nvidia.com). Equinix itself frames the offering as removing "months of pre-deployment infrastructure planning" for enterprises that lack in-house expertise to design, build, and operate high-density liquid-cooled data center space (Source: blog.equinix.com). Neither NVIDIA nor Equinix has published fixed per-rack or per-month pricing for Instant AI Factory publicly; enterprises evaluating this route should expect a custom quote scoped to committed capacity and contract length, similar to how hyperscale colocation pricing is typically negotiated.

Enterprises purchasing at true volume, dozens to hundreds of DGX B300 or GB300 NVL72 units for a dedicated AI factory, negotiate custom pricing directly with NVIDIA or its OEM partners (Dell, HPE, Lenovo, Supermicro, and others) rather than relying on list-style reseller quotes (Source: www.amcompute.com). The clearest public illustration of this scale is pharmaceutical company Eli Lilly's DGX SuperPOD deployment, covered in the Case Studies section below, which used **1,016 NVIDIA Blackwell Ultra GPUs** (Source: blogs.nvidia.com), a scale at which per-GPU pricing is understood to fall meaningfully below the effective rates implied by single-system or single-GPU list quotes, though neither Lilly nor NVIDIA has disclosed the transaction's dollar value.

Cloud and Usage-Based Pricing: Renting DGX B300 and HGX B300 Capacity

For teams that do not want to commit $300,000-plus in capital expenditure, renting DGX-class or HGX-class B300 capacity by the GPU-hour is increasingly viable as more cloud providers bring Blackwell Ultra online. GPU cloud pricing aggregator getdeploying.com, which tracks live listings across dozens of providers, reports that **"listings for the B300 reach $18.00/hr, often reflecting a premium for high availability. However, you might be able to find available instances from as low as $3.27/hr per GPU (36-month reservation)"** (Source: getdeploying.com). *Table 2* below summarizes a representative sample of named on-demand B300 listings gathered from getdeploying.com's live comparison table in July 2026; because that table is a frequently updated, dynamically rendered listing rather than static prose, the individual row citations below point to the comparison page as a whole rather than to a fragment-addressable quote.

PROVIDER	CONFIGURATION	PRICE PER GPU-HOUR	BILLING
DigitalOcean	8x B300 NVL, 2.3 TB total VRAM	$5.65	12-month reserved (Source: getdeploying.com)
Runpod (community cloud)	1x B300 SXM, 288 GB VRAM	$6.94	On-demand (Source: getdeploying.com)
Runpod (secure cloud)	1x B300 SXM, 288 GB VRAM	$7.39	On-demand (Source: getdeploying.com)
Nebius	1x B300 SXM5, 288 GB VRAM	$7.85	On-demand (Source: getdeploying.com)
Lyceum	1x B300, 288 GB VRAM	$7.99	On-demand (Source: getdeploying.com)
Sesterce	1x B300, 288 GB VRAM	$8.70	On-demand (Source: getdeploying.com)
Spheron Network	B300 SXM6 (6 configurations)	$9.16	On-demand, per-minute billing (Source: spheron.network)
AWS	8x B300, 2.1 TB total VRAM (p6-b300.48xlarge)	$17.80	On-demand (Source: getdeploying.com)
Oracle Cloud Infrastructure	4x B300 NVL72, 1.1 TB total VRAM	$18.00	On-demand (Source: getdeploying.com)

Table 2 shows a roughly threefold spread among mainstream on-demand listings, from $5.65 per GPU-hour on a reserved DigitalOcean contract up to $18.00 per GPU-hour on Oracle Cloud Infrastructure, with most specialty GPU clouds (Runpod, Nebius, Lyceum, Sesterce, Spheron) clustering between roughly $7 and $9 per GPU-hour and the large hyperscalers (AWS, Oracle) pricing noticeably higher, consistent with the pattern documented for older GPU generations where hyperscaler list pricing bundles managed networking, enterprise service-level agreements (SLAs), and broader platform integration into the hourly rate (Source: tech-insider.org). For comparison, Spheron notes that older-generation GPUs remain considerably cheaper on its own platform: **"on Spheron right now an H200 SXM runs about $3.70/hr on-demand and an H100 starts around $2.01/hr (PCIe) to $3.92/hr (SXM5)"** (Source: spheron.network), meaning renting a single B300 GPU on-demand currently costs roughly 1.8 to 2.6 times as much per hour as renting an H100 or H200 on the same marketplace, a premium buyers should weigh against the B300's throughput and memory advantages covered in the Data Analysis section.

Providers and analysts broadly expect this premium to compress over time as Blackwell Ultra supply catches up with demand, following the pattern set by the Hopper generation. Spheron notes that **"the H100 went from $8/hr in early 2024 to under $3/hr in 2026"** and predicts B300 pricing should follow a similar arc as **"Vera Rubin (R100) volume shipments pull frontier training and inference demand off Blackwell Ultra"** over the following six to twelve months (Source: spheron.network). For enterprises weighing rent versus buy, iFactoryApp estimates that on-premise ownership reaches break-even against cloud rental within roughly 18 months at moderate utilization above ~60 percent, though it adds the important caveat that regulated industries with data-sovereignty requirements should weight the decision toward on-premises ownership regardless of the pure cost math (Source: ifactoryapp.com).

DGX B300 vs DGX B200: Specifications and Value Comparison

The DGX B300's most direct point of comparison is its immediate predecessor, the **DGX B200**, which shipped on the original (non-Ultra) Blackwell architecture. NVIDIA's own product page frames the generational jump succinctly: DGX B300 **"boosts dense FP4 performance by 1.5x and attention performance by 2x over DGX B200"** (Source: nvidia.com). NVIDIA reseller Exxact Corporation, an NVIDIA Elite Partner, published a detailed side-by-side comparison confirming that DGX B300 memory **"has grown from 1.4TB to 2.3TB, allowing larger models to run in memory without partitioning or complex workarounds"** and that networking bandwidth expanded from 0.8 TB/s on DGX B200 to 1.6 TB/s on DGX B300 (Source: exxactcorp.com) (Source: exxactcorp.com). GPU parts reseller server-parts.eu independently corroborates the memory jump at the chip level, listing **192 GB HBM3e for the B200 versus up to 288 GB for the B300**, and reports the DGX-system-level total for DGX B300 at approximately 2.3 TB versus approximately 1.5 TB for DGX B200 (Source: www.server-parts.eu) (Source: www.server-parts.eu). Note that a third

independent source, acecloud.ai, cites a slightly different B200 per-GPU figure of **180 GB HBM3e** rather than the 192 GB figure used by server-parts.eu and Exxact, illustrating that even basic B200 memory specifications are not perfectly consistent across secondary sources (Source: acecloud.ai).

The upgrade is not unambiguous, however. Exxact's comparison flags that **"NVIDIA DGX B300 has lower INT8 and FP64 performance compared to the NVIDIA DGX B200, which will impact INT8 optimized inferencing and double-precision imperative scientific computing"**, reporting FP64 throughput falling from 296 teraFLOPS on DGX B200 to just 10 teraFLOPS on DGX B300 (Source: exxactcorp.com). iFactoryApp reports a similar but not identical FP64 collapse, from **37 teraFLOPS down to 1.25 teraFLOPS per GPU** (Source: ifactoryapp.com), and Glenn Klockwood's independent architecture notes cite yet another pair of figures, **1.2 teraFLOPS on B300 versus 37 teraFLOPS on B200** at the per-GPU level (Source: glennklockwood.com). The exact figures differ by source (system-level versus per-GPU accounting likely explains some of the gap), but all three independently arrive at the same directional conclusion: NVIDIA traded away the vast majority of Blackwell Ultra's double-precision compute to maximize FP4 and FP8 inference throughput, meaning the DGX B300 is a poor fit for traditional high-performance computing (HPC) workloads like computational fluid dynamics or molecular dynamics that require FP64 precision, despite being the stronger choice for LLM inference and reasoning workloads. Typical TDP (thermal design power, a proxy for heat and power draw) also rose from roughly 1,000W to approximately 1,400W per GPU between the two generations, per server-parts.eu's comparison (Source: www.server-parts.eu). *Table 3* summarizes the comparison.

DIMENSION	DGX B200	DGX B300	ADVANTAGE
GPU architecture	Blackwell	Blackwell Ultra	B300 (Source: nvidia.com)
GPU memory (system total)	~1.4-1.5 TB HBM3e	~2.1-2.3 TB HBM3e	B300, +50% to +64% (Source: exxactcorp.com) (Source: www.server-parts.eu)
Per-GPU TDP	~1,000W	~1,400W	B200 lower (Source: www.server-parts.eu)
Dense FP4 performance	Baseline	1.5x B200	B300 (Source: nvidia.com)
Attention-layer performance	Baseline	2x B200	B300 (Source: nvidia.com)
Networking bandwidth	0.8 TB/s	1.6 TB/s	B300 (Source: exxactcorp.com)
FP64 double precision	Higher (tens of TFLOPS)	Sharply reduced	B200 (Source: exxactcorp.com)
Cooling	Liquid cooling recommended, air-cooled variants exist	NVIDIA describes standard DGX B300 as air-cooled; some integrators call liquid cooling mandatory at the HGX/rack-scale level	Mixed, see discussion below (Source: nvidianews.nvidia.com) (Source: spheron.network)
System-level price (2026)	~$280,000-$320,000	~$300,000-$350,000 (up to $400,000-$500,000 at some resellers)	B200 cheaper (Source: ifactoryapp.com)
Lead time (2026)	~8-16 weeks	~8-20 weeks depending on source	B200 slightly faster (Source: ifactoryapp.com)

The cooling row deserves explicit discussion because it is one of the report's clearest cross-source contradictions. NVIDIA's own GTC press release for the Blackwell Ultra DGX SuperPOD states directly that **"air-cooled NVIDIA DGX B300 systems harness the NVIDIA B300 NVL16 architecture to help data centers everywhere meet the computational demands of generative and agentic AI applications"** (Source:

nvidianews.nvidia.com), and NVIDIA's official datasheet, hosted by distributor Ingram Micro, echoes that **"DGX B300's air-cooled design allows for easy integration into existing data center infrastructure"** (Source: na.ingrammicro.com). Yet GPU rental marketplace Spheron Network states flatly that **"at 1,400W per GPU (11.2 kW for an 8-GPU system, before CPUs and networking), air cooling isn't viable. The DGX B300 and HGX B300 require direct liquid cooling (DLC)"** (Source: spheron.network), and iFactoryApp similarly lists direct liquid cooling as **"mandatory (DLC)"** for DGX B300 versus merely **"recommended"** for DGX B200 (Source: ifactoryapp.com). This report cannot resolve the contradiction definitively from public sources, but the most plausible explanation is that the standard 10U DGX B300 SKU documented in NVIDIA's own physical specifications, which lists fan airflow figures of up to 1,500 cubic feet per minute (CFM) rather than a liquid-loop specification, is genuinely air-cooled by design (Source: docs.nvidia.com), while higher-density HGX B300 baseboard integrations and the rack-scale GB300 NVL72 (which NVIDIA itself confirms is liquid-cooled (Source: nvidianews.nvidia.com) genuinely require direct liquid cooling because of higher aggregate rack power density. Buyers should confirm the cooling requirement directly with their chosen OEM before finalizing a data center power and cooling budget.

Release Timeline, Lead Times, and Availability

NVIDIA first announced the Blackwell Ultra platform, including both the DGX B300 and the rack-scale GB300 NVL72, at its GTC conference on **March 18, 2025**, with founder and CEO Jensen Huang stating the platform was designed as **"a single versatile platform that can easily and efficiently do pretraining, post-training and reasoning AI inference"** (Source: nvidianews.nvidia.com). At that announcement, NVIDIA guided that **"Blackwell Ultra-based products are expected to be available from partners starting from the second half of 2025"** (Source: nvidianews.nvidia.com), and its follow-on DGX SuperPOD press release similarly stated systems were **"expected to be available from partners later this year"** (2025) (Source: nvidianews.nvidia.com). Reseller Exxact Corporation confirmed on October 1, 2025, that the **"NVIDIA DGX B300 is available to order from Exxact and will start shipping later this year, Q4 2025"** (Source: exxactcorp.com), consistent with NVIDIA's original guidance.

Actual volume shipments, however, appear to have slipped slightly into early 2026. Spheron Network states directly that **"NVIDIA shipped the B300 (officially 'Blackwell Ultra') in January 2026"**, adding that **"the DGX B300 user guide went live on January 20th, and cloud providers started listing instances within weeks"** (Source: spheron.network). This is independently corroborated by NVIDIA's own documentation: the DGX B300 User Guide PDF is dated **"Jan 20, 2026"** on its cover page (Source: docs.nvidia.com), and NVIDIA's product page currently states the system is **"Available Now"** with systems described as **"Shipping Now"** as of the report's July 2026 publication date (Source: nvidia.com) (Source: nvidia.com). GPU pricing aggregator getdeploying.com lists the DGX B300's broader release window as **"Q4 2025"** in its own specifications summary, splitting the difference between NVIDIA's original guidance and Spheron's January 2026 volume-shipping date (Source: getdeploying.com).

Lead times, meaning the time between placing an order and receiving hardware, have lengthened somewhat as Blackwell Ultra demand has outpaced early supply. Spheron reported in its most recent update that **"lead times are currently 8-12 weeks"** for DGX B300 systems ordered directly through NVIDIA's partner network, adding that this **"requires liquid cooling infrastructure and NVIDIA-certified installation"** (Source: spheron.network). iFactoryApp's more recent (May 2026) comparison quotes a longer window, stating that as of April 2026 DGX B300 lead times run 12 to 20 weeks compared against DGX B200 at 8 to 16 weeks, and separately notes that **"B200 backlog estimated at 3.6M units through mid-2026. B300 shipping since January 2026 with faster cloud ramp"** (Source: ifactoryapp.com) (Source: ifactoryapp.com). Taken together, buyers ordering a DGX B300 in mid-to-late 2026 should budget for a realistic delivery window of **two to five months**, with the shorter end of that range applying to smaller orders through well-stocked partners and the longer end applying to larger custom configurations or partners further down NVIDIA's allocation queue. Marketplace listings such as Uvation's confirm this uncertainty directly to prospective buyers, instructing customers to **"contact our sales team for bulk order inquiries and lead time details"** rather than publishing a fixed delivery date (Source: marketplace.uvation.com).

Comparative Context and Market Positioning

The DGX B300 competes for enterprise AI infrastructure budget against three distinct categories of alternative: the prior-generation DGX B200 already covered above, Hopper-generation systems (H100 and H200) still widely deployed and available at lower cost, and NVIDIA's own rack-scale GB300 NVL72 aimed at the largest hyperscale buyers. Independent cloud provider acecloud.ai's own comparison table places the B300 chip clearly ahead of H200 on memory and bandwidth: **288 GB HBM3e for B300 versus 180 GB for B200 and 141 GB for H200** (Source: acecloud.ai), with per-GPU HBM bandwidth reaching **"up to 8 TB/s"** for B300 versus **4.80 TB/s** for H200 (Source: acecloud.ai). acecloud.ai frames the practical implication plainly: NVIDIA's own materials describe Blackwell Ultra's 288 GB of HBM3e as **"3.6 times the on-package memory of H100 and 50% more than the base Blackwell generation,"** which the firm says means **"more of the model can stay on GPU without paging or sharding across too many nodes"** and **"more KV cache can remain on GPU during long context inference"** (Source: acecloud.ai) (Source: acecloud.ai). For workloads that fit comfortably inside a smaller memory footprint, GPU marketplace Spheron Network's own guidance recommends sticking with the cheaper H200, noting that **"the H200 delivers strong performance at around $3.70/hr. No reason to pay the B300 premium for memory you won't use"** (Source: spheron.network).

Against its own sibling product, the GB300 NVL72, the DGX B300 occupies a deliberately different market position, sitting between HGX-class OEM baseboard servers and full rack-scale systems in NVIDIA's own platform hierarchy. Server-platform analyst American Compute explains this hierarchy directly: **"MGX is a general modular reference architecture... OEMs build servers around HGX, which is NVIDIA's proprietary GPU baseboard... DGX is NVIDIA's own complete server using the same HGX baseboard, plus a fixed CPU, memory, and support stack chosen by NVIDIA"** (Source: www.amcompute.com). The firm also notes a cost structure detail relevant to buyers comparing DGX against building a custom HGX server: **"the GPUs on the HGX baseboard are the most expensive component in the server, accounting for roughly 60-70% of total hardware cost in a typical 8-GPU deployment"** (Source: www.amcompute.com), meaning even the wide reseller price spread documented in Table 1 is unlikely to reflect NVIDIA discounting the silicon itself so much as OEMs varying markup on the remaining 30 to 40 percent of the bill of materials. Networking hardware vendor NADDOD's technical comparison of the interconnect stack across the family confirms the architectural distinction between generations: DGX B300 memory capacity per GPU is 288 GB versus 192 GB on B200, and network card bandwidth doubles from ConnectX-7's 400 Gbps to ConnectX-8's 800 Gbps generation over generation (Source: www.naddod.com).

At the ecosystem level, the number of vendors bringing Blackwell Ultra hardware to market signals a healthy, competitive supply chain rather than a NVIDIA-only channel. NVIDIA's original platform announcement lists server partners **Cisco, Dell Technologies, Hewlett Packard Enterprise, Lenovo and Supermicro**, alongside **Aivres, ASRock Rack, ASUS, Eviden, Foxconn, GIGABYTE, Inventec, Pegatron, Quanta Cloud Technology, Wistron and Wiwynn**, as expected to ship Blackwell Ultra-based servers (Source: nvidianews.nvidia.com), while cloud service providers **Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure**, together with GPU-focused clouds **CoreWeave, Crusoe, Lambda, Nebius, Nscale, Yotta and YTL**, are named as launch partners for Blackwell Ultra instances (Source: nvidianews.nvidia.com). This breadth of supply, roughly a dozen server OEMs and a dozen cloud providers, is a meaningful factor keeping DGX B300 pricing competitive rather than monopolistic, though it does not eliminate the wide reseller price dispersion documented earlier in this report.

Data Analysis and Evidence

Quantifying the DGX B300's real-world performance advantage requires third-party, auditable benchmark data rather than vendor marketing claims alone. The industry-standard **MLPerf Inference** benchmark, administered by MLCommons, gave Blackwell Ultra its debut in the version 5.1 round; NVIDIA's technical blog reports that **"NVIDIA submitted results in the available category using the GB300 NVL72 rack-scale system, the first-ever MLPerf submissions using the Blackwell Ultra architecture"** (Source: developer.nvidia.com). On the DeepSeek-R1 reasoning-model benchmark, per-GPU throughput climbed from **1,253 tokens per second on an 8-GPU DGX H200** to **4,024 tokens per second on GB200 NVL72** to **5,842 tokens per second on GB300 NVL72** in the offline scenario (Source: developer.nvidia.com). NVIDIA summarizes the generational jump as **"compared to the GB200 NVL72 submission, GB300 NVL72 delivered 45% higher performance per GPU on the new DeepSeek-R1 benchmark in the offline scenario and 25% in the server scenario"**, and states that against unverified Hopper-based measurements, **"Blackwell Ultra delivered about 5x higher throughput per GPU"** (Source: developer.nvidia.com). NVIDIA's own DGX B300 product page adds a more recent MLPerf Inference v6.0 (April 2026) data point, stating that **"systems powered by NVIDIA Blackwell Ultra GPUs delivered the highest throughput across the widest range of models and scenarios. On DeepSeek-R1, Blackwell Ultra systems delivered 2.5 million tokens per second, up to 2.7x higher token throughput compared to Blackwell Ultra debut submissions just six months prior"** (Source: nvidia.com). Separately, NVIDIA's HGX platform page claims HGX B300 **"delivers up to 2.6x higher training performance for large language models such as DeepSeek R1"** relative to the prior generation, tied to its expanded memory and NVLink bandwidth (Source: acecloud.ai).

On the cost-efficiency side, NVIDIA cites independent research firm SemiAnalysis's InferenceX benchmark suite to claim Blackwell Ultra delivers **"up to 50x higher throughput per megawatt and up to 35x lower cost per token than NVIDIA Hopper for low-latency agentic workloads"** as of Q1 2026 (Source: nvidia.com), and separately claims a specific unit-economics figure: **"NVIDIA Blackwell Ultra delivers AI Inference at $0.24 per million tokens at 102 TPS/user on DeepSeek-R1 using NVIDIA Dynamo TensorRT-LLM and MTP, according to SemiAnalysis InferenceX benchmarks as of April 2026"** (Source: nvidia.com). Spheron's own cost-per-token modeling, while cautioned by the company itself as **"early estimates based on NVIDIA's published improvement ratios"** rather than independently measured (Source: spheron.network), arrives at a directionally similar conclusion using its own July 2026 cloud pricing: relative to an H100 SXM baseline, B300 on-demand comes in at roughly **0.51x** the cost per token at FP8 precision and **0.34x** at FP4 precision (Source: spheron.network).

The demand backdrop behind this pricing sits inside NVIDIA's own financial disclosures. For the first quarter of fiscal year 2027, ended April 26, 2026, NVIDIA reported **"record revenue of $81.6 billion, up 85% from a year ago"** and **"record Data Center revenue of $75.2 billion, up 92% from a year ago"** (Source: nvidianews.nvidia.com) (Source: nvidianews.nvidia.com). Within that segment, **"Data Center compute revenue was a record $60.4 billion, up 77% from a year ago and up 18% sequentially,"** while **"Data Center networking revenue was a record $14.8 billion, up 199% from a year ago and up 35% sequentially"** (Source: nvidianews.nvidia.com). While NVIDIA does not break out DGX B300-specific revenue in its public filings, the scale and growth rate of Data Center revenue, most of it Blackwell and Blackwell Ultra silicon, corroborate the tight allocation and

elevated lead times documented earlier in this report: demand for Blackwell Ultra capacity is outstripping available supply broadly enough to show up materially in NVIDIA's consolidated financial results. *Table 4* consolidates the pricing survey data gathered across resellers and marketplaces referenced throughout this report.

CHANNEL	CONFIGURATION	PRICE (AS DOCUMENTED)	AS OF
iFactoryApp anchor	8-GPU DGX B300 system	$300,000-$350,000	Q1 2026 (Source: ifactoryapp.com)
Spheron Network / reseller high end	8-GPU DGX B300 system	$400,000-$500,000	July 2026 (Source: spheron.network)
American Compute (HGX-class OEM)	8-GPU HGX B200/B300-class server	$250,000-$400,000	March 2026 (Source: www.amcompute.com)
aiserver.eu	8-GPU, 2 TB RAM, 30 TB NVMe, 3-yr support	€535,000-€714,000	July 2026 (Source: aiserver.eu)
Servermall.com	8-GPU, 2,000 GB DDR5 RAM	€1,161,141 + €243,840 VAT	July 2026 (Source: servermall.com)
Standalone B300 GPU	Single Blackwell Ultra GPU	~$53,000	July 5, 2026 (Source: spheron.network)
GB300 NVL72 rack	72-GPU rack-scale system	~$3.7M-$4.0M (rumored)	2025-2026 (Source: glennklockwood.com)
Named cloud on-demand (low to high)	Single B300 GPU-hour	$5.65-$18.00/hr	July 2026 (Source: getdeploying.com)

Read together, Table 4 shows roughly a **six-fold spread** between the cheapest and most expensive documented paths to DGX B300-class hardware or capacity, from a $53,000 standalone GPU purchase up to a fully configured, VAT-inclusive European system quote north of $1.5 million equivalent. That spread is not evidence of a broken market so much as evidence of a genuinely heterogeneous one, where system RAM, storage, software licensing (NVIDIA AI Enterprise), support duration, VAT treatment, and reseller margin all move independently of the underlying GPU cost.

Case Studies and Real-World Examples

Eli Lilly: DGX SuperPOD for Drug Discovery

Pharmaceutical company **Eli Lilly** provides the most concrete public example of DGX B300 deployed at meaningful scale. NVIDIA's blog reports that Lilly is **"deploying the largest, most powerful AI factory wholly owned and operated by a pharmaceutical company, the world's first NVIDIA DGX SuperPOD with DGX B300 systems"**, announced at NVIDIA GTC Washington, D.C. and **"built with 1,016 NVIDIA Blackwell Ultra GPUs"** (Source: blogs.nvidia.com) (Source: blogs.nvidia.com). NVIDIA quantifies the resulting compute scale as **"over 9,000 petaflops of AI performance"**, or more than 9 quintillion math operations per second (Source: blogs.nvidia.com). The infrastructure feeds **Lilly TuneLab**, described as a federated-learning platform built on **"$1 billion worth of Lilly's proprietary data"**, which lets partner biotechs tap Lilly's models without exposing their own data (Source: blogs.nvidia.com). The deployment sits inside a wider **$50 billion commitment** to US manufacturing and R&D expansion, including a proposed **$4.5 billion** facility in Indiana called the **Lilly Medicine Foundry**, expected to help create roughly **13,000 high-wage manufacturing and construction jobs** (Source: blogs.nvidia.com) (Source: blogs.nvidia.com). Neither Lilly nor NVIDIA has disclosed the specific dollar value of the DGX SuperPOD purchase itself, but applying this report's documented per-GPU system-price range of roughly $37,500 to $43,750 to 1,016 GPUs implies a notional hardware value in the range of **$38 million to $44 million** before volume discounting, networking, and facility costs, illustrating the scale gap between a single DGX B300 chassis and an enterprise-grade AI factory.

CoreWeave: First Cloud Deployment of GB300 NVL72

GPU cloud provider **CoreWeave** announced on July 3, 2025, that it had become **"the first AI cloud provider to deploy the latest NVIDIA GB300 NVL72 systems for customers, with plans to significantly scale deployments worldwide"** (Source: www.coreweave.com). CoreWeave quantifies the customer-facing benefit as **"up to a 10x boost in user responsiveness, a 5x improvement in throughput per watt compared to the previous generation NVIDIA Hopper architecture, and a 50x increase in output for reasoning model inference"** (Source: www.coreweave.com). The company built the initial deployment in collaboration with **Dell, Switch, and Vertiv**, and integrated it directly with its own CoreWeave Kubernetes Service and Slurm-on-Kubernetes scheduling stack (Source: www.coreweave.com). CoreWeave went on to set additional Blackwell Ultra performance records, publishing results titled **"CoreWeave Sets New AI Training Records in MLPerf Training v6.0, Training DeepSeek-V3 in Approximately Two Minutes"** on the largest GB300 NVL72 cluster entered in that benchmark round (Source: www.coreweave.com). CoreWeave's GB300 rollout illustrates the cloud-rental path documented earlier in this report: enterprises that need Blackwell Ultra-class capacity without owning hardware can access it through a growing set of neocloud and hyperscale providers rather than purchasing a DGX B300 outright.

Equinix and Alembic: Colocated DGX Infrastructure for a Mid-Market SaaS Company

Not every DGX buyer is a pharmaceutical giant or a public GPU cloud. Data center operator **Equinix**, which partners with NVIDIA to offer the **NVIDIA Instant AI Factory** managed service described earlier in this report, documents a smaller-scale, mid-market example in marketing intelligence SaaS company **Alembic**. Equinix's blog describes how **"the company needed to bring a private AI stack closer to its customers to enable higher-performance inference at the edge"** and **"deployed DGX infrastructure in an Equinix AI-ready colocation data center to enhance the foundation models brought in from cloud providers with its own proprietary data"** (Source: blog.equinix.com) (Source: blog.equinix.com). Alembic leveraged Equinix's network fabric for high-speed connectivity to its own customers and partners rather than routing all inference traffic back through a public cloud region (Source: blog.equinix.com). This case illustrates the colocation middle path between buying a DGX B300 for an in-house data center and renting bare GPU-hours from a public cloud: Equinix's model lets an enterprise own or lease dedicated DGX hardware while outsourcing the power, cooling, and physical security requirements that make DLC-class systems expensive and complex to host in-house.

Implications and Future Directions

Three forward-looking dynamics should shape how buyers think about DGX B300 pricing over the remainder of 2026 and into 2027. First, the platform's successor is already on NVIDIA's public roadmap: iFactoryApp's roadmap chart places **Vera Rubin**, built on TSMC 3nm process technology with 288 GB of next-generation HBM4 memory and 13 TB/s of bandwidth, as an **"H2 2026 · Announced"** milestone, with a further **"Rubin Ultra"** generation targeted for 2027 (Source: ifactoryapp.com). Spheron independently confirms this timeline, noting that Rubin's R100 chip **"is expected to reach cloud providers in H2 2026"** (Source: spheron.network). Buyers placing large DGX B300 orders in late 2026 should factor in that Blackwell Ultra will likely be a two-to-three-year architecture rather than NVIDIA's current flagship indefinitely, which affects both resale value planning and the urgency of any "wait for the next generation" deferral strategy.

Second, infrastructure readiness, not price, is emerging as the binding constraint for many prospective buyers. The DGX B300's roughly 14 to 14.5 kilowatt power draw is **"roughly 2x what an equivalent H100 DGX system draws"** according to Spheron (Source: spheron.network), and buyers discussing real-world deployments on Reddit's r/HomeDataCenter community echo this concern in plain terms, with one thread on deploying 1.4 kW-class GPUs noting **"what's been killing us lately is the supporting infrastructure"** rather than the GPUs themselves (Source: www.reddit.com). This dynamic is precisely why colocation and managed-service paths like Equinix's Instant AI Factory and CoreWeave's cloud offering, both covered in the Case Studies section, are gaining traction relative to pure on-premises ownership: they let buyers absorb the DGX B300's power and cooling requirements as an operating expense rather than a capital project. Reddit's r/LocalLLM community, discussing a hypothetical individual-scale DGX B300 purchase, was broadly skeptical that a single-node purchase makes sense outside an enterprise use case, with one commenter advising simply **"this is crazy. Just rent it when you need it. vast.ai or runpod"** and another suggesting that anyone with several million dollars to spend **"should use a small fraction of that to hire an engineer to help you sort this out"** (Source: www.reddit.com) (Source: www.reddit.com), reinforcing this report's own guidance that the DGX B300 is best evaluated against a concrete enterprise workload rather than purchased speculatively.

Third, return-on-investment (ROI) data from the broader Dell AI Factory ecosystem, while not DGX B300-specific, gives a useful directional benchmark for enterprises building a business case around any Blackwell Ultra-class purchase. Dell reports that as of March 2026, **"over 4,000 customers deploying the Dell AI Factory, and early adopters seeing up to 2.6x ROI within the first year"** (Source: www.dell.com). Dell also flags a broader industry shift relevant to DGX B300 buyers specifically: **"as AI code assistants and agentic workflows drastically lower the cost and time to build custom applications, CIOs are increasingly choosing to develop AI capabilities in-house, on-premises, driving the need for owned infrastructure"** (Source: www.dell.com). If that shift continues, demand pressure on DGX B300 allocation and lead times documented in this report is more likely to persist through 2026 than to ease quickly.

Frequently Asked Questions (FAQs)

What is the price of the NVIDIA DGX B300? There is no single manufacturer list price. Market data points to roughly **$300,000 to $350,000** for a base 8-GPU configuration as of Q1 2026 (Source: ifactoryapp.com), with some resellers quoting **$400,000 to $500,000** for higher-spec configurations (Source: spheron.network), and European partner quotes ranging from roughly €535,000 to over €1.4 million including VAT (Source: aiserver.eu) (Source: servermall.com). A single B300 GPU purchased outright runs about **$53,000** (Source: spheron.network).

What are the NVIDIA DGX B300 specifications? Eight Blackwell Ultra SXM GPUs, dual Intel Xeon 6776P CPUs (128 cores), 2.1 to 2.3 TB of HBM3e GPU memory depending on the NVIDIA document consulted, 144 petaFLOPS of sparse FP4 inference performance, 72 petaFLOPS of FP8 training performance, 14.4 TB/s aggregate NVLink bandwidth, and roughly 14 to 14.5 kilowatts of power draw in a 10U chassis (Source: nvidia.com) (Source: docs.nvidia.com).

When was the NVIDIA DGX B300 released? NVIDIA announced the platform at GTC on **March 18, 2025** (Source: nvidianews.nvidia.com), guided initial partner availability for the second half of 2025 (Source: nvidianews.nvidia.com), and confirmed volume shipping began in **January 2026** (Source: docs.nvidia.com).

What is the NVIDIA DGX B300 lead time? Documented lead times range from **8 to 20 weeks** depending on the source and configuration, with Spheron reporting 8 to 12 weeks and iFactoryApp reporting 12 to 20 weeks as of April 2026 (Source: spheron.network) (Source: ifactoryapp.com).

How does the DGX B300 compare to the DGX B200? DGX B300 delivers 1.5x the dense FP4 performance and 2x the attention-layer performance of DGX B200, with roughly 50% more GPU memory, at the cost of a sharp reduction in FP64 double-precision compute and a higher price (Source: nvidia.com) (Source: exxactcorp.com).

What is NVIDIA Blackwell Ultra? Blackwell Ultra is the GPU architecture underlying both the DGX B300 and the rack-scale GB300 NVL72, announced by NVIDIA as **"the next evolution of the NVIDIA Blackwell AI factory platform... paving the way for the age of AI reasoning"** (Source: nvidianews.nvidia.com).

How much GPU memory does the DGX B300 have? NVIDIA's own materials are inconsistent: 2.1 TB on the primary product page (Source: nvidia.com), 2.3 TB (8 x 288 GB HBM3e) in the technical user guide (Source: docs.nvidia.com). Independent analysis suggests roughly 2.1 TB usable after reserved overhead (Source: glennklockwood.com).

Is the NVIDIA DGX B300 available now? Yes. NVIDIA's product page states the system is **"Available Now"** as of this report's publication (Source: nvidia.com), with volume shipments confirmed since January 2026 (Source: docs.nvidia.com).

Conclusion

The honest answer to "what does an NVIDIA DGX B300 cost" is a range, not a figure: roughly $300,000 to $350,000 for a base 8-GPU system at the most commonly cited market anchor, stretching to $400,000 to $500,000 at some resellers and beyond $1 million-equivalent for fully loaded European configurations, because NVIDIA does not publish a list price and routes every sale through partners who set their own margins, support terms, and software bundles. Buyers evaluating the DGX B300 in the second half of 2026 should treat this report's Table 1 through Table 4 as a starting checklist rather than a final quote: confirm whether a given price includes NVIDIA AI Enterprise software, verify the actual GPU memory figure (2.1 TB versus 2.3 TB) against the specific configuration being quoted, and clarify directly with the OEM whether the target deployment requires direct liquid cooling, since NVIDIA's own materials and third-party integrators do not fully agree on that point.

On substance, the generational case for DGX B300 over DGX B200 is strong for inference-heavy and reasoning-model workloads, with roughly 50 percent more GPU memory, doubled attention-layer performance, and independently benchmarked MLPerf gains of 25 to 45 percent per GPU over the closest prior-generation rack-scale system. That case is weaker for workloads requiring FP64 double-precision compute, where the DGX B300 gives up the vast majority of its predecessor's throughput, and weaker still for buyers whose models comfortably fit within a smaller memory footprint, where a cheaper H200 remains the more economical choice per multiple sources' own guidance. Lead times of two to five months, elevated cloud rental rates relative to Hopper-generation hardware, and mandatory or near-mandatory liquid-cooling infrastructure at scale mean the DGX B300 is best suited to organizations with a clear, memory-bound or reasoning-model-specific inference workload and the facilities budget to match, rather than a default upgrade path for every enterprise AI buyer. With NVIDIA's Vera Rubin platform already on the roadmap for the second half of 2026, buyers should expect today's pricing and lead-time picture to keep shifting through the rest of the year.

Tags: nvidia dgx b300, dgx b300 price, blackwell ultra, dgx b200 vs b300, nvidia dgx pricing, gb300 nvl72, ai infrastructure pricing, gpu cloud pricing, dgx superpod, nvidia lead times

DISCLAIMER

This document is provided for informational purposes only. No representations or warranties are made regarding the accuracy, completeness, or reliability of its contents. Any use of this information is at your own risk. GPUSmith shall not be liable for any damages arising from the use of this document. This content may include material generated with assistance from artificial intelligence tools, which may contain errors or inaccuracies. Readers should verify critical information independently. All product names, trademarks, and registered trademarks mentioned are property of their respective owners and are used for identification purposes only. Use of these names does not imply endorsement. This document does not constitute professional or legal advice. For specific guidance related to your needs, please consult qualified professionals.