DGX vs HGX vs NVL72 vs MGX: NVIDIA Platform Comparison 2026

Published July 29, 2026 37 min read

DGX vs HGX vs NVL72 vs MGX: NVIDIA Platform Comparison 2026

Executive Summary

NVIDIA sells the same generation of graphics processing units (GPUs) through four distinct system layers, and confusing them is easy because NVIDIA reuses chip-generation names like "B200" across products that are not interchangeable (Source: modal.com). **NVIDIA HGX** is a licensable 8-GPU baseboard, built around NVLink and NVSwitch, that original equipment manufacturers (OEMs) such as Dell, HPE, Supermicro, and Lenovo integrate into their own branded servers (Source: amcompute.com). **NVIDIA DGX** is NVIDIA's own sealed, turnkey version of essentially the same hardware, using the identical HGX baseboard but with a fixed CPU, memory, and support stack chosen by NVIDIA rather than an OEM (Source: amcompute.com). **NVIDIA GB200 NVL72** (and its successor **GB300 NVL72**) is a rack-scale system that abandons the 8-GPU server boundary entirely, wiring 72 GPUs and 36 Grace CPUs into a single NVLink domain that delivers 130 terabytes per second (TB/s) of GPU-to-GPU bandwidth (Source: nvidia.com). **NVIDIA MGX** is not a GPU-carrying product at all but an open, modular rack and tray specification, unveiled in May 2023, that lets OEM and ODM partners assemble more than 100 different server configurations from a common blueprint (Source: nvidianews.nvidia.com). As of July 2026, MGX has become the physical chassis standard underneath the newest rack-scale systems, and NVIDIA's newest DGX B300 system is "deployable in NVIDIA MGX racks for the first time," formally bridging what once looked like four separate product lines (Source: nvidia.com).

Price and power draw scale together with performance. An 8-GPU HGX B200 server from a systems integrator is listed from about $483,000 to $792,000 depending on cooling and configuration (Source: bizon-tech.com) (Source: viperatech.com). Those configuration- and seller-specific listings conflict with a separate reported $250,000 to $400,000 range for complete Blackwell 8-GPU HGX servers (Source: amcompute.com), so they should not be treated as mutually consistent or as a single indicative market price. A single DGX B200 system draws approximately 14.3 kilowatts (kW) at peak with 1,440 gigabytes (GB) of GPU memory (Source: nvidia.com), while a full GB200 NVL72 rack draws roughly 120 kW (Source: spheron.network) and is reported by Wolfe Research's supply-chain analysis to sell for about $3 million, with the newer GB300 NVL72 at roughly $4.3

million (Source: au.finance.yahoo.com). NVIDIA has not published official list prices for the DGX B200, DGX B300, or either NVL72 rack (Source: amcompute.com), so every rack- and system-level dollar figure in this report is a market estimate from a named analyst or reseller, not a manufacturer price, and is flagged as such.

The decision between platforms reduces to three variables: how many GPUs a workload's largest single job needs inside one memory-coherent domain, how much control an organization wants over CPU, chassis, and support, and how much liquid-cooling and power infrastructure a given data center can support. Independent MLPerf Training v5.0 results, verified by MLCommons, showed GB200 NVL72 delivering up to 2.6 times more performance per GPU than the prior Hopper generation, and completing Llama 3.1 405B pretraining 2.2 times faster than Hopper at the same 512-GPU scale (Source: developer.nvidia.com). Hyperscalers moved quickly: Microsoft Azure brought its ND GB200 v6 virtual machines to general availability atop a 4,000-GPU GB200 Grace Blackwell cluster (Source: techcommunity.microsoft.com), and CoreWeave became one of the first clouds to put GB200 NVL72 online at scale for customers including Cohere, IBM, and Mistral AI (Source: blogs.nvidia.com). Most enterprise buyers running single-model workloads never need a full rack, and most data center GPU buyers overall end up on HGX-based OEM servers rather than DGX or rack-scale systems (Source: amcompute.com).

NVIDIA's Data Center segment posted record revenue of $75.2 billion in its first fiscal quarter of 2027 (the three months ended April 26, 2026), up 92% year over year (Source: nvidianews.nvidia.com), and Reuters has described NVIDIA as commanding "more than 80% of the AI chip market" (Source: reuters.com), underscoring how much of the AI infrastructure economy runs through these four product families. Looking forward, NVIDIA projects that its Vera Rubin NVL72 platform, built on the third-generation MGX rack design, will deliver AI inference at one-tenth the cost per million tokens of GB200 NVL72. NVIDIA states that this workload-specific comparison is subject to change and is based on Kimi-K2-Thinking using 32K/8K input/output sequence lengths (Source: nvidia.com), meaning the calculus this report walks through, which platform fits which workload, will keep shifting generation to generation even as the four-way taxonomy of chip, baseboard, turnkey system, and rack-scale supercomputer persists.

Introduction and Background

Buyers researching NVIDIA's data center hardware quickly run into four overlapping product names describing different layers of the same physical stack rather than four competing choices at the same level. The confusion is compounded by NVIDIA's practice of reusing GPU-generation names across chip, baseboard, and system product lines; a search for "NVIDIA B200" surfaces posts about the B200 chip itself, HGX B200 baseboards, and DGX B200 turnkey systems, three distinct products despite the shared name (Source: modal.com). Understanding the difference matters because the four platforms carry very different prices, degrees of vendor lock-in, and power and cooling requirements; choosing incorrectly can mean paying rack-scale prices for a workload that never needed more than eight GPUs, or bottlenecking a genuinely rack-scale training job on hardware never designed to act as one coherent memory domain.

NVIDIA's turnkey-systems lineage traces to April 2016, when the company unveiled the original DGX-1, marketed as "the world's first deep learning supercomputer," combining eight Tesla P100 GPUs in a single 3U chassis with claimed throughput equivalent to 250 x86 servers (Source: nvidianews.nvidia.com). That original DGX-1 sold for $129,000, a price MIT Technology Review reported as steep enough that some large customers preferred buying individual GPUs and integrating their own clusters rather than paying a premium for a sealed system (Source: technologyreview.com), a convenience-versus-cost tension that persists in the DGX-versus-HGX decision a decade later. NVIDIA founder Jensen Huang personally delivered one of the first DGX-1 units to a then-small research lab called OpenAI in 2016, a moment NVIDIA later said "launched the era of AI supercomputers" (Source: nvidianews.nvidia.com).

HGX emerged as the industry-facing counterpart: rather than selling every 8-GPU system directly, NVIDIA ships HGX baseboards to OEMs including Dell, HPE, Supermicro, and Lenovo, who cannot modify the baseboard itself but choose everything else around it, chassis height, cooling, host CPU, storage, and networking (Source: amcompute.com). By May 2023, NVIDIA recognized that neither the fixed DGX appliance nor the high-end HGX baseboard served the long tail of data centers wanting accelerated computing in more varied server shapes, and unveiled MGX, a modular reference architecture letting manufacturers mix GPUs, CPUs, and networking across more than 100 possible server variants while cutting OEM development costs by up to three-quarters and shrinking development time to as little as six months (Source: nvidianews.nvidia.com). The rack-scale NVL72 systems arrived with the Blackwell architecture in March 2024, physically wiring 72 GPUs into a single non-blocking NVLink domain, and the GB200 compute tray itself is built on the MGX design (Source: developer.nvidia.com), collapsing what looked like four separate product lines into a layered stack. This report examines each platform in turn, compares them on specification and cost, reviews the quantitative evidence on performance and market adoption, walks through named real-world deployments, and closes with guidance on matching a workload to the right layer of NVIDIA's stack as of July 2026.

NVIDIA DGX: The Turnkey Appliance

Capabilities

NVIDIA DGX is the company's own sealed, fixed-configuration AI system, sold complete with a validated hardware stack, an integrated software layer, and direct enterprise support, rather than the baseboard-plus-integrator model HGX uses (Source: amcompute.com). The prior Blackwell 8-GPU model, **DGX B200**, packages eight NVIDIA Blackwell GPUs with 1,440 GB of total GPU memory and 64 TB/s of HBM3e bandwidth (Source: nvidia.com), delivering 3 times the training performance and 15 times the inference performance of the previous-generation DGX H100 (Source: nvidia.com). The current Blackwell Ultra 8-GPU DGX system, **DGX B300**, boosts dense FP4 performance 1.5 times and attention performance 2 times over DGX B200 (Source: nvidia.com) and is the first DGX system deployable inside NVIDIA MGX racks (Source: nvidia.com). NVIDIA specifies eight Blackwell Ultra SXM GPUs with 2.1 TB of total GPU memory, eight ConnectX-8 VPI ports, and two dual-port BlueField-3 DPUs. Its system-level FP4 Tensor Core specification is 144 PFLOPS sparse or 108 PFLOPS dense (NVIDIA DGX B300 specifications. At rack scale, **DGX GB200** repackages the NVL72 architecture as a turnkey NVIDIA-supported product, scaling to 72 Blackwell GPUs and 36 Grace CPUs per rack with up to 30.2 TB of total fast memory (Source: nvidia.com).

Every DGX system ships with NVIDIA AI Enterprise software, NVIDIA Mission Control for AI factory operations, and NVIDIA's DGX OS, plus three years of enterprise business-standard hardware and software support (Source: nvidia.com); DGX systems always include two BlueField DPUs, while OEM HGX builds vary DPU count and model by design (Source: amcompute.com). NVIDIA's DGX SuperPOD extends the same philosophy to multi-rack scale, described as "leadership-class AI infrastructure" configurable with any DGX system and scalable to tens of thousands of GPUs (Source: nvidia.com), and is delivered pre-built, cabled, and factory-tested so months of integration compress to weeks (Source: amcompute.com), while DGX BasePOD is a simpler reference architecture combining DGX systems with third-party storage without the full NVIDIA networking stack (Source: amcompute.com).

Adoption

DGX is an NVIDIA-supported option for deploying AI infrastructure. NVIDIA reports 8 of the top 10 global telecommunications companies, 7 of the top 10 global pharmaceutical companies, all 10 of the top 10 global car manufacturers, and 9 of the top 10 U.S. government institutions run DGX systems (Source: nvidia.com). Named customers include Lockheed Martin, which centralized machine-learning operations on a DGX SuperPOD (Source: nvidia.com) and Sony, which installed a DGX SuperPOD in its R&D center (Source: nvidia.com). Independent research corroborates DGX's broader footprint, citing Shell, BMW, Sony, and Lockheed Martin among its customer base and separately noting 10 of the top 10 global car manufacturers run DGX-based infrastructure (Source: amcompute.com).

Strengths and Limitations

DGX's chief strength is reduced integration risk: one NVIDIA-validated configuration, one support contract, and pre-tuned performance out of the box. Its chief limitation is cost relative to HGX-based OEM alternatives: full 8-GPU DGX B300 systems are quoted by resellers as high as $400,000 to $500,000 (Source: spheron.network), and DGX generally "costs more than an equivalent OEM server because it includes NVIDIA's software stack and direct support," a premium buyers pay for validation rather than materially different silicon (Source: amcompute.com).

NVIDIA HGX: The OEM Baseboard

Capabilities

HGX is a licensable 8-GPU baseboard, not a complete system: it bundles NVIDIA GPUs in the SXM form factor with NVLink interconnect and NVSwitch routing chips onto a single board that OEMs wrap with their own CPUs, chassis, cooling, storage, and support (Source: amcompute.com). As of July 2026, NVIDIA's specification page lists HGX in a single baseboard configuration available with eight Rubin, Blackwell, or Blackwell Ultra SXM GPUs (Source: nvidia.com). The current-generation **HGX B300** (Blackwell Ultra) offers 2.3 TB of total memory (eight 288 GB GPUs), 144 sparse PFLOPS of FP4 Tensor Core performance, and 14.4 TB/s of total NVLink bandwidth (Source: nvidia.com); at the per-GPU level, Blackwell Ultra pushes dense NVFP4 compute to 15 PFLOPS, a 1.5 times increase over standard Blackwell and 7.5 times increase over Hopper (Source: developer.nvidia.com). Supermicro's HGX B300 systems can be packed at up to 144 liquid-cooled GPUs in a single 21-inch OCP ORV3 rack, sustaining each GPU at up to 1,100 watts (W) of thermal design power (Source: supermicro.com). The prior-generation **HGX B200** offers 1.4 TB of memory at the same 144 sparse PFLOPS FP4 figure (Source: nvidia.com), and Supermicro's HGX B200 air-cooled systems deliver up to 15 times the inference and 3 times the training performance of the prior Hopper generation (Source: supermicro.com). Looking ahead, the next-generation **HGX Rubin NVL8** pairs eight Rubin GPUs with either a Vera CPU or an x86 baseboard, and NVIDIA projects it will deliver up to 10 times more "token

factory" throughput than HGX B200 while matching its training performance with four times fewer GPUs (Source: nvidia.com). NVIDIA has shipped HGX boards since 2017 across four GPU generations, with each generation bringing more GPU memory and faster NVLink bandwidth (Source: amcompute.com).

Adoption

HGX lets OEMs such as Dell, HPE, Supermicro, and Lenovo build branded servers around NVIDIA's reference baseboard rather than requiring buyers to purchase a finished DGX system (Source: amcompute.com). Dell and Supermicro are described as the most common HGX OEMs, differentiated mainly by chassis height, cooling design, CPU vendor choice between Intel Xeon and AMD EPYC, and support tiers (Source: amcompute.com). Supermicro's front-I/O liquid-cooled HGX B200 system uses "DLC-2" direct liquid-cooling technology that the company states captures up to 92% of the heat generated by server components, enabling up to 40% data center power savings (Source: supermicro.com).

Strengths and Limitations

HGX's strength is choice: buyers select their preferred OEM, CPU architecture, storage, and support contract while running the same validated GPU baseboard NVIDIA itself uses inside DGX. Reseller listings illustrate the resulting price variability: one fully configured 8-GPU HGX B200 system starts at $483,091 (Source: bizon-tech.com), while another lists a comparable configuration at $792,000 (Source: viperatech.com), a roughly $300,000 spread driven by cooling, storage, and support differences rather than the GPU baseboard itself, since NVIDIA fixes GPU count, NVSwitch layout, and power delivery to the GPUs regardless of which OEM assembles the box (Source: amcompute.com). The limitation is integration burden: unlike DGX, the OEM is responsible for validating the full stack, and unlike MGX, HGX offers a single high-end configuration rather than dozens of chassis and cooling variants.

NVIDIA GB200 / GB300 NVL72: The Rack-Scale System

Capabilities

GB200 NVL72 is described by NVIDIA as "an exascale computer in a single rack," interconnecting 72 Blackwell GPUs and 36 Grace CPUs through the NVLink Switch System to deliver 130 TB/s of low-latency GPU-to-GPU communication (Source: nvidia.com). Independent hardware analysis puts the rack at 72 B200-class GPUs, 36 Grace Arm central processing units (CPUs), 13.4 TB of unified GPU memory, and 1.44 exaflops of sparse FP4 compute in a single liquid-cooled enclosure (Source: spheron.network). The system is built from 18 compute trays, each holding two Grace CPUs and four Blackwell GPUs, and nine NVLink switch trays, with the compute tray design itself based on the MGX reference architecture and delivering 80 petaflops of AI performance and 1.7 TB of fast memory per tray (Source: developer.nvidia.com). Each Grace Blackwell Superchip pairs two Blackwell GPUs with one Grace CPU over a 900 gigabytes per second (GB/s) NVLink-Chip-to-Chip (C2C) interconnect, giving applications coherent access to a unified memory space rather than treating CPU and GPU memory as separate pools connected by PCIe (Source: developer.nvidia.com). Fifth-generation NVLink provides 1.8 TB/s of bidirectional GPU-to-GPU bandwidth and can extend NVLink domains up to 576 GPUs in a non-blocking compute fabric (Source: developer.nvidia.com). NVIDIA also sells a smaller **GB200 NVL4** module, four GPUs plus two Grace CPUs bridged over NVLink, compatible with liquid-cooled MGX modular servers for converged HPC and AI workloads that do not need a full rack (Source: nvidia.com).

The successor **GB300 NVL72**, built on Blackwell Ultra silicon, integrates 72 Blackwell Ultra GPUs and 36 Grace CPUs and is described by NVIDIA as "purpose-built for test-time scaling inference and AI reasoning tasks," delivering up to a 50 times overall increase in AI factory output compared with Hopper-based platforms (Source: nvidia.com). Supermicro's GB300 NVL72 implementation uses a 250 kW liquid-to-liquid cooling distribution unit (CDU) or a 200 kW liquid-to-air sidecar option and doubles networking throughput to 800 Gb/s per GPU compared with the prior generation (Source: supermicro.com).

Adoption

Rack-scale NVL72 deployment is concentrated among hyperscale cloud providers and AI labs needing coherent memory across dozens of GPUs. CoreWeave became "one of the first cloud providers to bring NVIDIA GB200 NVL72 systems online for customers at scale," with Cohere, IBM, and Mistral AI training and deploying models on the platform (Source: blogs.nvidia.com); the company also became the first cloud provider to deploy GB300 NVL72, with initial systems announced in July 2025 and cloud instances generally available from August 19, 2025 (Source: spheron.network). Azure followed with the first large-scale GB300 NVL72 cluster for OpenAI workloads in October 2025, and AWS launched EC2 P6e-GB300

UltraServers with general availability on December 2, 2025 (Source: spheron.network). Dell Technologies shipped the first Dell PowerEdge XE9712 racks with GB200 NVL72 to support CoreWeave's Cloud Services Platform, part of a strategic agreement Dell described as equipping "CoreWeave's enterprise customers with the speed and scalability to accelerate AI-driven projects" (Source: dell.com).

Strengths and Limitations

NVL72's chief strength is treating dozens of GPUs as one coherent accelerator, directly benefiting models too large to fit a single 8-GPU server's combined memory; the 13.4 TB of pooled GPU memory allows a 671-billion-parameter model to run entirely within one rack (Source: spheron.network). NVIDIA's own comparisons against H100-based infrastructure claim a 30 times speedup in real-time large language model (LLM) inference, a 4 times speedup in LLM training, a 25 times gain in energy efficiency, and an 18 times gain in data processing throughput (Source: nvidia.com). Its principal limitations are infrastructure prerequisites most buyers cannot casually meet: a full rack weighs approximately 1.36 metric tons and occupies a non-standard 48U OCP Open Rack V3 form factor rather than a conventional 19-inch data center rack (Source: spheron.network), draws roughly 120 kW to 132 kW under full load requiring dedicated three-phase power (Source: spheron.network), and mandates direct liquid cooling since standard rear-door heat exchangers built for 30 to 40 kW racks cannot handle the load, according to a Supermicro rack-scale datasheet describing the design as an exascale system delivering up to 25 times more energy efficiency than the previous generation (Source: supermicro.com).

NVIDIA MGX: The Modular Reference Architecture

Capabilities

MGX is not a GPU-carrying product; it is an open, modular reference architecture covering chassis mechanicals, rack form factors, power distribution, and connectors, compatible with the Open Compute Project (OCP) and Electronic Industries Alliance rack standards (Source: nvidianews.nvidia.com). It provides "an open modular reference architecture that enables OEMs, ODMs, and ecosystem partners to build accelerated systems faster," spanning everything "from single-node servers to rack-scale AI factories" (Source: nvidia.com). At launch in May 2023, MGX supported 1U, 2U, and 4U chassis in air- and liquid-cooled variants, the then-current NVIDIA GPU portfolio, Grace and GH200 Grace Hopper Superchips or x86 CPUs, and BlueField-3 DPU or ConnectX-7 networking (Source: nvidianews.nvidia.com). As of July 2026, NVIDIA has advanced to a "third-generation MGX rack architecture" unifying compute and networking servers, cooling, power, and connectors into one common rack-scale design (Source: nvidia.com), and MGX has become the mechanical home for non-GPU accelerators too: NVIDIA's product page notes MGX "supports NVIDIA Groq 3 LPX racks scaling 256 LPU accelerators in a single liquid-cooled MGX rack" (Source: nvidia.com). MGX is explicitly designed for multi-generational compatibility, so a chassis built for one GPU generation can be reused across future generations with less redesign than a one-off custom platform, protecting an OEM's engineering investment across hardware cycles (Source: amcompute.com).

Adoption

MGX launched with ASRock Rack, ASUS, GIGABYTE, Pegatron, QCT, and Supermicro as first adopters, with QCT and Supermicro first to market in August 2023 (Source: nvidianews.nvidia.com). SoftBank Corp. announced plans to use MGX to dynamically allocate GPU resources between generative AI and 5G workloads across multiple hyperscale data centers in Japan (Source: nvidianews.nvidia.com). NVIDIA's Vera Rubin NVL72 materials cite "over 80 MGX ecosystem partners" supporting the current rack generation (Source: nvidia.com). Partners including Supermicro, QCT, ASUS, and GIGABYTE build and sell the actual MGX-based servers to end customers (Source: amcompute.com).

Strengths and Limitations

MGX's strength is its open modular reference architecture: NVIDIA says it enables OEMs, ODMs, and ecosystem partners to build accelerated systems faster, while its common rack-scale design can reduce redesign across hardware generations (Source: nvidia.com). Its limitation is scope: MGX is "not a replacement for HGX," since HGX targets 8-GPU training nodes with maximum interconnect bandwidth while MGX covers inference servers, mixed CPU-GPU nodes, and smaller deployments (Source: amcompute.com); it describes the chassis and power layer, not the compute silicon, so buyers still separately choose which HGX-class baseboard or superchip module goes inside an MGX-compliant rack.

Feature Comparison

Table 1 below summarizes the specification, cost, and control trade-offs across all four platforms as of July 2026, drawing on NVIDIA's own published specifications alongside multiple independent OEM, integrator, and analyst sources for pricing that NVIDIA does not officially publish.

DIMENSION	HGX (B200/B300 BASEBOARD)	DGX (B200/B300 SYSTEM)	GB200/GB300 NVL72 (RACK-SCALE)	MGX (MODULAR REFERENCE ARCHITECTURE)
What it is	GPU baseboard hardware for OEM integration (Source: amcompute.com)	NVIDIA's fixed hardware system built on the HGX baseboard (Source: amcompute.com)	Full liquid-cooled 72-GPU rack, not a baseboard (Source: amcompute.com)	Server reference architecture / design spec (Source: amcompute.com)
GPU count per unit	8 per current baseboard (NVIDIA HGX Platform	8 (system) or 72 (DGX GB200/GB300) (Source: nvidia.com)	72, single NVLink domain (Source: nvidia.com)	Varies, 1-GPU servers to 72-GPU racks (Source: amcompute.com)
GPU memory (total)	1.4 TB (B200) / 2.3 TB (B300) (Source: nvidia.com)	1,440 GB (B200) / 2.3 TB (B300) (Source: nvidia.com)	13.4 TB (Source: spheron.network)	Set by installed GPU module
NVLink bandwidth	14.4 TB/s aggregate (Source: nvidia.com)	14.4 TB/s (B200/B300 system) (Source: amcompute.com)	130 TB/s rack-wide, all-to-all (Source: spheron.network)	Set by GPU/baseboard installed
Power draw	Server-dependent; B300 GPU up to 1,100 to 1,400W each (Source: supermicro.com)	~14.3 kW (DGX B200) (Source: nvidia.com); ~14 kW (DGX B300) (Source: spheron.network)	120 to 132 kW per rack under full load (Source: spheron.network)	Set by configuration
Cooling	OEM choice; DLC-2 liquid captures 92% of heat (Source: supermicro.com)	Air or liquid depending on model	Mandatory direct liquid cooling (Source: spheron.network)	1U to 4U, air or liquid (Source: nvidianews.nvidia.com)
Buyer / integrator	OEMs: Dell, HPE, Supermicro, Lenovo (Source: amcompute.com)	Direct from NVIDIA or select partners	Hyperscalers, large clouds (CoreWeave, Azure, AWS) (Source: spheron.network)	OEM/ODM partners including Supermicro, QCT, ASUS (Source: amcompute.com)
Indicative price (market estimate, not NVIDIA list price)	$250,000 to $400,000 (Blackwell 8-GPU) (Source: amcompute.com)	$400,000 to $500,000 (DGX B300 system) (Source: spheron.network)	~$3 million (GB200) to $4.3 million (GB300), Wolfe Research estimate (Source: au.finance.yahoo.com)	Not separately priced; adds to GPU/CPU cost

The comparison shows a substantial step in deployment scope moving from HGX/DGX up to NVL72: an 8-GPU system is a server-level deployment, while NVL72 is a liquid-cooled 72-GPU rack-scale system with additional NVLink-switch and Grace-CPU infrastructure. Actual price, power, cooling, and installation requirements are configuration-specific and should be confirmed in a current vendor quote and facility specification. MGX does not appear as a separately priced product because it is a reference architecture rather than a finished system.

Performance and Benchmarks

Independent, standardized performance evidence comes primarily from MLPerf, the benchmark suite maintained by MLCommons, an industry consortium publishing verified training and inference results across vendors. In the MLPerf Training v5.0 round, published June 4, 2025, Blackwell-based submissions delivered the fastest time to train across all seven benchmarks in the suite, spanning LLM pretraining, LLM fine-tuning, text-to-image generation, recommender systems, graph neural networks, natural language processing, and object detection (Source: developer.nvidia.com). This round marked the first MLPerf Training submissions using the GB200 NVL72 rack-scale system (Source: developer.nvidia.com).

Table 2 below summarizes the headline MLPerf Training v5.0 results that anchor the performance claims made throughout this report, comparing Blackwell-based submissions against the prior Hopper generation across the benchmark suite's most-cited tests.

BENCHMARK	HOPPER TIME-TO-TRAIN	BLACKWELL (GB200 NVL72) TIME-TO-TRAIN	SPEEDUP
Llama 3.1 405B pretraining (512 GPUs)	269.12 min	121.09 min	2.2x (Source: developer.nvidia.com)
Llama 2 70B LoRA fine-tuning (8 GPUs)	27.93 min	11.14 min	2.51x (Source: developer.nvidia.com)
Stable Diffusion v2 pretraining (8 GPUs)	33.97 min	12.86 min	2.64x (Source: developer.nvidia.com)
R-GAT graph neural network (8 GPUs)	11.18 min	4.97 min	2.25x (Source: developer.nvidia.com)

These four results, drawn from the same MLPerf Training v5.0 round and independently verified by MLCommons Association, show that Blackwell results vary by workload and submission configuration. The 512-GPU Llama pretraining result used GB200 NVL72 at scale. The 8-GPU fine-tuning, image-generation, and graph results also used eight Blackwell GPUs as part of a GB200 NVL72 system, not standalone HGX or DGX hardware; NVIDIA describes the fine-tuning submission that way (NVIDIA MLPerf Training v5.0 analysis. Those results support Blackwell-versus-Hopper comparisons for the tested configurations, but do not establish equivalent performance for every HGX or DGX configuration. Platform selection should therefore be validated with workload-specific measurements and deployment constraints.

On the headline Llama 3.1 405B pretraining benchmark, GB200 NVL72 completed training 2.2 times faster than Hopper-generation hardware at the same 512-GPU submission scale (Source: developer.nvidia.com), reaching up to 1,960 teraFLOPS (TFLOPS) of training throughput (Source: developer.nvidia.com).

MLPerf Inference results extend this pattern into the Blackwell Ultra generation. In MLPerf Inference v6.0, published in April 2026, systems powered by Blackwell Ultra GPUs delivered the highest throughput across the widest range of models and scenarios, and on the DeepSeek-R1 reasoning model specifically, Blackwell Ultra systems delivered 2.5 million tokens per second, up to 2.7 times higher than Blackwell Ultra's debut submissions just six months earlier, driven by software updates to NVIDIA's TensorRT-LLM inference library (Source: nvidia.com). Independent third-party benchmarking from SemiAnalysis's InferenceX project found Blackwell Ultra delivering inference at $0.24 per million tokens at 102 tokens-per-second per user on DeepSeek-R1 using NVIDIA Dynamo and TensorRT-LLM (Source: nvidia.com), and separately measured standard Blackwell systems at approximately $0.02 per million tokens on the GPT-OSS-120B model, roughly 4.5 times cheaper than Hopper-powered systems at $0.09 per million tokens (Source: nvidia.com).

Real-world deployment data corroborates the MLPerf results directionally. Microsoft Azure measured over 860,000 tokens per second of inference throughput on a single GB200 NVL72 rack running the Llama 2 70B model, a 9 times increase per rack compared with the prior-generation ND H100 v5 virtual machine series (Source: techcommunity.microsoft.com). Cohere reported up to 3 times more training performance for 100-billion-parameter models on GB200 NVL72 compared with Hopper GPUs, even before Blackwell-specific software optimizations (Source: blogs.nvidia.com). At the physical-simulation level, GB200 NVL72 speeds database join queries up to 18 times faster than CPUs and 6 times faster than H100 GPUs on TPC-H-derived benchmarks, and Cadence SpectreX circuit simulations are projected to run 13 times faster on a GB200 Grace Blackwell Superchip than on a traditional CPU (Source: developer.nvidia.com). Independent GPU cloud analysis suggests that at rack-scale utilization on a 671-billion-parameter model, GB200 NVL72's compute and NVLink advantage can cut inference cost 3 to 5 times compared with an equivalent number of Hopper-

generation nodes (Source: spheron.network). These MLPerf figures are independently verified by MLCommons Association, distinguishing them from vendor-projected claims pending independent benchmark confirmation, such as HGX Rubin NVL8's stated 10 times token-factory throughput improvement over HGX B200 (Source: nvidia.com).

Data Analysis and Evidence

The commercial scale behind these four platforms is substantial and growing quickly. NVIDIA's Data Center segment, encompassing DGX, HGX, MGX-based, and NVL72 system revenue, generated a record $75.2 billion in the first quarter of fiscal year 2027 (three months ended April 26, 2026), up 92% year over year and up 21% sequentially (Source: nvidianews.nvidia.com). Total company revenue reached $81.6 billion, up 85% year over year, with a GAAP gross margin of 74.9% (Source: nvidianews.nvidia.com). Reuters reported that NVIDIA "commands more than 80% of the AI chip market" (Source: reuters.com), and previously reported NVIDIA's chief financial officer Colette Kress saying in mid-2024 that demand for Blackwell chips could exceed supply "well into next year" (Source: reuters.com).

Analyst supply-chain research provides visibility into the rack-scale segment specifically. Wolfe Research's supply-chain checks found GPU baseboards account for roughly 75% of a finished NVL72 rack's total price (Source: au.finance.yahoo.com), and estimated Blackwell rack shipments reached roughly 1,000 units per week by the end of calendar 2025, implying 50,000 to 60,000 racks shipped across 2026 (Source: au.finance.yahoo.com). Independently, TrendForce forecasts global AI server shipments growing more than 28% year over year in 2026, outpacing 12.8% growth in total server shipments (Source: trendforce.com), with GPU-based systems accounting for 69.7% of AI server shipments and NVIDIA's GB300-based systems expected to drive most of that volume (Source: trendforce.com). Grand View Research separately estimates the global AI server market at $157.0 billion in 2026, up from $131.7 billion in 2025, projecting growth to $598.1 billion by 2033 at a 21.2% compound annual growth rate (Source: grandviewresearch.com), with GPU-based servers holding the largest revenue share at over 53.0% in 2025 (Source: grandviewresearch.com) and North America holding the largest regional share at 38.2% (Source: grandviewresearch.com).

Table 3 below consolidates the pricing and market-size estimates gathered from independent analysts and market-research firms, since NVIDIA does not publish list prices for any of the four platforms covered in this report.

METRIC	ESTIMATE	SOURCE AND DATE
GB200 NVL72 rack price	~$3 million	Wolfe Research, 2026 (Source: au.finance.yahoo.com)
GB300 NVL72 rack price	~$4.3 million	Wolfe Research, 2026 (Source: au.finance.yahoo.com)
1,000-GPU B200 deployment (all-in)	$45 million to $50 million	Introl deployment guide, 2026 (Source: introl.com)
1,000-GPU GB200 deployment (all-in)	~$200 million	Introl deployment guide, 2026 (Source: introl.com)
Global AI server market, 2026	$157.0 billion	Grand View Research, 2026 (Source: grandviewresearch.com)
Global AI infrastructure spending, 2026 (forecast)	$487 billion	IDC, via digitalapplied.com, 2026 (Source: digitalapplied.com)
AI server shipment growth, 2026 (YoY)	+28%	TrendForce, January 2026 (Source: trendforce.com)

The spread between these figures illustrates why buyers should treat any single dollar estimate cautiously. Wolfe Research's approximately $3 million GB200 NVL72 rack estimate and Introl's approximately $200 million estimate for a 1,000-GPU GB200 deployment are materially divergent when normalized by GPU count: 1,000 GPUs equal about 14 NVL72 racks, and $3 million per rack implies roughly $42 million in rack hardware before infrastructure. Introl does not provide an itemized reconciliation of its total, so the two estimates have non-comparable scopes and should not be treated as mutually validating. The market-size and infrastructure-spending rows also use different methodologies and should not be added together or treated as cross-validating. Read individually, each row is a named, dated estimate; read as a set, they indicate a broader AI server market forecast to grow at double-digit rates through the end of the decade.

On pricing specifically, published market estimates diverge meaningfully depending on source and configuration, a discrepancy this report notes rather than resolves. Independent deployment-cost analysis put a complete 1,000-GPU B200 buildout at $45 million to $50 million including infrastructure, versus roughly $200 million for an equivalent 1,000-GPU GB200 deployment once the 40% to 50% infrastructure premium for power and cooling upgrades is included (Source: introl.com). The Introl analysis separately described the GB200 NVL72 rack as approaching $3 million, "among the most expensive computing systems ever mass-produced" (Source: introl.com), similar to Wolfe Research's independent $3 million rack estimate. However, Introl does not itemize how that rack estimate relates to its separate approximately $200 million all-in estimate for 1,000 GPUs, so those figures should not be treated as reconciled. On next-generation pricing, Wolfe Research assumes NVIDIA's Rubin Ultra racks will price around $10 million, noting each additional $1 million per rack above its estimate could add $10 billion to $12 billion in annual revenue at projected shipment volumes (Source: au.finance.yahoo.com). By contrast, IDC's separate AI infrastructure spending tracker measured full-year 2025 AI infrastructure spending, including servers, storage, and networking, at $318 billion worldwide, more than double 2024's $153 billion, and forecasts 2026 spending will reach $487 billion, a 53% year-over-year increase, of which servers alone made up 97.6% of the total in the fourth quarter of 2025 (Source: digitalapplied.com). Gartner separately forecasts total worldwide AI spending, across the full stack of hardware, software, and services, at $2.59 trillion in 2026, of which AI infrastructure alone accounts for more than 45% (Source: digitalapplied.com). Readers should treat all NVL72 and DGX system-level dollar figures in this report as third-party market estimates rather than official NVIDIA list prices, since NVIDIA has not published retail pricing for either the rack-scale systems or standalone DGX units as of July 2026.

Case Studies and Real-World Examples

CoreWeave: First Cloud-Scale GB200 NVL72 Deployment

CoreWeave was among the first cloud providers to bring GB200 NVL72 systems online for customers at production scale, with the deployment underpinning work for three named AI companies. Cohere used the platform to develop its enterprise AI agent product, North, reporting up to 3 times faster training for 100-billion-parameter models compared with previous-generation Hopper hardware even before Blackwell-specific optimization (Source: blogs.nvidia.com). IBM used one of the first GB200 NVL72 deployments, scaling to thousands of Blackwell GPUs, to train its open-source Granite model family, which underpins the IBM watsonx Orchestrate agent platform (Source: blogs.nvidia.com). Paris-based Mistral AI used the CoreWeave GB200 NVL72 deployment to accelerate development of models including Mistral Large, reporting a 2 times performance improvement for dense model training without any additional tuning (Source: blogs.nvidia.com). Dell Technologies supplied the physical racks for this deployment, shipping the first Dell PowerEdge XE9712 GB200 NVL72 servers, with CoreWeave co-founder and Chief Strategy Officer Brian Venturo saying Dell is "a strategic partner when it comes to delivering world-class performance at scale" (Source: dell.com).

Microsoft Azure: ND GB200 v6 General Availability

Microsoft brought its Azure ND GB200 v6 virtual machine series, built on GB200 NVL72, to general availability atop what it described as one of the first cloud-hosted 4,000-GPU GB200 Grace Blackwell supercomputing clusters (Source: techcommunity.microsoft.com). Ian Buck, NVIDIA's vice president of Hyperscale and HPC, said "the NVIDIA GB200 NVL72, with its unparalleled performance and connectivity, tackles the most complex AI workloads, enabling businesses to innovate faster and more securely" (Source: techcommunity.microsoft.com). Azure separately offers GB200 NVL72 capacity as one of several major hyperscale providers, alongside CoreWeave, Oracle Cloud, and Google Cloud, all listing GB200 NVL72 availability by early 2026 (Source: spheron.network).

Eli Lilly: DGX SuperPOD With DGX B300 for Drug Discovery

Pharmaceutical company Eli Lilly deployed what NVIDIA describes as its "largest AI factory for drug discovery," an NVIDIA DGX SuperPOD built with DGX B300 systems intended to enable breakthroughs in genomics, medicine, and molecular design (Source: nvidia.com). The deployment illustrates the 8-GPU DGX product line's role at the largest scale of enterprise deployment, using DGX SuperPOD's turnkey multi-rack architecture rather than a bespoke rack-scale NVL72 build, since Lilly's computational drug-discovery workloads are well suited to many parallel DGX B300 nodes rather than requiring a single 72-GPU coherent memory domain.

SoftBank Corp.: MGX for Dynamic 5G and AI Resource Allocation

SoftBank Corp. was named as a launch partner for MGX in 2023, planning to use the modular platform's flexibility to dynamically shift GPU resources between generative AI workloads and 5G network functions across multiple hyperscale data centers in Japan (Source: nvidianews.nvidia.com). Separately, on the DGX side of NVIDIA's portfolio, SoftBank's Ashiq Khan, Vice President and Head of the Unified Cloud and Platform Division, said the company is "pioneering homegrown LLMs for the Japanese language, aiming at 390 billion parameters," using DGX SuperPOD and the NVIDIA AI Enterprise software stack (Source: nvidia.com), illustrating how a single large enterprise customer can span multiple layers of NVIDIA's product taxonomy simultaneously.

University of Florida: DGX SuperPOD at Public-Research Scale

The University of Florida's HiPerGator AI cluster, built on NVIDIA DGX SuperPOD with Blackwell architecture, supports over 60% of the university's research projects and has served nearly 7,000 users, processing more than 33 million research requests in the preceding year (Source: nvidia.com). Chemistry professor Adrian Roitberg said the system "will allow researchers to perform quantum-accurate molecular simulations of proteins to help find cures to diseases like COVID-19," work he said "would've taken more than 6,000 years" without the platform, reducing it to a single day (Source: nvidia.com).

OpenAI: The DGX-1 Origin Point

Though it predates the current HGX/DGX/NVL72/MGX taxonomy by nearly a decade, NVIDIA's delivery of one of the first DGX-1 systems to OpenAI in 2016 remains the clearest illustration of the turnkey-appliance philosophy that still defines DGX today. NVIDIA later characterized the moment as one that "launched the era of AI supercomputers and unlocked the scaling laws that drive modern AI" (Source: nvidianews.nvidia.com), and the underlying DGX-1 pricing dynamic, a $129,000 sealed system priced at a premium over self-assembled equivalents but justified for organizations valuing speed of deployment over customization (Source: technologyreview.com), is structurally identical to the DGX-versus-HGX trade-off buyers evaluate in 2026.

Implications and Future Directions

The clearest structural trend is convergence rather than divergence: MGX, originally positioned as a lightweight alternative to HGX for smaller, more varied deployments, is now the physical chassis underneath NVIDIA's most powerful rack-scale systems. NVIDIA's own developer documentation states plainly that the GB200 compute tray "is based on the new NVIDIA MGX design" (Source: developer.nvidia.com), and by the DGX B300 generation, NVIDIA advertises that system as "deployable in NVIDIA MGX racks for the first time" (Source: nvidia.com). Buyers evaluating these four names in 2026 and beyond should expect the distinction to keep narrowing to two effective axes: how many GPUs need to sit in one coherent NVLink domain, an 8-GPU HGX/DGX-class decision versus a 72-GPU-plus NVL72-class decision, and who integrates and supports the final system, an OEM under an HGX or MGX-derived design versus NVIDIA directly under the DGX brand.

The next generation, Vera Rubin, is already reshaping the economics that motivate platform choice. NVIDIA states Vera Rubin NVL72, built on the third-generation MGX rack design, delivers AI training with one-fourth the GPUs and AI inference at one-tenth the cost per million tokens compared with today's GB200 NVL72 (Source: nvidia.com), and up to 10 times more tokens per megawatt when deployed alongside the new NVIDIA Groq 3 LPX inference-accelerator rack (Source: nvidia.com). Wolfe Research's supply-chain analysis projects Rubin-generation rack shipments reaching roughly 55,000 units in 2026, climbing to 55,000 Rubin racks and 15,000 Rubin Ultra racks by 2027, alongside slower growth for standalone HGX platforms (Source: au.finance.yahoo.com), a directional signal that rack-scale NVL72-class systems, not 8-GPU HGX or DGX boxes, are becoming NVIDIA's primary unit of sale even as smaller systems remain the more accessible and common entry point for most enterprise buyers. Consistent with that trajectory, Dell separately launched new servers powered by Blackwell Ultra chips in May 2025, offered in air- and liquid-cooled versions supporting up to 192 chips by default and up to 256 with customization, enabling AI model training reported as up to four times faster than the prior generation (Source: grandviewresearch.com). Power infrastructure, not silicon, increasingly gates who can adopt the rack-scale tier at all: NVIDIA's own roadmap materials describe a shift to 800 volt direct current (VDC) power distribution beginning in 2027 specifically to support megawatt-scale IT racks beyond what today's GB200 and GB300 NVL72 designs can draw (Source: nvidia.com), meaning the practical ceiling on rack-scale adoption over the next several years will be set as much by data-center electrical engineering as by GPU availability.

Frequently Asked Questions (FAQs)

What is the difference between NVIDIA DGX and HGX? HGX is a licensable 8-GPU baseboard that OEMs such as Dell, Supermicro, and HPE integrate into their own branded servers, while DGX is NVIDIA's own sealed, fixed-configuration system built on the same HGX baseboard but sold, supported, and serviced directly by NVIDIA (Source: amcompute.com).

What is NVIDIA NVL72? NVL72 refers to a rack-scale system, currently sold as GB200 NVL72 and GB300 NVL72, that connects 72 GPUs and 36 Grace CPUs into a single NVLink domain delivering 130 TB/s of GPU-to-GPU bandwidth, effectively making an entire rack behave as one very large accelerator rather than eight independent servers (Source: nvidia.com).

What is NVIDIA MGX? MGX is an open, modular reference architecture for server and rack design, covering chassis mechanicals, power distribution, and rack form factors, that lets manufacturers build over 100 different accelerated computing configurations from a common blueprint rather than designing each server from scratch (Source: nvidianews.nvidia.com).

How does GB200 NVL72 compare to HGX B200? GB200 NVL72 connects 72 GPUs in one NVLink domain at 130 TB/s of bandwidth (Source: spheron.network), versus HGX B200's 8 GPUs at 14.4 TB/s of aggregate NVLink bandwidth (Source: nvidia.com); GB200 NVL72 is designed for workloads that need more coherent GPU memory than eight GPUs can offer, while HGX B200 remains the more accessible unit for standard 8-GPU training and inference jobs.

Should I buy DGX or HGX? DGX suits buyers who want a single NVIDIA-supported, pre-validated appliance and are willing to pay a premium for reduced integration risk; HGX suits buyers who want to choose their own OEM, CPU vendor, and support contract, typically at a lower total cost for equivalent GPU hardware (Source: amcompute.com).

How is MGX different from DGX? MGX describes a modular rack and chassis architecture that any OEM can build to under an open specification compatible with the Open Compute Project (Source: nvidianews.nvidia.com), while DGX is a specific, NVIDIA-branded finished product; as of the DGX B300 generation, DGX systems can themselves be deployed inside MGX-compliant racks, making the two complementary rather than competing (Source: nvidia.com).

Which platform is best for AI training versus inference? MLPerf Training v5.0 results show GB200 NVL72's rack-scale NVLink domain delivering the largest training speedups for the biggest models, up to 2.6 times higher performance per GPU than Hopper-generation hardware (Source: developer.nvidia.com), while smaller models and most production inference workloads run efficiently on 8-GPU HGX or DGX systems; NVIDIA states DGX B200 alone delivers 15 times the inference performance of the prior-generation DGX H100 (Source: nvidia.com).

What are the main use cases for each NVIDIA platform? HGX and DGX B200/B300 fit single-team model training, fine-tuning, and inference at the 8-GPU scale; DGX SuperPOD and DGX BasePOD fit multi-rack enterprise AI Centers of Excellence, as with Eli Lilly's drug-discovery deployment (Source: nvidia.com); GB200/GB300 NVL72 fit frontier model pretraining and high-throughput reasoning inference at hyperscale cloud providers (Source: nvidia.com); and MGX fits organizations, from telecom operators to edge data centers, that need standardized, multi-generational-compatible chassis across varied hardware mixes, as with SoftBank's dynamic 5G and AI allocation plans (Source: nvidianews.nvidia.com).

Conclusion

DGX, HGX, GB200/GB300 NVL72, and MGX are four product families that address different layers of accelerated-system design. HGX is NVIDIA's 8-GPU baseboard platform for OEM integration; DGX is NVIDIA's turnkey system; GB200 and GB300 NVL72 are rack-scale systems that connect 72 GPUs in a single NVLink domain; and MGX is NVIDIA's modular reference architecture for partner-built systems. MGX is complementary to the other families, not a universal chassis layer: NVIDIA specifically identifies the GB200 compute tray as MGX-based and DGX B300 as deployable in MGX racks (Source: developer.nvidia.com, (Source: nvidia.com). For workloads that fit in an 8-GPU node, HGX-based OEM systems and DGX offer different integration and support models. Workloads requiring a much larger coherent GPU domain may justify NVL72-class infrastructure, subject to a current supplier quote and the necessary power, cooling, networking, and deployment capacity. MGX helps partners standardize designs across configurations and generations; it does not erase the distinctions among the four families (Source: nvidia.com).

References

- NVIDIA HGX Platform, nvidia.com
- NVIDIA MGX Platform, nvidia.com
- NVIDIA DGX Platform, nvidia.com
- NVIDIA GB200 NVL72, nvidia.com
- NVIDIA GB300 NVL72, nvidia.com
- NVIDIA Vera Rubin NVL72, nvidia.com

- NVIDIA DGX B200 / DGX B300 specifications, nvidia.com
- NVIDIA MGX Server Specification press release, nvidianews.nvidia.com
- NVIDIA First Quarter Fiscal 2027 Financial Results, nvidianews.nvidia.com
- NVIDIA Blackwell MLPerf Training v5.0 results, developer.nvidia.com
- NVIDIA GB200 NVL72 architecture blog, developer.nvidia.com
- Inside NVIDIA Blackwell Ultra technical blog, developer.nvidia.com
- CoreWeave GB200 NVL72 deployment announcement, blogs.nvidia.com
- Azure ND GB200 v6 general availability announcement, techcommunity.microsoft.com
- HGX, DGX, MGX: NVIDIA's Server Platforms, amcompute.com
- GB200 NVL72 Guide and NVIDIA B300 Blackwell Ultra Guide, spheron.network
- NVIDIA Blackwell HGX and GB200/GB300 NVL72 Solutions, supermicro.com
- Decoding Nvidia's Blackwell Products, modal.com
- Complete Guide to NVIDIA B200 vs GB200 Deployment, introl.com
- CoreWeave and Dell Technologies press release, dell.com
- Wolfe Research NVIDIA rack-scale pricing analysis via Yahoo Finance, au.finance.yahoo.com
- TrendForce Global AI Server Shipments Forecast, trendforce.com
- Grand View Research AI Server Market Report, grandviewresearch.com
- AI Spending Forecasts 2026, digitalapplied.com
- Reuters, Nvidia AI chip demand coverage, reuters.com
- MIT Technology Review, DGX-1 coverage, technologyreview.com
- BIZON X9000 G4 HGX B200 server listing, bizon-tech.com
- ViperaTech Supermicro HGX B200 server listing, viperatech.com

---

Tags: nvidia dgx, nvidia hgx, gb200 nvl72, nvidia mgx, blackwell gpu, ai server comparison, ai infrastructure, nvidia gb300 nvl72

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will GPUSmith or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.