GPU SMITH
COMPUTE ENGINEERING

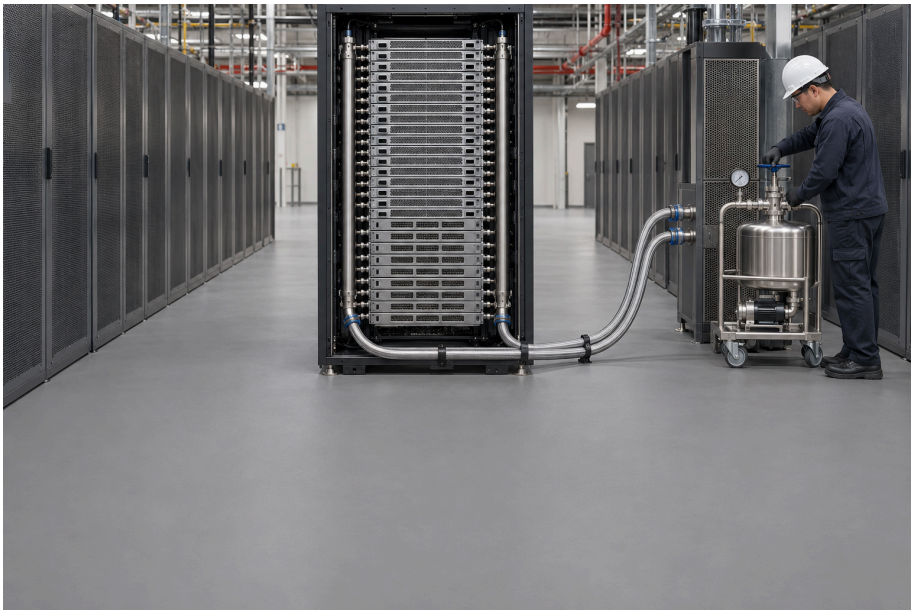
GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

GB300 NVL72 Cost per Rack: Ex-Works to Deployed (2026)

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-11-26 · gb300 nvl72 cost per rack · gb300 nvl72 price · gb300 nvl72 cost per gpu · nvidia gb300 nvl72 · gb300 nvl72 total cost of ownership · gb300 nvl72 vs gb200 nvl72 · blackwell ultra rack pricing · dgx b300 vs nvl72 · nvl72 power and cooling cost

Original article: https://gpusmith.com/articles/en/gb300-nvl72-cost-per-rack-per-gpu



In this article

1. Executive Summary
2. Introduction and Background
3. What "Rack Price" Means: Scope Before Dollars
4. Source Ledger: Who Said What, and When
5. Bill of Cost: What Each Line Covers
6. Comparative Context and Market Positioning
7. Power, Cooling and Facility Costs
8. Data Analysis and Evidence
9. Sensitivity and Staleness: Memory Prices and How to Refresh the Ledger
10. Buyer Checklist: Terms to Request in a Quote
11. Implications and Future Directions
12. Frequently Asked Questions (FAQs)
13. Conclusion

Executive Summary

NVIDIA publishes no list price for the **GB300 NVL72**, so every rack figure in circulation is an estimate with its own scope and date. The company's product page directs buyers to sales (www.nvidia.com), and independent roundups state that no list price exists for either the GB200 or GB300 rack (gpuperhour.com) (www.morphllm.com). The practical question for a business case is therefore which scope a number describes, not which number is "correct".

Dated figures cluster in three bands. Loop Capital's implied **$3.7M to $4.0M** per server dates from March 2025, before volume shipments (www.tweaktown.com). Wolfe Research's note of 2026-01-30 cited reported prices of about **$4.3M** for GB300 NVL72 against about $3M for GB200 (ca.finance.yahoo.com), and a Morgan Stanley bill-of-materials analysis relayed in May 2026 put the rack near **$4M** for hyperscale buyers (www.tomshardware.com). A single self-published Data Gravity review of anonymized small-order purchase orders (2026-08-31), whose figures it says were adjusted modestly to preserve the buyer's anonymity, reported **just under $5.0M** ex-works and about **$5.7M** deployed (www.datagravity.dev). The highest figures are a single unnamed source at $6M to $6.5M (www.tomshardware.com) and an unsourced installed-with-networking ballpark of $8M to $8.6M (www.amcompute.com).

Converted to 72 GPU slots (72 Blackwell Ultra packages, each two dies (developer.nvidia.com)), these become roughly **$60K** ($4.3M), **$69K** ($5.0M), **$79K** ($5.68M) and **$111K to $119K** ($8.0M to $8.6M) per slot, by our arithmetic. Over five years straight-line, capital alone is $1.36 to $2.73 per GPU-hour at full utilization. The same $5.68M rack costs **$2.56 to $2.88** per utilized GPU-hour at 80% to 90% utilization once colocation is added, and $3.84 at 60% (www.datagravity.dev).

Three cautions apply. First, the deployed figure excludes construction, core switching and storage (www.datagravity.dev), and no source itemizes those lines for GB300. Second, scope labels differ: one source warns that many quotes represent only the ODM (original design manufacturer) price without a proper warranty (www.tomshardware.com). Third, figures go stale quickly: NVIDIA reportedly told large customers that

server prices would rise by more than 15% in many cases for early 2027 shipments (www.tomshardware.com), and DRAM contract prices are still rising on TrendForce's forecast (www.trendforce.com). A defensible business case should carry a range, label each line "published" or "not published", and re-date every input.

Introduction and Background

The **GB300 NVL72** is NVIDIA's liquid-cooled rack-scale system that integrates 72 Blackwell Ultra GPUs and 36 Grace CPUs (www.nvidia.com) in a single NVLink domain. It is built from 18 compute trays, each with four GPUs and two CPUs (docs.nvidia.com), and 9 NVLink switch trays (docs.nvidia.com). NVIDIA's reference architecture rates the full rack at up to 142 kW (docs.nvidia.com). The target query, "gb300 nvl72 cost per rack", is a commercial-investigation search: the reader usually wants a number to place in a budget, a lender model, or a rent-versus-buy comparison against 8-GPU DGX or HGX B300 systems.

This report treats that number as a ledger problem. Prices for the rack are not catalogued anywhere; they surface in analyst notes, trade-press relays of those notes, anonymized purchase-order reviews, reseller listings and cloud-provider rate cards. Each has a different scope (ex-works, delivered, installed, cluster-level), a different date and a different level of verifiability. The report assigns each figure a source type and a staleness label, builds a bill of cost with explicit "published" and "not published" lines, and converts the results to cost per GPU slot and per GPU-hour using inputs that the reader can reproduce.

GPU Smith is an independent AI infrastructure engineering consultancy. It is not a chip manufacturer, cloud provider or reseller, and it does not sell the systems discussed here. The report is reference analysis only. It does not rely on any client engagement, and it does not claim delivery or benchmark results. All factual citations point to third-party sources (the few links to GPU Smith articles are navigational and are not evidence for any figure), and readers who plan a purchase should obtain written quotes and have their own finance and facilities engineers review the assumptions. All figures are as of 26 November 2026 unless a different date is shown.

What "Rack Price" Means: Scope Before Dollars

A single phrase, "GB300 NVL72 price", hides at least five different purchases. Separating them is the most valuable step in any business case.

- **Ex-works (EXW) rack:** the rack leaves the factory, and everything from the loading dock onward is the buyer's problem. Data Gravity uses exactly this framing: "a purchased rack is EXW" (www.datagravity.dev).
- **Rack with in-rack cooling distribution:** a coolant distribution unit (CDU) can sit inside the rack, in a row, or be absent from the quote. Cisco's catalogue for the Supermicro rack shows two part numbers differing only in cooling scope, one without any CDU (www.cisco.com). Data Gravity's purchase-order sample likewise included configurations with and without an in-rack CDU (www.datagravity.dev).
- **Delivered and installed rack:** adds freight, placement, coolant fill, power and cooling hookup and commissioning. Lenovo notes that its rack ships without water, so the customer must fill it (lenovopress.lenovo.com).

- **Cluster-level cost:** adds scale-out networking, storage and management. NVIDIA's DGX SuperPOD reference architecture defines a scalable unit as 8 DGX GB300 systems (docs.nvidia.com), with Quantum-X800 InfiniBand compute fabric and storage from certified partners, so a cluster is much more than a stack of racks.
- **Facility-inclusive cost:** adds the building, power and heat rejection, which no rack quote includes.

What is in the rack

Rack contents differ little across vendors but are not identical. Lenovo lists 18 GPU compute trays, 9 switch trays and a choice of 6 or 8 power shelves (lenovopress.lenovo.com). Cisco's datasheet for the Supermicro rack lists 8 power shelves (www.cisco.com), and NVIDIA's reference architecture specifies 8 shelves of 33 kW (docs.nvidia.com). A quote that specifies 6 shelves is therefore a different configuration, with a different power ceiling, from one that specifies 8.

What "72 GPUs" means

The figure of 72 counts packages, not dies. Each Blackwell Ultra GPU is composed of two reticle-sized dies joined by NVIDIA's High-Bandwidth Interface (developer.nvidia.com), and the pair functions as a single CUDA-programmed accelerator (developer.nvidia.com). Each package carries 288 GB of HBM3e (developer.nvidia.com). Dividing a rack price by 72 therefore gives cost per GPU package, which matches how 8-GPU DGX B300 systems count GPUs (www.nvidia.com). Dividing by 144 would count dies, which no vendor price sheet does.

Warranty and service terms are part of scope

Warranty terms change what a price buys. Lenovo's rack carries a 3-year customer-replaceable-unit and onsite warranty with next-business-day 9x5 coverage (lenovopress.lenovo.com), with support entitled at rack level by rack serial number (lenovopress.lenovo.com). HPE's page states three years of hardware diagnostic support and repair, but only 90 days of software and initial setup support (buy.hpe.com). Cisco's offer of the Supermicro platform follows standard Supermicro warranty terms (www.cisco.com). A reseller listing for the rack in India states "made to order" and "price on application", configured and quoted per deployment (rdp.in). These pages show that a public number, if it exists, is indicative at best.

Source Ledger: Who Said What, and When

Table 1 below lists every dated GB300 NVL72 price figure found in this research, with its scope, originator, type and a reliability label. "Secondhand" means the page the figure was read on is not the originator. Ledger rows are ordered by the figure's own date, not by price.

Figure (GB300 NVL72 unless noted)	Scope as stated	Source and date	Type	Reliability and staleness label
$3.7M to $4.0M per server	Implied per-server price in an order estimate	Loop Capital via TweakTown, 2025-03-26 (www.tweaktown.com); repeated as "rumored" by a researcher's notes dated 2026-07-01 (glennklockwood.com)	Analyst note, second hand	Predates volume shipments; repeated, not re-measured. **Stale**
About $4.3M (GB200 about $3M)	"Reported prices", scope not stated	Wolfe Research via Investing.com, 2026-01-30 (ca.finance.yahoo.com)	Analyst, second hand	Wolfe adds that NVIDIA's boards are about 75% of final rack price (ca.finance.yahoo.com). **Aging (10 months)**
About $4M per rack	Hyperscale buyer, bill-of-materials basis	Morgan Stanley Research via Tom's Hardware, 2026-05-21 (www.tomshardware.com)	Analyst, second hand	Original note not reached. **Aging (6 months)**
$6M to $6.5M	Inference-optimized system; one unnamed source	Tom's Hardware, about March 2026 (www.tomshardware.com)	Media	Same article warns many quotes are ODM-only without proper warranty (www.tomshardware.com). **Outlier, scope unknown**
Just under $5.0M ex-works, about $32K more with 250 kW CDU	Small orders, without in-rack CDU	Data Gravity purchase-order review, 2026-08-31 (www.datagravity.dev)	PO review, anonymized	Figures adjusted to protect buyer anonymity; cannot be independently checked (www.datagravity.dev). **Most recent, unverifiable**
About $5.68M deployed	Ex-works plus five line items; excludes construction, core switching and storage	Data Gravity author calculation, 2026-08-31 (www.datagravity.dev)	Author calculation	Line items are shown as an image, not text. **Derived**
$8M to $8.6M	"Includes the NVLink switch trays, networking and installation"	American Compute, October 2026 (www.amcompute.com)	Reseller or appraiser ballpark	No originator named. **Treat as scope-expanded outlier**
GB200 NVL72: $3.1M rack-only, $3.9M all-in	Hyperscaler; all-in adds networking, storage, other	SemiAnalysis, 2025-08-20 (newsletter.semianalysis.com)	Research firm, originator	Different product, but the only published rack-only to all-in delta. **Aging**

The ledger shows that sources agree less than their headline ranges suggest. Spheron's roundup, for instance, gives GB300 as anywhere from $3.7M to $6.5M (www.spheron.network), a spread of 1.76 times (our arithmetic: 6.5 divided by 3.7). That spread is mostly a scope and date effect, not a price dispersion for one identical purchase. Few rows carry a precise scope label for the GB300 NVL72. The clearest are the purchase-order review ("without in-rack CDU") and the American Compute ballpark ("includes networking and installation"), and neither can be independently verified. Other rows state only a basis, or say that scope is not stated.

Reading the ledger: four rules

- **Prefer the originator:** the Wolfe and Morgan Stanley figures were reached only through relays, and the Loop Capital figure through TweakTown. Where the relay is the only access, the article labels it secondhand.
- **Treat repeats as one datapoint:** the $3.7M to $4M figure appears in at least three places, but all trace to the same March 2025 estimate (www.spheron.network).
- **Distinguish transacted from estimated:** Data Gravity describes its numbers as transacted prices for small orders and says hyperscale buyers will do better (www.datagravity.dev). An analyst bill-of-materials figure is a different kind of number.
- **Date-stamp every row:** a figure from early 2026 predates the memory price movements discussed later in this report.

A preliminary note on the $6M to $6.5M figure: Data Gravity treats it as a media estimate and argues that purchase orders sit lower (www.datagravity.dev). The two cannot both describe the same purchase scope, and neither can be verified by a reader.

Bill of Cost: What Each Line Covers

Table 2 below is a bill-of-cost template. A figure appears only where a dated source exists. Everything else is marked "not published", and the reader should obtain a written quote for it.

Cost line	Status	What the sources say	Date and type
Rack, ex-works	Published (range)	$3.7M to $6.5M across analysts and media; just under $5.0M in PO review (www.datagravity.dev)	2025-2026; mixed
In-rack CDU	Partly published	About $32K more for a 250 kW CDU (www.datagravity.dev); Cisco sells the rack with or without a CDU (www.cisco.com)	PO review, 2026-08
In-rack liquid-cooling content	Published (not additive)	$49,860 per rack, cold plates, manifolds and related parts inside the BOM (www.tomshardware.com)	Morgan Stanley via media, 2025-11
Facility CDU, heat rejection	Not published	Cisco lists CDU options up to 250 kW in-rack, 1.8 MW in-row, and a 200 kW sidecar (www.cisco.com), with no prices; in-row units and sidecars are quoted separately	Vendor datasheet
Scale-out networking (NICs, switches, optics)	Share only	Networking is 15% of total cluster cost for a 3-layer InfiniBand GB300 NVL72 cluster (newsletter.semianalysis.com); GB200 all-in minus rack-only is about $0.8M (newsletter.semianalysis.com)	SemiAnalysis, 2025-2026
Storage	Not published	NVIDIA gives throughput guidelines per scalable unit, not prices; storage comes from certified partners (docs.nvidia.com)	Vendor reference
Install and commissioning	Total only	About $0.68M between ex-works and deployed (our arithmetic: $5.68M less just under $5.0M) (www.datagravity.dev)	Author calculation, 2026-08

Cost line	Status	What the sources say	Date and type
Spares	Ratio only	Lenovo recommends at least 2% spare-to-production ratio for compute and switch trays (lenovopress.lenovo.com)	Vendor guide, 2026-08
Warranty and support	Terms published, price not	Three years of repair and onsite support; 90 days of setup support (buy.hpe.com)	Vendor page
Power and colocation	Published inputs	See the power section below	CBRE, EIA, 2025-2026
Facility construction	Published per MW	$14 to $16 per watt for advanced US AI facilities (reports.turnerandtownsend.com)	Turner and Townsend, 2026-27

Table 2 makes the central point of this report concrete: the largest published number for any single non-rack line is the aggregate installation gap of roughly $0.7M, and its five components are not itemized in text. A buyer who needs cost per GPU at the cluster level has published evidence for the rack and for the facility, but only shares and ratios for the network, and nothing for storage.

Rack content that is easy to mis-scope

Data Gravity reports compute trays at about 87% of rack value, and NVIDIA content including the NVLink fabric above 92% (www.datagravity.dev). If correct, only a single-digit share of the rack price is power shelves, busbar, cold plates, steel, cabling and integration. This matters because it limits how much an integrator can discount: most of the price sits in components that NVIDIA supplies. Wolfe's supply-chain checks put the boards NVIDIA sells at about 75% of the final rack price (ca.finance.yahoo.com). The two shares differ in definition (board versus compute tray plus fabric), so they are not directly comparable.

Scale-out networking is the line most often missing from rack quotes. Each GPU has an 800 Gb/s ConnectX-8 connection (www.nvidia.com), and SemiAnalysis noted that an NVL72 rack retains one scale-out optical port per GPU, 72 per rack (newsletter.semianalysis.com). Within a three-layer network, optical transceivers account for 60% of networking cost when NVIDIA transceivers are used (newsletter.semianalysis.com). Applying the 15% cluster share to a hypothetical $5.0M rack does not yield a price, because the share is of total cluster cost, not of rack cost. No unit prices for these parts were found in a primary source.

Comparative Context and Market Positioning

Against GB200 NVL72

The GB200 NVL72 comparison shows what scope can do. SemiAnalysis put the GB200 rack at $3.1M alone and about $3.9M all-in (newsletter.semianalysis.com), a 26% uplift (our arithmetic: 3.9 divided by 3.1). Wolfe's reported $3M to $4.3M implies GB300 about 43% above GB200 (4.3 divided by 3.0) (ca.finance.yahoo.com). Tom's Hardware's unnamed source implies about 1.8 to 2.3 times (our arithmetic: 6.0 divided by 3.4 to 6.5 divided by 2.8), from $2.8M to $3.4M for GB200 against $6M to $6.5M for GB300 (www.tomshardware.com). Data Gravity argues the widely quoted "$3 million rack" described GB200 and answers the wrong question

(www.datagravity.dev). NVIDIA states the GB300 NVL72 delivers 1.5 times the dense FP4 Tensor Core FLOPS of Blackwell GPUs (www.nvidia.com), so a higher price does not by itself imply a worse price per unit of compute. Morgan Stanley's in-rack cooling content rises 20% from GB200 to GB300 (www.datagravity.dev), but cooling content is only about 1% of a $5.0M rack (our arithmetic: $49,860 divided by $5.0M).

Against 8-GPU DGX and HGX B300 systems

Eight-GPU B300 systems are priced far more transparently, mainly through reseller and OEM configurators. NVIDIA's DGX B300 uses 8 Blackwell Ultra SXM GPUs (www.nvidia.com) and draws about 14 kW (www.nvidia.com). Nine such systems hold 72 GPUs and draw about 126 kW (our arithmetic), close to the 132 kW to 140 kW that HPE gives for one rack (buy.hpe.com). Published system prices for 8-GPU B300 servers, none of them NVIDIA list prices (aircloud.dev), include:

- **$485,918 starting price** for an HGX B300 server at one reseller, with eight 800 Gbps InfiniBand links (petronellatech.com).
- **$515,808.58 as configured** for a Colfax 8U server with an 8x B300 baseboard (colfax-intl.com).
- **$500,000 to $550,000** as a working range in a 2026 cluster-cost guide, updated 2025-11-19 (aircloud.dev).
- **$550,000 to $650,000** in a reseller survey cited by an aggregator in November 2026 (www.morphllm.com).
- **$785,606.50** as Lenovo's usual customer sale price for an 8x B300 ThinkSystem configuration as of 15 June 2026 (lenovopress.lenovo.com).

Dividing by 8 gives $60.7K to $98.2K per GPU (our arithmetic). Nine systems at $500,000 to $550,000 cost $4.5M to $4.95M; at Lenovo's configuration, $7.07M. These are not like-for-like purchases. The 8-GPU systems have no 72-GPU NVLink domain, no included liquid cooling in the quoted configurations and no scale-out switches, while the rack includes the NVLink switch trays. The per-GPU comparison is therefore directional. It does indicate that the per-slot premium for the rack is not obviously large, and that configuration choices (CPU, memory, storage, support term) move 8-GPU prices by tens of percent.

Against rented capacity

Rented GB300 pricing is sparse. CoreWeave lists GB300 NVL72 as contact sales (www.morphllm.com). Together AI lists $15.79 per GPU-hour on demand (www.together.ai). Oracle lists the GB300 NVL72 instance at $18.00 per GPU-hour (www.oracle.com). AWS lists B300 capacity blocks at $16.146 per accelerator (aws.amazon.com), with the next price update scheduled for January 2027 (aws.amazon.com). NVIDIA's chief financial officer said on the 17 November 2026 call that B300 rental rates had risen more than 60% since introduction (www.nasdaq.com). Rental rate cards carry no date stamp of their own, and the AWS page confirms they change. They are valid comparators only for the date on which they were read.

Power, Cooling and Facility Costs

Rack power is the starting input. NVIDIA's reference architecture cites up to 142 kW (docs.nvidia.com), Vertiv's reference design supports up to 142 kW (www.vertiv.com), and Schneider Electric's design caps at 142 kW per rack (www.se.com). HPE gives 132 kW to 140 kW (buy.hpe.com), and Lenovo rates a 135 kW TDP and up to

155 kW peak (lenovopress.lenovo.com). Vertiv reports a field deployment with each rack drawing about 140 kW at full load (www.vertiv.com). This report uses 140 kW as a planning figure.

Electricity

At 140 kW running continuously, a rack draws 1,226,400 kWh a year (140 times 8,760). The U.S. Energy Information Administration (EIA) reported September 2026 average prices of 9.52 cents per kWh for industry (www.eia.gov) and 14.27 cents for commercial customers (www.eia.gov). Texas industrial customers paid 6.99 cents (www.eia.gov). At those rates, IT load alone costs about $116,800, $175,000 and $85,700 a year respectively (our arithmetic). Uptime Institute's 2025 survey of data center operators reports a weighted average power usage effectiveness of 1.54 (datacenter.uptimeinstitute.com); applying that ratio raises the three figures to about $179,800, $269,500 and $132,000. Liquid-cooled facilities typically do better than the survey average, so these are upper-range planning values.

Colocation

Renting space and power instead of building is the usual alternative. CBRE reported Northern Virginia asking rates of $160 to $185 per kW for requirements of 10 MW and above (www.cbre.com), per month according to the trade-press summary of the report. CBRE's H2 2025 report put 250 to 500 kW requirements at a record $196.25 per kW per month (www.cbre.co.uk), and its midyear 2026 outlook expects them above $215 (www.cbre.co.uk). Primary-market vacancy fell to a record low of 1.4% in H1 2026 (www.cbre.com). At 140 kW and $160 to $185, that is $22,400 to $25,900 a month, or $268,800 to $310,800 a year (our arithmetic). Data Gravity's own colocation input of about $318,000 a year (www.datagravity.dev) is about 2% above the top of this range (our arithmetic: 318,000 divided by 310,800).

Building a facility

If the buyer must build rather than lease, the per-rack facility cost approaches the rack's own price. Cushman and Wakefield's 2026 guide puts all-in greenfield cost in the United States and Canada at $17.6M per MW, excluding chips (finviz.com). Turner and Townsend puts advanced US AI facilities at $14 to $16 per watt (reports.turnerandtownsend.com). JLL notes that tenant technology fit-out for AI can cost as much as $25M per MW (www.jll.com). At 0.14 MW per rack, those construction figures equal $1.96M to $2.46M per rack (our arithmetic), or roughly 40% to 50% of a $5.0M rack. Data Gravity excludes construction from its deployed figure (www.datagravity.dev), so any buyer without existing liquid-cooled capacity must add this line.

Cooling hardware

Cooling decisions affect scope. Cisco's datasheet lists an in-rack CDU up to 250 kW, an in-row CDU up to 1.8 MW and a liquid-to-air sidecar up to 200 kW (www.cisco.com). HPE offers both in-row and a newly validated in-rack CDU (buy.hpe.com). Lenovo says the coolant connection can go to an in-rack or in-row CDU (lenovopress.lenovo.com). Vertiv reported CDUs delivering 121 kW of heat rejection for a GB300 NVL72 in a customer deployment (www.vertiv.com). No source gave a facility-side CDU or chiller price, so that line stays "not published" in Table 2.

Data Analysis and Evidence

This section converts the ledger to per-slot and per-GPU-hour values. *Table 3* below uses one rack of 72 GPU slots. Capital per year is the price divided by five (five-year straight line), and GPU-hours per year are 72 times 8,760, or 630,720. The per-GPU-hour column is capital only, at 100% utilization, and is therefore a floor. Every figure in the table is arithmetic on the displayed inputs.

Scope scenario	Rack price	Per GPU slot (price / 72)	Capital per year (5-year)	Capital per GPU-hour at 100%
Loop Capital low (www.tweaktown.com)	$3.7M	$51.4K	$0.74M	$1.17
Morgan Stanley hyperscale (www.tomshardware.com)	$4.0M	$55.6K	$0.80M	$1.27
Wolfe reported (ca.finance.yahoo.com)	$4.3M	$59.7K	$0.86M	$1.36
Data Gravity ex-works (www.datagravity.dev)	$5.0M	$69.4K	$1.00M	$1.59
Data Gravity deployed (www.datagravity.dev)	$5.68M	$78.9K	$1.136M	$1.80
Single-source media (www.tomshardware.com)	$6.0M to $6.5M	$83.3K to $90.3K	$1.20M to $1.30M	$1.90 to $2.06
Installed with networking (www.amcompute.com)	$8.0M to $8.6M	$111.1K to $119.4K	$1.60M to $1.72M	$2.54 to $2.73

The table reproduces the three conversions given in the Executive Summary and extends them. The same physical rack becomes about $60K, $69K or $79K per GPU slot depending on whether the reader starts from Wolfe, from ex-works purchase orders or from the deployed calculation. The $8M to $8.6M ballpark yields $111K to $119K per slot, which matches the $111,000 to $119,000 per GPU figure on the same page (www.amcompute.com). The spread from lowest to highest row is 2.3 times (8.6 divided by 3.7). A finance model that picks one row without stating its scope is in effect picking one of the five purchases described earlier.

Operating-cost overlay

Adding Data Gravity's colocation input of about $318,000 a year (www.datagravity.dev) to straight-line capital for the $5.68M deployed case gives $1,454,000 a year, or $2.31 per GPU-hour at 100% utilization (our arithmetic). At 90% and 80% utilization that becomes $2.56 and $2.88, and Data Gravity reports $3.84 at 60% (www.datagravity.dev). Data Gravity puts five years of electricity at about $1.1M, under a fifth of deployed capital, at an assumed $0.15 per kWh (www.datagravity.dev). At the EIA commercial rate with the 1.54 PUE applied above, five years of electricity is about $1.35M, roughly 24% of the $5.68M deployed cost (our arithmetic), so the 'under a fifth' result does not hold at that rate. For a buyer with a low-cost power contract, the rack price and utilization therefore dominate the result; electricity is secondary.

Useful-life assumption

The five-year life is an assumption, not a measurement. CoreWeave's 10-K states that, effective 1 January 2023, it moved its computing equipment used in data centers from five to six years of useful life, so this is a 2023 accounting estimate and not new evidence about Blackwell-era hardware (s205.q4cdn.com), and AWS's chief executive reportedly stated a 5 to 6 year life for chips, servers and networking gear (fortune.com). A six-year life lowers the capital column by one sixth ($1.36 becomes $1.13 per GPU-hour at $4.3M). Rental rates have been rising for B300 (www.nasdaq.com), which supports longer lives in the current market, but the report offers no forecast.

Rent versus own, as arithmetic

Full-load all-in cost for the deployed case is $2.31 per GPU-hour. Against Together AI's $15.79 (www.together.ai), that cost equals about 15% of the rental rate (2.31 divided by 15.79), so a buyer breaks even on these inputs at about 15% utilization. The comparison omits the cost of capital, cluster networking, storage, staff and the facility build, all of which would raise the owned figure, and rental rates are not contracted prices. The result shows direction, not a recommendation.

Sensitivity and Staleness: Memory Prices and How to Refresh the Ledger

Memory is the main driver of price movement in this period. A Morgan Stanley analysis relayed in May 2026 put memory content in a GB300 rack at $373,939 (wccftech.com), about 9% of the roughly $4M rack (our arithmetic), against about $2M for a Vera Rubin rack, up 435% (www.tomshardware.com). For GB300 the exposure is smaller but real. Samsung and SK hynix reportedly raised 2026 HBM3E supply prices by close to 20% before the year began (www.tomshardware.com). TrendForce projects blended HBM average selling price up 121% year over year in 2027 (www.trendforce.com), and on 25 November 2026 expected quarterly contract price increases to narrow progressively (www.trendforce.com). SK hynix has said its DRAM and HBM volumes through 2029 are effectively sold out (koreajoongangdaily.com).

NVIDIA is also signalling higher pricing. Tom's Hardware, citing Bloomberg, reported in August 2026 that NVIDIA told large customers that prices of servers with its AI chips would rise by more than 15% in many cases (www.tomshardware.com). NVIDIA's own results show 74.8% GAAP gross margin for the quarter ended October 2026 (investor.nvidia.com) and guidance of 71.0% to 72.0% for the next quarter (investor.nvidia.com), with purchase commitments of $390 billion (www.nasdaq.com). These are corporate figures, not rack prices, but they are the best primary indicators of cost-side pressure.

If the 15% increase applied to the $5.0M ex-works figure, the rack would cost $5.75M and $79.9K per slot (our arithmetic, illustrative; the report does not claim it applies to GB300 specifically). TrendForce expects Vera Rubin ASP to be roughly double that of GB300 (www.trendforce.com), so some buyers may find GB300 pricing relatively stable even as the market moves.

A refresh protocol

- **Re-date every row quarterly:** replace a figure only when a newer source with the same scope appears; do not average across dates.
- **Track three triggers:** memory contract prices, NVIDIA's gross margin guidance and rental-rate changes.
- **Never mix scopes:** keep the ex-works, deployed and cluster-level columns separate.
- **Mark unverifiable inputs:** flag the purchase-order review and any unnamed-source figure as unverifiable in the model.
- **Request written quotes:** a business case with only public estimates is a screening tool, not a procurement basis.

Buyer Checklist: Terms to Request in a Quote

The checklist below converts the scope findings into questions for the vendor.

- **Delivery term:** ex-works, delivered duty paid or installed; who owns freight, insurance and placement.
- **Rack contents:** number of power shelves (6 or 8) (lenovopress.lenovo.com), rack type, busbar, management switches, cables and manifolds.
- **CDU:** none, in-rack (capacity and model) or in-row; whether the quote includes the CDU or lists it separately (www.cisco.com).
- **Coolant:** who supplies the fluid and fill service, since racks may ship dry (lenovopress.lenovo.com).
- **Integration level:** rack-level or cluster-level validation; Supermicro describes L11 and L12 testing before shipment (www.supermicro.com), and Dell described its first rack as fully assembled and tested (www.dell.com).
- **Warranty:** term (three years is common), response time and whether support is entitled by rack serial number (lenovopress.lenovo.com).
- **Setup support:** duration and scope; HPE's page lists 90 days (buy.hpe.com).
- **Spares:** whether the quote includes the 2% spare ratio Lenovo recommends (lenovopress.lenovo.com), and who stores it.
- **Networking:** NICs, switches, optics and cables for the scale-out fabric, and the number of layers assumed.
- **Storage:** vendor, throughput target and whether it is in scope.
- **Software and support entitlements:** NVIDIA software, management tools and liquid-cooling maintenance, where offered.
- **Quote validity:** the number of days the price holds, given the price movements above.
- **Price basis:** whether the quote is a list price, a negotiated price or an indicative price (buy.hpe.com).
- **Lead time and payment terms:** delivery date, milestone payments and cancellation terms.
- **Tax and duty:** whether taxes, import duties or export-control compliance are the buyer's responsibility.

Implications and Future Directions

For enterprise buyers, the main implication is that the scope label matters more than the dollar value. A buyer comparing a rack quote with an 8-GPU system quote should convert both to cost per GPU package, and align them on cooling, networking, warranty and facility. For lenders, the most useful inputs are the dated, scope-labelled rows in Table 1 and the explicit gaps in Table 2, because unlabelled numbers create valuation risk. For operators, the rack price is only the first of several lines, and construction or colocation can match it in a five-year view (www.datagravity.dev).

Rack-scale systems are shifting how vendors sell. NVIDIA's DGX SuperPOD reference architecture sizes a scalable unit of eight racks at 1.2 MW TDP (docs.nvidia.com). The Enterprise reference architecture uses a single rack as the scalable unit (docs.nvidia.com). Both indicate that rack price will increasingly be one line in a larger, negotiated statement of work.

Three developments could reduce the value of this report's figures within months: a change in NVIDIA's pricing policy, further memory price movement, and a shift in availability as Vera Rubin ships. TrendForce expects Vera Rubin system prices to be roughly double those of GB300 (www.trendforce.com), and Wolfe's note put Rubin racks at $5M to $6M (ca.finance.yahoo.com). If that ladder holds, GB300 prices could soften in relative terms but not necessarily in absolute terms. As GB300 pricing converges on transacted values, the value of anonymized purchase-order evidence should fall and the value of verifiable quotes should rise.

GPU Smith's own perspective, as an independent adviser, is that no public figure should be used as a quote substitute. The ledger approach reduces the risk of anchoring on one number, and it makes assumptions visible to finance reviewers. Readers should have their own electrical, mechanical and financial specialists review any plan that depends on these inputs.

Frequently Asked Questions (FAQs)

How much does a GB300 NVL72 rack cost?

There is no list price (www.nvidia.com). Dated estimates run from about $3.7M to $6.5M, with a purchase-order review at just under $5.0M ex-works and about $5.68M deployed (www.datagravity.dev). An installed-with-networking ballpark reaches $8M to $8.6M (www.amcompute.com).

What is the GB300 NVL72 price per GPU?

Dividing by 72 GPU packages gives about $60K at $4.3M, $69K at $5.0M and $79K at $5.68M (our arithmetic). The $8M to $8.6M ballpark equals $111K to $119K per GPU (www.amcompute.com).

What does a GB300 NVL72 deployment cost beyond the rack?

Data Gravity's calculation adds about $0.68M for five line items but excludes construction, core switching and storage (www.datagravity.dev). Facility build costs are $14 to $16 per watt for advanced US AI sites (reports.turnerandtownsend.com).

What is the total cost of ownership?

On Data Gravity's inputs, the $5.68M deployed rack with colocation costs $2.56 to $2.88 per utilized GPU-hour at 80% to 90% utilization (www.datagravity.dev). Table 3 shows the capital-only floor.

How does GB300 NVL72 compare with GB200 NVL72?

Analyst figures imply about 43% more for GB300 (Wolfe) (ca.finance.yahoo.com), against a 26% rack-only to all-in uplift for GB200 (newsletter.semianalysis.com).

What are the power and cooling requirements?

Rack power is up to 142 kW (docs.nvidia.com). At EIA's September 2026 industrial rate, IT load costs about $116,800 a year (www.eia.gov). Colocation of 140 kW at $160 to $185 per kW per month costs $268,800 to $310,800 a year (www.cbre.com).

How does it compare with a DGX B300 purchase?

Nine 8-GPU systems match the 72-GPU count. At $500,000 to $550,000 each (aircloud.dev) they cost $4.5M to $4.95M before networking and facilities.

Conclusion

No single rack price exists for the GB300 NVL72, and the figures in circulation measure different things. The defensible approach is to carry three columns: ex-works rack, installed rack and cluster-level cost per GPU, each with its own source, date and reliability label. On current public evidence the ex-works range is roughly $4M to $5M, the deployed estimate is about $5.7M under one author's calculation (the ex-works and deployed figures both trace to the same anonymized, unverifiable purchase-order review), and cluster and facility lines remain unpublished. Per GPU slot, that spans about $60K to $79K for the rack alone. These are screening values. Because memory prices, NVIDIA's pricing and rental rates are all moving, a purchasing decision needs written quotes with stated delivery terms, cooling scope, warranty and validity period. Re-date the ledger every quarter, keep unverifiable figures flagged, and treat any number older than six months as stale.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.nvidia.com/en-us/data-center/gb300-nvl72
2. https://gpuperhour.com/blog/gb200-gb300-nvl72-explained
3. https://www.morphllm.com/nvidia-gb300
4. https://www.tweaktown.com/news/104223/apple-becomes-nvidia-customer-orders-estimated-1-billion-of-new-gb300-nvl72-ai-servers/index.html
5. https://ca.finance.yahoo.com/news/wolfe-lifts-nvidia-target-25-145211458.html
6. https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-memory-costs-soar-485-percent-latest-ai-systems-now-cost-usd7-8-million-to-build-memory-now-comprises-25-percent-of-the-total-cost-rubin-gpus-a-mere-usd50-000-apiece
7. https://www.datagravity.dev/p/how-much-does-an-nvidia-nvl72-cost

8. https://www.tomshardware.com/tech-industry/artificial-intelligence/price-of-nvidias-vera-rubin-nvl72-racks-skyrockets-to-as-much-as-usd8-8-million-apiece-but-server-makers-margins-will-be-tight-nvidia-is-moving-closer-to-shipping-entire-full-scale-systems
9. https://www.amcompute.com/compare/gb300-vs-b300
10. https://developer.nvidia.com/blog/inside-nvidia-blackwell-ultra-the-chip-powering-the-ai-factory-era
11. https://www.tomshardware.com/pc-components/dram/nvidia-reportedly-warns-biggest-customers-of-15-percent-price-hikes-on-ai-servers
12. https://www.trendforce.com/presscenter/news/20261125-13321.html
13. https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/components.html
14. https://www.cisco.com/site/us/en/solutions/global-partners/supermicro/nvidia-gb300-nvl72-ds.html
15. https://lenovopress.lenovo.com/lp2357-lenovo-nvidia-gb300-nvl72-rack-scale-ai
16. https://docs.nvidia.com/pdf/dgx-spod-gb300-ra.pdf
17. https://www.nvidia.com/en-us/data-center/dgx-b300
18. https://buy.hpe.com/be/en/compute/rack-scale-system/nvidia-nvl-system/p/1014890105
19. https://rdp.in/gpu-mart?p=1651
20. https://glennklockwood.com/garden/processors/B300
21. https://newsletter.semianalysis.com/p/h100-vs-gb200-nvl72-training-benchmarks
22. https://www.spheron.network/blog/gb300-nvl72-vs-gb200-nvl72-pricing-availability-2026
23. https://www.tomshardware.com/pc-components/cooling/cooling-system-for-a-single-nvidia-blackwell-ultra-nvl72-rack-costs-a-staggering-usd50-000-set-to-increase-to-usd56-000-with-next-generation-nvl144-racks
24. https://newsletter.semianalysis.com/i/178153689/when-will-cpo-be-ready-for-primetime
25. https://reports.turnerandtownsend.com/data-centre-construction-cost-index-2026-us/ai-data-centers
26. https://newsletter.semianalysis.com/p/nvidias-optical-boogeyman-nvl72-infiniband
27. https://aircloud.dev/fr/blog/hgx-b300-cluster-cost
28. https://petronellatech.com/hardware/nvidia-hgx
29. https://colfax-intl.com/Servers/CX8860s-X6B3
30. https://lenovopress.lenovo.com/lp2368-on-premise-vs-cloud-generative-ai-total-cost-of-ownership
31. https://www.together.ai/pricing
32. https://www.oracle.com/cloud/compute/pricing
33. https://aws.amazon.com/ec2/capacityblocks/pricing
34. https://www.nasdaq.com/articles/nvidia-nvda-q3-2027-earnings-call-transcript
35. https://www.vertiv.com/en-us/about/news-and-insights/corporate-news/vertiv-develops-energy-efficient-cooling-and-power-reference-architecture-for-the-nvidia-gb300-nvl72
36. https://www.se.com/ww/en/about-us/newsroom/news/press-releases/schneider-electric-announces-new-reference-designs-featuring-integrated-power-management-and-liquid-cooling-controls-supporting-nvidia-mission-control-and-nvidia-gb300-nvl72-68ca8fdd5e6c6f9d3a096133
37. https://www.vertiv.com/en-us/about/news-and-events/articles/blog-posts/how-vertiv-and-coreweave-are-orchestrating-ai-innovation
38. https://www.eia.gov/electricity/monthly/update/end-use.php

39. https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a
40. https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2025.Annual.Survey.Report.pdf?version=0
41. https://www.cbre.com/insights/books/north-america-data-center-trends-h1-2026/northern-virginia-data-center-market
42. https://www.cbre.co.uk/insights/books/north-america-data-center-trends-h2-2025
43. https://www.cbre.co.uk/insights/books/us-real-estate-market-outlook-midyear-review-2026/data-centers
44. https://www.cbre.com/insights/books/north-america-data-center-trends-h1-2026
45. https://finviz.com/news/388803/cushman-wakefield-releases-2026-data-center-development-cost-guide-citing-21-rise-in-per-mw-construction-costs
46. https://www.jll.com/en-us/insights/market-outlook/data-center-outlook
47. https://s205.q4cdn.com/133937190/files/doc_financials/2025/q4/CoreWeave-Inc-FY25-10-K-7.pdf
48. https://fortune.com/2026/04/15/data-centers-hyperscalers-spending-billions-on-hardware-thats-worthless-in-3-years
49. https://wccftech.com/nvidia-vera-rubin-rack-hit-with-memory-price-surge-pushing-hbm4-lpddr5x-bill-to-2m-of-7-8m-total
50. https://www.trendforce.com/presscenter/news/20260929-13255.html
51. https://koreajoongangdaily.com/business/sk-hynix-says-dram-hbm-booked-through-2029-as-profit-soars-to-62b/12942008
52. https://investor.nvidia.com/news/press-release-details/2026/NVIDIA-Announces-Financial-Results-for-Third-Quarter-Fiscal-2027
53. https://www.trendforce.com/presscenter/news/20260828-13204.html
54. https://www.supermicro.com/manuals/brochure/Brochure-DCBBS.pdf
55. https://www.dell.com/en-us/blog/dell-delivers-market-s-first-nvidia-gb300-nvl72-to-coreweave
56. https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/network-logical-architecture.html

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.