NVIDIA GPU Form Factors Explained: SXM vs PCIe vs NVL72 vs MGX

Published July 22, 2026 47 min read

Failed to authenticate: OAuth session expired and could not be refreshedAt the base of the stack is the **physical form factor** of the GPU itself: how it plugs into a server. NVIDIA's data center GPUs, including the H100, H200, and Blackwell-generation B200, ship in two primary form factors. **PCIe** cards install into any standard server's PCI Express slot, the same physical interface used by network cards and storage controllers, and have been the default expansion mechanism in servers for two decades (Source: amcompute.com). **SXM** (Server PCI Express Module) is NVIDIA's proprietary alternative: a GPU module that mounts directly onto a specialized baseboard rather than a slot, wired for much higher power delivery and a dedicated high-speed interconnect called NVLink. An H100 SXM and an H100 PCIe card use identical silicon manufactured on the same process, so the performance gap between them comes entirely from how each chip connects to the rest of the system, not from the chip itself (Source: amcompute.com) (Source: deploybase.ai). Notably, the GPU die itself, code-named GH100 in the Hopper generation, is a separate concept from any of these packaging choices: the same GH100 silicon underlies the PCIe card, the SXM module, and the dual-GPU NVL board, which is why buyers occasionally see "GH100" and "H100" used interchangeably in technical documentation (Source: thundercompute.com).

Above the individual GPU sits the **baseboard and system layer**. NVIDIA formalized its modular, multi-GPU server approach around the transition from the Pascal generation (Tesla P100) to Volta (V100), when it began manufacturing and selling a standardized eight-GPU SXM baseboard directly to server vendors rather than leaving multi-GPU assembly to individual manufacturers (Source: servethehome.com). That standardized baseboard became **HGX**, and NVIDIA's own fully integrated server built on top of it became **DGX**, first released in 2016 (Source: civo.com). ServeTheHome, a hardware review outlet that has tracked NVIDIA's multi-GPU platforms since the Pascal generation, notes that from the H100 generation onward "what you cannot do... is just buy H100 GPUs and install them. NVIDIA makes the SXM baseboards now," reflecting how tightly NVIDIA controls the SXM supply chain compared to the open PCIe standard (Source: servethehome.com). Separately, **MGX** addresses a different problem: not maximum eight-GPU interconnect density, but flexible, mixed-vendor, mixed-scale system design, spanning single-node inference boxes up to full rack-scale AI factories (Source: amcompute.com). Finally, at rack scale, NVIDIA's **NVL** naming (NVL2, NVL4, NVL72, and beyond) describes how many GPUs share a single non-blocking NVLink domain, culminating in the 72-GPU GB200 and GB300 NVL72 racks that behave, from a software perspective, like one enormous GPU. This report walks through each of these layers in turn, quantifies the differences with vendor specifications and independent pricing data, and closes with real-world deployments and forward-looking guidance for infrastructure buyers evaluating NVIDIA hardware as of mid-2026.

PCIe vs SXM: The Foundational Form Factor Decision

What PCIe and SXM Actually Are

PCIe (Peripheral Component Interconnect Express) is an open industry standard expansion interface used throughout the computing industry, not something NVIDIA invented. A PCIe GPU is a dual-slot, air-cooled card that plugs into any server or workstation with a compatible slot, drawing power through standard connectors and communicating with the CPU and other GPUs over the shared PCIe bus (Source: amcompute.com). **SXM (Server PCI Express Module)** is NVIDIA's own high-performance socket, not compatible with any standard motherboard; SXM modules mount onto NVIDIA's HGX baseboard, a purpose-built board with dedicated high-speed wiring between GPU positions (Source: amcompute.com). DeployBase, a GPU infrastructure guide publisher, describes SXM and PCIe simply as "GPU connection types" that "determine how GPUs communicate with CPU and other components," a framing that usefully separates the connector question from the underlying chip's compute capability (Source: deploybase.ai). The naming itself, "Server PCI Express Module," can mislead buyers into thinking SXM is just another PCIe variant; in practice it is a distinct mechanical and electrical interface that cannot be retrofitted into a PCIe slot.

The defining functional difference is **GPU-to-GPU communication**. A PCIe GPU communicates with other GPUs either indirectly through the CPU or, at best, through an optional two-GPU NVLink bridge; it has no native path to a broader multi-GPU fabric. Server-parts.eu, a GPU hardware reseller, summarizes this in its comparison table as "Limited, via CPU, or NVLink bridge in pairs (model dependent)" for PCIe versus "NVLink + NVSwitch full mesh" for SXM (Source: server-parts.eu). An SXM GPU, mounted on an HGX baseboard, connects to every other GPU on that baseboard through NVLink and an NVSwitch crossbar chip, giving each GPU a direct, full-bandwidth path to every other GPU in the same node (Source: amcompute.com). For scaling across multiple servers, SXM-based systems typically add InfiniBand networking on top of NVLink, following NVIDIA's SuperPOD reference architecture, whereas PCIe systems generally do not use InfiniBand for inter-GPU scaling at all (Source: mbuzztech.com).

Quantified Performance and Power Differences

Table 1 below compares the NVIDIA H100, the generation for which both PCIe and SXM variants remain widely deployed and priced, across the specifications that matter most for workload planning.

SPECIFICATION	H100 PCIE	H100 SXM	H100 NVL (DUAL-GPU PCIE MODULE)
GPU memory	80 GB HBM3	80 GB HBM3	94 GB HBM3 per GPU (188 GB per pair) (Source: spheron.network)
Memory bandwidth	2.0 TB/s (Source: cloudgputracker.com)	3.35 TB/s	3.9 TB/s (Source: spheron.network)
TDP (thermal design power)	350 W (Source: mbuzztech.com)	700 W	700 to 800 W per two-GPU module (Source: spheron.network)
Multi-GPU interconnect	Optional 2-GPU NVLink bridge or PCIe bus only	NVLink + NVSwitch, full mesh, up to 8 GPUs (Source: server-parts.eu)	NVLink bridge, 2-GPU pair only, 600 GB/s (Source: spheron.network)
NVLink bandwidth (if present)	None (PCIe Gen5 x16 only)	900 GB/s per GPU (Source: cloudgputracker.com)	600 GB/s (pair total)
Typical purchase price (mid-2026)	$25,000 to $30,000 (Source: northflank.com)	$35,000 to $40,000 (Source: northflank.com)	Priced as HGX/DGX-class hardware
On-demand cloud rate (mid-2026)	$1.40/hr (Runcrate) (Source: runcrate.ai)	$1.50 to $5.12/hr (Source: runcrate.ai) (Source: inferencebench.io)	Not commonly listed separately
Server compatibility	Any standard PCIe server	HGX or DGX systems only (Source: server-parts.eu)	Any PCIe server with NVLink bridge support

The performance gap is not academic. Independent specification aggregators report the H100 SXM's memory bandwidth advantage at roughly 68% over PCIe (3.35 TB/s versus 2.0 TB/s) and a CUDA core count advantage of roughly 16% (16,896 versus 14,592 enabled cores) (Source: cloudgputracker.com) (Source: cloudgputracker.com). CloudGPUTracker's own buying guidance concludes that "for AI training, the key factors are VRAM size, memory bandwidth, and tensor core performance," and that because the two H100 variants carry identical 80 GB capacity, "performance characteristics become the deciding factor" rather than memory size alone (Source: cloudgputracker.com). At the interconnect level the difference is far larger: PCIe 5.0 tops out at roughly 128 GB/s of bus bandwidth, while NVLink on the same generation delivers 900 GB/s, a roughly 7x advantage that becomes decisive for workloads that must synchronize gradients or activations across many GPUs (Source: amcompute.com). Northflank summarizes the tradeoff for buyers succinctly: "PCIe is easier to deploy and shows up in more off-the-shelf systems. SXM offers better performance with higher bandwidth and power, often used in tightly coupled multi-GPU servers" (Source: northflank.com). A third variant, the **H100 NVL**, is a PCIe-form-factor product that pairs two H100 boards with a dedicated NVLink bridge and boosted 94 GB HBM3 memory per GPU, intended specifically for large language model inference workloads such as Llama 2 70B that benefit from extra memory headroom without requiring a full eight-GPU HGX system (Source: spheron.network). A single H100 NVL GPU holds 94 GB of HBM3, and Llama 3 70B at FP16 precision requires roughly 140 GB for weights alone, so a paired NVL module provides 188 GB total, enough to hold the model plus meaningful key-value cache headroom without the four-way tensor parallelism an 80 GB SXM GPU would otherwise require (Source: spheron.network).

Cloud Rental Economics

For teams that rent rather than buy, the cost differential compresses considerably because cloud providers price capacity dynamically. On-demand H100 SXM rates across major clouds range from roughly $1.50 per hour on lower-cost neoclouds up to $5.12 per hour on AWS, with GCP at $4.85 and Azure at $4.98 per hour (Source: inferencebench.io) (Source: runcrate.ai). H100 PCIe rentals run cheaper still, with providers such as Lambda listing $2.29 per hour on-demand versus $2.49 for the equivalent SXM instance, and Runcrate quoting H100 PCIe at $1.40 per hour against $1.50 for SXM, roughly a 7% discount for the PCIe variant on that platform (Source: inferencebench.io) (Source: runcrate.ai). Jarvislabs, a smaller provider, advertises H100 access starting at $2.69 per hour with per-minute billing, while noting that rates across the broader market span from $2.69 up to $9.984 per hour depending on the provider (Source: jarvislabs.ai) (Source: jarvislabs.ai). Newer Blackwell-generation rentals command a premium

over Hopper, with Runcrate listing the B200 at roughly $3.40 per hour, notably above its own Hopper-generation H100 and H200 rates on the same platform (Source: runcrate.ai). Because pricing varies so widely by provider, commitment length, and region, procurement teams should treat any single quoted rate as a snapshot rather than a market constant, and should validate current pricing directly with providers before budgeting.

When Each Form Factor Makes Sense

Server-parts.eu, a hardware reseller specializing in GPU procurement, summarizes the decision along workload lines: PCIe suits **inference, fine-tuning, and mixed workloads** where GPUs operate largely independently, while SXM (via HGX) suits **large model training and high-performance computing (HPC)** where tight GPU-to-GPU synchronization dominates runtime (Source: server-parts.eu). MBUZZ Technologies, a systems integrator, adds that SXM GPUs "are ideal for premium AI and HPC services" using NVLink, NVSwitch, and InfiniBand together in clusters such as NVIDIA SuperPODs, while "PCIe GPUs are better suited for cost-effective, general-purpose instances, particularly for tasks like inference, rendering, or workloads where GPUs function independently" (Source: mbuzztech.com). DeployBase frames the buy-versus-rent decision similarly for sustained workloads, recommending teams "own PCIe for stable production training" once utilization exceeds roughly 1,000 hours per month, at which point the economics favor capital purchase over continued hourly rental (Source: deploybase.ai).

This maps to a broader principle applicable across every form factor NVIDIA sells: workloads that fit comfortably on a single GPU, or where GPUs process independent batches without frequent synchronization, rarely benefit enough from NVLink to justify SXM's higher power draw, higher unit cost, and more restrictive server compatibility. Workloads that require distributed tensor parallelism, large mixture-of-experts (MoE) models, or multi-node training runs generally cannot achieve competitive throughput on PCIe alone, because the collective communication overhead across the slower interconnect dominates step time; independent interconnect research describes this as "the cross-rack collapse," where "the moment a collective walks off the fast fabric onto PCIe or Ethernet, per-GPU bandwidth drops 10 to 20 times" (Source: blog.prompt20.com). NVIDIA's Blackwell generation reinforces this split structurally: as of mid-2026, the newest Blackwell and Blackwell Ultra GPUs (B200 and B300) are sold almost exclusively in SXM form via HGX and DGX systems, with PCIe-class enterprise inference instead served by the separate RTX PRO Server line built on MGX (Source: mbuzztech.com).

HGX, DGX, and MGX: The Baseboard and System Layer

HGX: The Proprietary Multi-GPU Baseboard

NVIDIA HGX is the baseboard that holds SXM GPUs and wires them together with NVLink and NVSwitch. It is not sold to consumers as a standalone product; it is the reference design that OEMs such as Dell, HPE, Supermicro, and Lenovo integrate into their own branded servers, choosing their own chassis, cooling system, storage, and CPU pairing around NVIDIA's fixed GPU module (Source: civo.com). Civo, a cloud infrastructure provider, describes HGX as bringing "together the full power of NVIDIA GPUs, NVLink, NVIDIA networking, and fully optimized AI and HPC software stacks" to deliver maximum performance, while the OEM "decides on the chassis, cooling, power delivery, and network configuration" (Source: civo.com). Historically, HGX configurations have shipped in four-GPU and eight-GPU variants across every generation from A100 through the newest Rubin platform (Source: amcompute.com). The most current HGX Blackwell specifications show HGX B300 (Blackwell Ultra) delivering 144 petaFLOPS (PFLOPS) of sparse FP4 Tensor Core performance and 2.1 TB of total memory, versus HGX B200's otherwise similar 144 PFLOPS FP4 figure but with 1.4 TB of total memory, with both generations using fifth-generation NVLink at 1.8 TB/s of GPU-to-GPU bandwidth and 14.4 TB/s of total NVLink bandwidth across the eight-GPU baseboard (Source: nvidia.com). For enterprises building custom clusters directly from HGX rather than buying DGX, NVIDIA's own Enterprise Reference Architecture uses HGX B300 as its building block, with each eight-GPU system grouped into scalable units that can reach up to 1,024 GPUs across 128 systems (Source: civo.com).

DGX: NVIDIA's Own Turnkey Server

NVIDIA DGX uses the identical HGX baseboard as OEM partner systems but wraps it in a fixed, NVIDIA-selected configuration of CPUs, memory, networking, storage, and software, sold and supported directly by NVIDIA (Source: amcompute.com). ServeTheHome confirms this lineage directly, noting that "the NVIDIA DGX V100 and DGX A100 generations used the HGX baseboards and then built a server around" them, with NVIDIA rotating OEM manufacturing partners generation to generation while keeping the configuration largely fixed (Source: servethehome.com). The current flagship, **DGX B200**, packs eight Blackwell GPUs interconnected with fifth-generation NVLink, delivering three times the training performance and 15 times the inference performance of the prior DGX H100 generation, with 1,440 GB of total GPU memory at 64 TB/s of aggregate HBM3e bandwidth and a maximum system power draw of approximately 14.3 kilowatts (kW), according to NVIDIA's own benchmarks and specifications (Source: nvidia.com). Retail listings from European reseller aiserver.eu price the DGX B200 at approximately €455,000 to €628,000 (roughly $490,000 to $676,000 at mid-

2026 exchange rates), with the newer DGX B300 listed higher at €535,000 to €714,000, illustrating that each Blackwell-class DGX generation has commanded a meaningful price premium over its predecessor (Source: aiserver.eu) (Source: aiserver.eu), while a Reddit discussion citing early market pricing estimated an eight-GPU DGX B200-class server's component cost alone near $500,000 based on comparable server builds (Source: www.reddit.com). Because DGX bundles NVIDIA's software stack, Base Command orchestration, and enterprise support directly, it typically costs more than an equivalent OEM HGX-based server; teams optimizing for lowest cost per GPU or already standardized on a specific OEM generally choose HGX-based partner servers instead (Source: amcompute.com). Civo documents a concrete enterprise example of this integrated approach: electronics and entertainment conglomerate Sony has deployed "large clusters of DGX A100 systems installed in their data centers" for tasks such as training deep learning models for super-resolution image processing, "cutting what was previously a month-long training workload down to a single day" (Source: civo.com). Civo also documents that NVIDIA has continued extending the DGX family downward in scale, announcing at GTC 2026 the DGX Station GB300, a "deskside supercomputer that packs 748GB of memory and up to 20 petaflops of compute," aimed at bringing data-center-class performance to individual developer desks (Source: civo.com).

MGX: The Open, Modular Reference Architecture

NVIDIA MGX solves a different problem than HGX and DGX. Where HGX is a fixed, high-density eight-GPU baseboard optimized for maximum NVLink interconnect, MGX is an open modular reference architecture that lets partners combine GPUs, CPUs (Grace, Vera, x86, or other Arm processors), DPUs, storage, and networking across more than 100 validated combinations, spanning single-node inference servers up to full rack-scale AI factories (Source: nvidia.com) (Source: amcompute.com). MGX supports both rack-scale and single-card PCIe GPU solutions, including Blackwell and Rubin generations, and both Arm-based CPUs such as Grace and Vera alongside x86 and other Arm processors, giving buyers a genuinely mixed-vendor path that HGX does not offer (Source: nvidia.com). MGX is not a replacement for HGX; rather, HGX targets eight-GPU training nodes needing maximum interconnect bandwidth, while MGX covers inference servers, mixed CPU-GPU nodes, and smaller or more heterogeneous deployments where flexibility matters more than peak density (Source: amcompute.com).

MGX's practical value proposition centers on supply chain and engineering economics. NVIDIA states that MGX reduces research and development costs by $2 million to $4 million per platform for system builders through shared reference designs, and enables data center operators to scale from eight-GPU nodes to 144-GPU racks using consistent power and cooling interfaces while achieving up to 50% lower total cost of ownership (TCO), attributed to 94% power supply efficiency and reusable liquid-cooling plumbing (Source: developer.nvidia.com). MGX also enables roughly 80% of rack components, including busbars, coldplates, and power whips, to be pre-integrated at the factory, cutting deployment timelines from roughly 12 months down to under 90 days according to NVIDIA, and addresses Blackwell's power density directly: a full GB200 NVL72 rack requires up to 120 kW, which MGX's liquid-cooled busbars and manifolds handle while holding coolant temperature differentials under 15 degrees Celsius even at 1,400 amp loads (Source: developer.nvidia.com). More than 200 ecosystem partners have adopted MGX components as of mid-2025, including Asus, Cisco, Gigabyte, HPE, Lenovo, Supermicro, and Wistron (Source: developer.nvidia.com).

MGX's reach extends beyond conventional GPU servers. NVIDIA's newest inference accelerator rack, the **Groq 3 LPX** (a rack of 256 language processing unit, or LPU, accelerators produced with Groq), is built on the **MGX ELT rack** design, illustrating that MGX now underpins non-GPU accelerator hardware within the same standardized rack, power, and cooling envelope as GPU-based systems. Each LPX rack delivers 128 GB of static random-access memory (SRAM) for low-latency processing and 12 TB of DDR5 memory for large models, with 40 petabytes per second (PB/s) of SRAM bandwidth and 640 TB/s of chip-to-chip scale-up bandwidth across the rack, all built on the same MGX mechanical and power framework used for Blackwell and Rubin GPU racks (Source: nvidia.com). The distinction to keep straight: **HGX is a component** (a GPU baseboard), **DGX is a finished product** (NVIDIA's own server), and **MGX is a design methodology** (an open framework other companies' finished products are built from).

Comparing the Three Platforms Directly

Table 2 summarizes how HGX, DGX, and MGX differ across the dimensions that matter to buyers.

DIMENSION	HGX	DGX	MGX
What it is	A GPU baseboard with SXM sockets, NVLink, and NVSwitch	A complete, NVIDIA-built and NVIDIA-supported server	An open modular reference design, not a fixed product
Who builds the final system	OEMs (Dell, HPE, Supermicro, Lenovo, and others)	NVIDIA	OEMs and ODMs (Asus, Cisco, Gigabyte, Quanta, Wistron, and 200+ partners) (Source: developer.nvidia.com)
GPU form factor	SXM only, 4 or 8 GPUs per baseboard	SXM (same HGX baseboard)	SXM and PCIe, single-node to rack-scale
Primary use case	Maximum-density 8-GPU training nodes	Ready-to-deploy enterprise AI with direct support (Source: civo.com)	Flexible, mixed-vendor, mixed-scale deployments including inference
Support model	Handled by the OEM (Source: civo.com)	Direct NVIDIA enterprise support	Handled by the OEM/ODM
Approximate 8-GPU price point (Blackwell generation)	Varies by OEM, typically below equivalent DGX	~€455,000 to €628,000 (DGX B200) (Source: aiserver.eu)	Varies widely by OEM and configuration

The pattern across the table is that HGX and DGX both optimize for the same physical GPU topology (eight-GPU SXM with full NVLink mesh) but differ in who assembles and supports the surrounding system, while MGX is architecturally distinct, prioritizing configurability over fixed maximum density. Servermall, a server hardware supplier, frames the buying decision plainly: "if you need an affordable and flexible server for inference, testing, RAG or several applied models, PCIe is usually enough. If the workload requires 4 to 8 GPUs with fast communication between them, it is worth looking at SXM/HGX. DGX makes sense when the company needs not only graphics cards, but also a ready-made hardware and software system with support" (Source: servermall.com). A buyer choosing between an HGX-based Dell PowerEdge XE9680 and an NVIDIA DGX B200 is choosing a support and integration model, not a different GPU; a buyer choosing MGX-based hardware is choosing a different design philosophy entirely, one built for heterogeneous and evolving fleets rather than a single fixed configuration.

Superchips: Grace Hopper and Grace Blackwell Explained

What Makes a Superchip Different From a Standard GPU

The term "superchip" describes a specific NVIDIA product category, not a marketing synonym for "powerful GPU." A superchip is a module that physically and electrically fuses an NVIDIA Grace CPU with one or more GPUs using **NVLink-C2C (chip-to-chip)**, a coherent, high-bandwidth, low-latency interconnect that lets the CPU and GPU share memory as if they were a single processor. The **GH200 Grace Hopper Superchip**, the first of this family, combines the Grace CPU (up to 72 Arm Neoverse V2 cores) with a Hopper GPU (up to 144 streaming multiprocessors and 96 GB of HBM3 memory) via NVLink-C2C, delivering up to 900 GB/s of coherent bandwidth, seven times faster than a PCIe Gen5 link between a separate CPU and GPU (Source: developer.nvidia.com). By contrast, a "standard" H100 or B200 GPU, whether PCIe or SXM, is paired with a conventional x86 or separate Arm CPU over ordinary PCIe, with no shared coherent memory space; the CPU and GPU remain two distinct memory domains that require explicit data copies. Each GH200 module also ships with up to 480 GB of LPDDR5X CPU memory, giving the GPU direct access to seven times more fast memory than its own 80 GB of HBM3, or nearly eight times more with 141 GB HBM3e variants (Source: xenon.com.au).

This coherence matters operationally because it removes the need to copy data back and forth between CPU and GPU memory for many workloads. NVIDIA reports that GH200's shared per-process page table lets CPU and GPU threads access all system-allocated memory regardless of physical location, accelerating data processing workloads such as Apache Spark with RAPIDS by up to 36 times versus a 16-node premium CPU cluster, and reports embedding generation speedups of up to 30 times for retrieval-augmented generation (RAG) pipelines, driven by the Grace CPU's 72 power-efficient cores preprocessing data and NVLink-C2C moving it to the Hopper GPU seven times faster than PCIe (Source: nvidia.com). A dual-superchip

configuration, **GH200 NVL2**, fully connects two GH200 modules over NVLink for 288 GB of combined high-bandwidth memory and 10 TB/s of memory bandwidth, up to 3.5 times more GPU memory capacity than a single H100 server (Source: nvidia.com). Independent academic research on Grace Hopper's data movement, published on arXiv by researchers at ETH Zurich and the Swiss National Supercomputing Centre, measured point-to-point NVLink-C2C bandwidth at approximately 450 GB/s per direction (900 GB/s bidirectional total), consistent with NVIDIA's own published figures, and documented the architecture of the "Alps" supercomputer in Switzerland, whose compute nodes each pack four GH200 superchips fully interconnected by NVLink alongside a Slingshot network interface for inter-node scaling, demonstrating a second independent production deployment of the quad-GH200 design pattern beyond JUPITER (Source: arxiv.org).

The Grace Blackwell Superchip and Its Role in NVL72

The successor design, the **GB200 Grace Blackwell Superchip**, is the building block of the GB200 NVL72 rack, connecting one Grace CPU to two Blackwell GPUs, again via NVLink-C2C (Source: nvidia.com). Compared to the earlier GH200's one-to-one CPU-to-GPU ratio, GB200's Grace CPU pairs with two GPUs, a two-to-one GPU-to-CPU ratio that industry analysts note reflects direct customer feedback: many workloads found the GH200's one Grace CPU per GPU ratio provided more CPU capability than needed, driving up cost relative to benefit, which is one reason GH200 shipped in comparatively low volumes next to the eight-GPU, two-CPU HGX H100 configuration (Source: newsletter.semianalysis.com). SemiAnalysis describes the core GB200 module, known internally as the "Bianca board," as containing "two Blackwell B200 GPUs and a single Grace CPU," with a CPU-to-GPU ratio of 1:2 on a board compared to GH200's 1:1 ratio (Source: newsletter.semianalysis.com). Not every deployment uses the standard Grace-paired board: SemiAnalysis reports that hyperscaler Meta primarily uses a custom "Ariel" board variant, which swaps the standard two-GPU Bianca configuration for one Grace CPU paired with a single Blackwell GPU, doubling Grace CPU content per GPU to support Meta's memory-intensive recommendation-system training and inference workloads that require larger embedding tables (Source: newsletter.semianalysis.com). A separate variant, code-named "Miranda," replaces the Grace CPU entirely with x86 processors in an otherwise NVL72-compatible rack, trading the 900 GB/s NVLink-C2C bidirectional bandwidth of the Grace pairing for conventional PCIe-based CPU-to-GPU connectivity, which SemiAnalysis notes results in materially lower CPU-to-GPU bandwidth and a TCO profile that remains "questionable" versus the Grace-based configuration (Source: newsletter.semianalysis.com). This underscores that "superchip" specifically denotes the Grace-CPU-plus-GPU coherent package; an x86 CPU paired with the same Blackwell GPU in the same rack is not a superchip, even though it may sit in an otherwise identical NVL72 chassis.

Superchip vs Standard GPU: Practical Decision Criteria

The choice between a superchip-based system and a standard PCIe or SXM GPU server hinges on whether the workload is genuinely CPU-GPU coupled or purely GPU-bound:

- **Choose a superchip (GH200/GB200) when:** the workload requires frequent, fine-grained data movement between CPU and GPU memory, such as graph neural networks, large-scale data analytics, RAG pipelines with heavy preprocessing, or scientific simulations that oversubscribe GPU memory using CPU memory as an extension. GH200 delivers up to 8 times faster graph neural network training than an H100 PCIe system paired with a conventional CPU, according to NVIDIA benchmarks based on the GraphSAGE model (Source: nvidia.com).
- **Choose a standard GPU (PCIe or SXM/HGX) when:** the workload is dominated by GPU-side compute with predictable, batchable data transfer, such as most LLM training and inference, computer vision, and rendering, where the CPU's role is comparatively light and a conventional x86 host CPU is sufficient.
- **Choose GB200 NVL72 (superchip at rack scale) when:** the model itself exceeds what a single node's GPU memory or NVLink domain can hold, requiring the 72-GPU non-blocking fabric to keep tensor-parallel and expert-parallel communication off the slower scale-out network.

NVL72 and Rack-Scale Architecture

From Eight GPUs to Seventy-Two

The **NVL** designation (NVLink domain size) describes how many GPUs share a single non-blocking, all-to-all NVLink fabric, independent of how many physical servers those GPUs occupy. Through the Hopper generation, that domain was capped at eight GPUs, the maximum an HGX baseboard's NVSwitch chips could connect at full bandwidth, and independent interconnect researchers describe the resulting 2024 to 2026 shift to 72-GPU domains as "the most consequential interconnect change since NVLink itself shipped on Pascal" (Source: blog.prompt20.com). The **GB200 NVL72** breaks that ceiling by moving the NVSwitch fabric out of the server and into dedicated switch trays that span an entire rack, extending the non-blocking NVLink domain from 8 to 72 GPUs. An NVL72 rack contains 18 compute trays, each with two Grace CPUs and four Blackwell GPUs (two GB200 superchips), plus nine NVLink switch trays and two top-of-rack switches for management, all connected through a passive copper cable

backplane, with each switch tray housing two NVLink NVSwitches that together deliver 57.6 terabits per second (Tbps) of full-duplex bandwidth in a single rack unit (1U) design (Source: docs.nvidia.com) (Source: docs.nvidia.com). SemiAnalysis's independent component-level teardown corroborates this layout precisely, describing "18 1U compute trays and 9 NVSwitch trays" per rack, with each NVSwitch tray housing "two 28.8Tb/s NVSwitch5 ASICs" (Source: newsletter.semianalysis.com).

The result, as NVIDIA describes it, is that the 72-GPU domain "acts as a single, massive GPU," delivering 130 TB/s of low-latency GPU communication bandwidth across the rack (Source: nvidia.com). Independent research describes the same phenomenon in engineering terms: "an NVL72 rack is a 72-way crossbar inside one machine," where "NVLink 5 plus NVSwitch 4 turns the rack into a single SMP-like fabric where any GPU can reach any other at approximately 1.8 TB/s aggregate" (Source: blog.prompt20.com). NVIDIA's benchmark claims, measured against an HGX H100 cluster scaled over InfiniBand, put GB200 NVL72 at up to 30 times faster real-time LLM inference, 4 times faster LLM training, and 25 times better energy efficiency, alongside 18 times faster database query performance versus CPU-only systems. Because these are NVIDIA-published, workload-specific comparisons rather than independent third-party benchmarks, buyers should treat the exact multipliers as vendor-reported ceilings under favorable configurations rather than guaranteed uplift for every workload; SemiAnalysis, an independent semiconductor research firm, notes that only one hyperscaler currently plans NVL72 as its primary deployment configuration, with most others favoring the lower-density NVL36x2 variant (two 36-GPU racks joined into one 72-GPU domain) because most data centers cannot yet support NVL72's roughly 120 kilowatt (kW) per-rack power density even with direct-to-chip liquid cooling (Source: newsletter.semianalysis.com) (Source: blog.prompt20.com). NVL36x2, by contrast, draws approximately 66 kW per rack for a combined 132 kW across the paired-rack configuration, roughly 10 kW more than a single NVL72 rack in total, but at half the per-rack density most existing facilities can actually support (Source: newsletter.semianalysis.com).

GB200 NVL72 vs HGX: What Actually Changes

The most consequential difference between an NVL72 rack and a traditional HGX-based deployment is where the NVLink domain boundary sits. An HGX H100 or HGX B200 server tops out at eight GPUs in one non-blocking domain; scaling beyond eight GPUs on HGX-class hardware means falling back to InfiniBand or Ethernet networking between servers, which runs at a fraction of NVLink's bandwidth and materially higher latency. NVL72 moves that boundary from 8 to 72 GPUs, meaning workloads that previously had to be partitioned across the slower scale-out network within a single training job can now stay inside the fast NVLink fabric for nine times as many GPUs. Independent research quantifies the practical consequence for model training: on an 8-GPU HGX node, practical tensor-parallel group size is capped at 8 with total high-bandwidth memory (HBM) in domain around 1.1 TB (H100-class), whereas on an NVL72 rack, practical tensor-parallel groups scale up to 72 with roughly 13.5 TB of HBM in one coherent domain (Source: blog.prompt20.com). *Table 3* contrasts the two approaches directly.

ATTRIBUTE	TRADITIONAL HGX (8-GPU NODE)	GB200 NVL72 (RACK-SCALE)
GPUs per NVLink domain	8	72
CPUs in the domain	Typically 2 x86 CPUs, separate from GPUs	36 Grace CPUs, coherently linked to GPUs via NVLink-C2C
NVLink bandwidth (aggregate)	7.2 TB/s (NVLink 4 Switch, 8 GPUs) (Source: nvidia.com)	130 TB/s (NVLink 5 Switch, NVL72)
Cooling	Typically air-cooled	Fully liquid-cooled (mandatory) (Source: blog.prompt20.com)
Power per rack	Roughly 40 kW for air-cooled H100 racks (Source: newsletter.semianalysis.com)	Approximately 120 kW to 140 kW peak per independent teardowns (Source: newsletter.semianalysis.com)
Cross-node scaling beyond the domain	InfiniBand or Ethernet between 8-GPU nodes	Quantum-X800 InfiniBand or Spectrum-X800 Ethernet between NVL72 racks

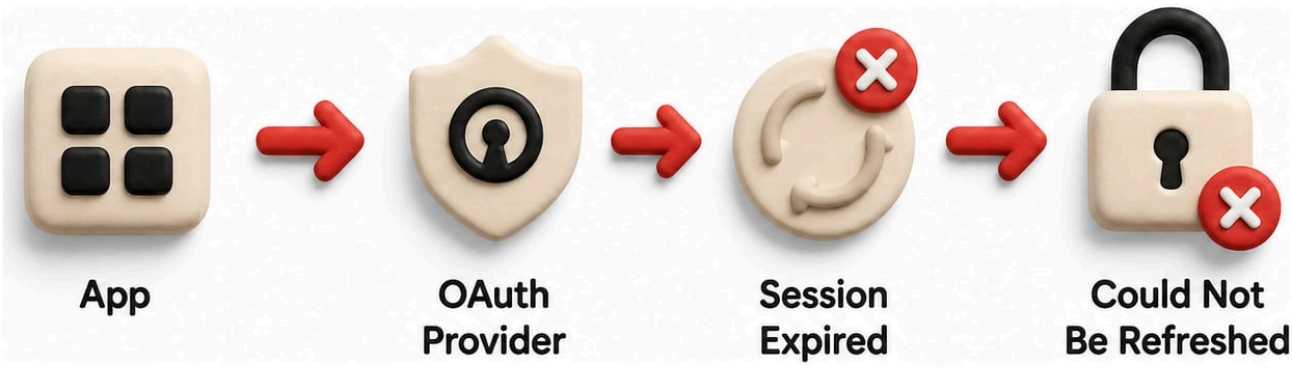
Deployment complexity rises accordingly: SemiAnalysis characterizes the shift as introducing "dozens of different deployment variants with tradeoffs and a significant complexity increase generation on generation" even though NVIDIA markets NVL72 as a standardized rack (Source: newsletter.semianalysis.com). NVIDIA has responded by contributing the GB200 NVL72 rack, compute tray, and switch tray liquid-cooling designs to the Open Compute Project (OCP), and partnering with data center infrastructure firm Vertiv on a joint reference architecture intended to cut

implementation time for data centers deploying Blackwell by up to 50%, while the same OCP contribution communicates a GPU-to-GPU communication speed of 1.8 TB/s per GPU across the 72-GPU domain, consistent with fifth-generation NVLink's per-GPU rate (Source: developer.nvidia.com).

The GB300 NVL72 and Beyond

NVIDIA's follow-on rack, **GB300 NVL72**, replaces the Blackwell GPUs with Blackwell Ultra GPUs and is purpose-built for test-time-scaling inference and AI reasoning workloads, delivering up to a 50 times overall increase in AI factory output performance compared to Hopper-based platforms, according to NVIDIA (Source: nvidia.com). Looking further ahead, NVIDIA's roadmap describes an HGX Rubin NVL8 platform and a Vera Rubin NVL72 rack using sixth-generation NVLink, preliminary specifications the company says will enable 3.6 TB/s of per-GPU bandwidth (2 times the prior generation) and 260 TB/s of aggregate rack bandwidth, roughly double NVL72's current 130 TB/s, and which remain explicitly "subject to change" as directional roadmap guidance rather than shipped specifications as of July 2026 (Source: nvidia.com).

Data Analysis and Evidence



Financial Scale of the Form-Factor Transition

The commercial stakes behind these architectural choices are substantial. NVIDIA's Data Center segment, which encompasses PCIe, SXM, HGX, DGX, MGX, and superchip products, generated record revenue of $51.2 billion in the third quarter of fiscal year 2026, up 66% from a year earlier and up 25% sequentially, a growth rate the company attributes to three simultaneous platform shifts: accelerated computing, more powerful AI models, and the rise of agentic applications (Source: sec.gov). Within that figure, Data Center compute revenue reached $43.0 billion, up 56% year over year, while Networking revenue, which captures the NVLink compute fabric underpinning GB200 and GB300 rack-scale systems specifically, hit a record $8.2 billion, up 162% year over year, driven largely by "the introduction and continued growth of NVLink compute fabric for GB200 and GB300 systems" (Source: sec.gov). NVIDIA also disclosed that Blackwell Ultra had become its leading architecture across all customer categories in the same quarter, while H20 (a China-specific export-compliant chip) sales were "insignificant," reflecting how quickly the installed base rotates through each new form-factor generation (Source: sec.gov). NVIDIA's inventory also grew to $19.8 billion, up from $15.0 billion the prior quarter, with total supply-related commitments of $50.3 billion, which the company attributes directly to "ordering to secure long lead-time components, meet the demand for Blackwell, and support future architecture ramps," a signal that the physical supply chain behind SXM, HGX, and NVL72 hardware remains a binding constraint on how quickly new form factors can reach customers (Source: sec.gov).

Interconnect Bandwidth Across Generations

The generational trend in NVLink bandwidth per GPU illustrates how quickly the interconnect side of the form-factor decision has moved. Per-GPU NVLink bandwidth rose from 900 GB/s on fourth-generation NVLink (Hopper, H100/H200) to 1,800 GB/s on fifth-generation NVLink (Blackwell, B200/GB200), and NVIDIA's roadmap targets 3,600 GB/s on sixth-generation NVLink (Rubin platform), a fourfold increase in three generations (Source: nvidia.com). At the switch level, aggregate rack bandwidth has grown even faster because the domain size itself expanded from 8 to 72 GPUs: NVLink 4 Switch tops out at 7.2 TB/s across an 8-GPU domain, NVLink 5 Switch reaches 130 TB/s across a 72-GPU NVL72 domain, and the projected NVLink 6 Switch targets 260 TB/s across the same 72-GPU footprint. By comparison, PCIe bus bandwidth has grown far more slowly: PCIe

5.0 provides roughly 128 GB/s and PCIe 4.0 roughly 64 GB/s, meaning the gap between NVLink and PCIe, already a 7x difference in the Hopper generation, is set to widen further as NVLink continues its faster generational cadence (Source: deploybase.ai). DeployBase's own latency measurements show the same widening pattern: roughly 1 microsecond of latency on SXM/NVLink versus roughly 3 microseconds on PCIe 5.0 and roughly 5 microseconds on PCIe 4.0, a 3x to 5x latency advantage for SXM that compounds with its bandwidth lead in collective operations (Source: deploybase.ai).

The Cost of Interconnect at Rack Scale

Rack-scale NVLink is not free, and its cost structure is itself a data point worth documenting. SemiAnalysis's component-level teardown estimates a GB200 NVL72 rack uses 5,184 discrete copper NVLink cables (one differential pair per cable, with each of the 72 GPUs requiring 72 differential pairs for full bidirectional 900 GB/s connectivity), a deliberate design choice over optical transceivers, which the firm estimates NVIDIA calculated would have added roughly 20 kW of power draw per rack and introduced materially worse reliability at the 1.6 terabit transceiver speeds required (Source: newsletter.semianalysis.com) (Source: newsletter.semianalysis.com). At roughly $850 per 1.6 terabit transceiver, SemiAnalysis calculates that an all-optical alternative would have cost approximately $550,800 per rack in transceivers alone before markup, or roughly $2.2 million per rack once NVIDIA's typical 75% gross margin is applied, a cost structure the firm identifies as the primary reason a proposed 256-GPU DGX H100 NVL configuration never reached commercial shipment (Source: newsletter.semianalysis.com). Independent cluster-economics analysis further estimates that at 16,000-GPU scale, an NVL72-based deployment carries higher per-rack capital expenditure than an equivalent HGX-based cluster but completes the same large mixture-of-experts training step roughly 15% to 25% faster, owing to expert-parallel bandwidth and the elimination of pipeline bubbles, concluding that "for trillion-parameter MoE training, NVL72 wins decisively," while dense models under roughly 200 billion parameters remain cost-competitive on conventional eight-GPU HGX nodes (Source: blog.prompt20.com).

Adoption Signal: What the H100 Generation Tells Us

Even as Blackwell-generation hardware becomes NVIDIA's leading architecture, the H100, released in 2022, remains heavily used: Stanford's 2026 AI Index Report found the H100 was used to train 28 notable AI models in 2025 alone, more than any other single accelerator, underscoring how long a given form-factor generation remains commercially relevant after a successor launches (Source: thundercompute.com). This matters directly for form-factor purchasing decisions: a buyer choosing PCIe versus SXM, or HGX versus MGX, today is making a decision whose useful life will likely span multiple newer NVIDIA generations, not just the current one. Reinforcing this point, Jarvislabs' pricing analysis notes H100 hourly rates today already span a roughly fourfold range from $2.69 to nearly $10.00 depending on provider, meaning form-factor selection interacts directly with which rental market a buyer can access, not just which hardware they choose (Source: jarvislabs.ai).

Case Studies and Real-World Examples

CoreWeave: First Cloud Provider to Offer GB200 NVL72 at General Availability

CoreWeave, a specialized AI cloud infrastructure provider, announced on February 4, 2025 that it was the first cloud provider to make GB200 NVL72-based instances generally available, built on the GB200 Grace Blackwell Superchip (Source: investors.coreweave.com). CoreWeave's deployment paired rack-level NVLink connectivity with NVIDIA Quantum-2 InfiniBand networking delivering 400 Gb/s of bandwidth per GPU through a rail-optimized topology, scaling to clusters of up to 110,000 GPUs (Source: investors.coreweave.com). CoreWeave co-founder and Chief Strategy Officer Brian Venturo described the launch as "another achievement of our series of firsts" following the company's earlier distinction as the first cloud provider to deploy NVIDIA H200 GPUs in August 2024 (Source: investors.coreweave.com) (Source: investors.coreweave.com). CoreWeave cited the same NVIDIA-published performance figures used throughout this report (up to 30x faster inference, up to 4x faster training, up to 25x lower TCO for real-time inference), and separately announced in early 2025 that it would deliver one of the first GB200 Superchip-enabled AI supercomputers to **IBM** for training the company's Granite model family (Source: investors.coreweave.com). IBM's Priya Nagpurkar, Vice President of Hybrid Cloud and AI Platform Research, said the partnership with CoreWeave, including IBM Spectrum Scale Storage, "demonstrates our commitment to advancing a hybrid cloud strategy for AI" (Source: investors.coreweave.com). NVIDIA's Vice President of Hyperscale and HPC, Ian Buck, framed the collaboration as enabling "organizations of all sizes to push the boundaries of AI" through the combination of CoreWeave's Kubernetes-based orchestration tooling and NVIDIA's rack-scale hardware (Source: investors.coreweave.com).

Microsoft Azure: ND GB200 v6 Reaches General Availability

Microsoft Azure announced the general availability of its **ND GB200 v6** virtual machine (VM) series on March 18, 2025, built on GB200 NVL72 and among the first cloud offerings of a 4,000-GPU GB200-powered supercomputing cluster (Source: techcommunity.microsoft.com). Microsoft reported that its own performance validation, using the Llama 70B model on GB200 NVL72, achieved over 860,000 tokens per second of throughput, a 9x increase per rack compared to the prior-generation ND H100 v5 VM series (Source: techcommunity.microsoft.com). Each ND GB200 v6 rack delivers 400 Gb/s of dedicated InfiniBand bandwidth per GPU, 1.6 terabits per second (Tb/s) per VM, and 28.8 Tb/s per GB200 NVL72 rack in a non-blocking fat-tree network designed to scale to hundreds of thousands of GPUs (Source: techcommunity.microsoft.com). Microsoft's product manager Matt Vegas also described the underlying Blackwell architecture as delivering a "36% increase in High Bandwidth Memory (HBM) with 192GB and a 67% increase in HBM capacity with 8 TB/s per GPU" relative to the prior-generation Azure ND H200 v5 VMs (Source: techcommunity.microsoft.com). Image and video AI company **Black Forest Labs** was cited as an early customer expanding its Azure partnership specifically to leverage this infrastructure for generative media model development, with founder and CEO Robin Rombach stating the collaboration would help the company "build and deliver the best possible image and video models faster and at greater scale" (Source: techcommunity.microsoft.com).

Jülich Supercomputing Centre: JUPITER, Europe's First Exascale-Class System

The **JUPITER** supercomputer, hosted by the Jülich Supercomputing Centre in Germany and owned by the EuroHPC Joint Undertaking, is built on nearly 24,000 NVIDIA GH200 Grace Hopper Superchips interconnected with NVIDIA Quantum-2 InfiniBand, using Eviden's BullSequana XH3000 liquid-cooled architecture (Source: nvidianews.nvidia.com). Announced in June 2025 as Europe's fastest supercomputer, JUPITER is designed to run 1 quintillion FP64 (double-precision floating point) operations per second, putting it on track to become Europe's first exascale system, and ranks among the top five systems on the TOP500 list of the world's fastest supercomputers while being the most energy efficient of that group at 60 gigaflops per watt (Source: nvidianews.nvidia.com) (Source: nvidianews.nvidia.com). Independent tracking by Wikipedia's TOP500-sourced infobox places JUPITER 4th on the November 2025 TOP500 ranking, operational since June 2025, with a measured 1.000 exaFLOPS Rmax (sustained) and 1.226 exaFLOPS Rpeak (theoretical) performance, a total system cost of approximately €499 million, and a power draw of 18.2 megawatts (MW) (Source: en.wikipedia.org (Source: en.wikipedia.org (Source: en.wikipedia.org (Source: en.wikipedia.org. At full capacity, the system is designed to complete AI model training tasks that would otherwise take far longer in under a week, and its waste heat is captured through warm-water cooling and fed into the Jülich campus heating network rather than simply exhausted (Source: en.wikipedia.org (Source: en.wikipedia.org. The system supports climate and weather modeling through NVIDIA's Earth-2 platform, quantum algorithm research using CUDA-Q, computer-aided engineering through PhysicsNeMo and Omniverse, and pharmaceutical drug discovery through NVIDIA's BioNeMo platform (Source: nvidianews.nvidia.com). JUPITER illustrates the superchip form factor's value proposition at true supercomputing scale: choosing a coherent CPU-GPU superchip design over conventional discrete CPU and GPU nodes for a system built explicitly around mixed HPC and AI workloads that benefit from tight memory coupling.

Oracle Cloud Infrastructure: GB200 NVL72 at Rack Scale

Oracle Cloud Infrastructure (OCI) has also deployed GB200 NVL72 racks, joining CoreWeave, Microsoft, and IBM as an early hyperscale adopter of the platform, with NVIDIA's own materials describing Microsoft, CoreWeave, and Oracle Cloud Infrastructure as jointly deploying GB300 NVL72 systems "at scale for low-latency and long-context use cases, such as agentic coding and coding assistants." This cross-hyperscaler pattern, the same rack-scale architecture appearing simultaneously across CoreWeave, Azure, and OCI within roughly a year of GB200's initial launch, is itself evidence that NVL72 has moved from an experimental configuration to a standard deployment target for frontier-scale AI infrastructure among major cloud providers, aligning with the industry shift toward serving frameworks such as NVIDIA Dynamo, TensorRT-LLM, and the open source SGLang and vLLM projects for low-precision NVFP4 inference at scale.

Implications and Future Directions

The trajectory across every layer of NVIDIA's form-factor stack points toward larger coherent domains and tighter CPU-GPU integration, not incremental single-chip improvements. NVLink bandwidth per GPU has already quadrupled from 900 GB/s to a targeted 3,600 GB/s across three generations, and the NVLink domain size has grown ninefold from 8 to 72 GPUs in a single generation transition (Hopper to Blackwell). This has two direct implications for buyers. First, the practical gap between PCIe and SXM/HGX is widening, not narrowing, because NVLink's bandwidth grows faster generation over generation than PCIe's; procurement decisions that were marginal calls in the H100 generation may become clearer-cut in the Blackwell and Rubin generations. Second, the rise of MGX as a parallel, non-HGX modular architecture, now underpinning both GPU servers and non-GPU accelerator racks like Groq's LPX, suggests NVIDIA is deliberately decoupling "maximum interconnect density" (HGX/NVL72) from "flexible heterogeneous deployment" (MGX) as two permanently distinct product tracks rather than converging them, giving enterprise buyers a genuine choice rather than a single forced path.

Power and cooling economics are becoming the binding constraint on adoption speed for the highest-density options. NVL72's roughly 120 kW per-rack power draw, three times a typical air-cooled H100 rack's roughly 40 kW, requires liquid cooling infrastructure that many existing data centers cannot yet support, which is why SemiAnalysis expects most GB200 deployments through 2026 to favor the lower-density NVL36x2 configuration over full NVL72 (Source: newsletter.semianalysis.com). NVIDIA's own roadmap acknowledges this constraint directly, with the company describing a shift toward 800 volt direct current (VDC) power distribution and integrated energy storage beginning in 2027 specifically to support megawatt-scale racks beyond what today's facilities can deliver (Source: nvidia.com). Buyers evaluating NVL72-class hardware today should treat facility power and cooling readiness, not GPU availability, as the likely gating factor on deployment timelines.

Looking further out, NVIDIA's disclosed roadmap for the Rubin platform, including HGX Rubin NVL8 and Vera Rubin NVL72, alongside the Groq 3 LPX inference accelerator built on the MGX ELT rack, signals that the form-factor decision itself is becoming workload-specific rather than one-size-fits-all: LPX targets ultra-low-latency, high-throughput token generation for agentic AI systems specifically, projected to deliver up to 35 times higher throughput per megawatt for trillion-parameter models when paired with Vera Rubin NVL72, a use case distinct from either training-optimized NVL72 GPU racks or general-purpose HGX inference nodes (Source: nvidia.com). Enterprise infrastructure planning through 2027 should therefore expect an expanding rather than consolidating menu of NVIDIA form factors, with the practical challenge shifting from "which GPU" to "which combination of GPU, CPU, interconnect, and rack architecture best matches this specific workload's data movement pattern," a shift Servermall summarizes for buyers navigating the current menu as choosing "PCIe" for "affordable and flexible" single-node needs, "SXM/HGX" once "the workload requires 4 to 8 GPUs with fast communication," and DGX-class turnkey systems once support and predictability outweigh raw cost per GPU (Source: servermall.com).

Frequently Asked Questions (FAQs)

What is the main difference between NVIDIA SXM and PCIe GPUs? SXM is NVIDIA's proprietary socketed form factor that mounts to an HGX baseboard and connects up to eight GPUs via NVLink at up to 900 GB/s per GPU (fourth-generation NVLink, Hopper), while PCIe is a standard expansion card compatible with any server, limited to roughly 128 GB/s over the PCIe 5.0 bus with no native multi-GPU fabric beyond an optional two-GPU bridge (Source: amcompute.com) (Source: server-parts.eu).

What is NVIDIA NVL72 in simple terms? NVL72 is the name for a 72-GPU NVLink domain, most commonly seen in the GB200 NVL72 and GB300 NVL72 racks, where 72 Blackwell or Blackwell Ultra GPUs and 36 Grace CPUs are connected through rack-spanning NVLink switch trays so the entire rack functions as one very large GPU rather than 72 separate ones (Source: blog.prompt20.com).

What is the NVIDIA Grace Hopper Superchip? GH200 is a single module that fuses an Arm-based Grace CPU (up to 72 cores) with a Hopper GPU (up to 96 GB of HBM3) using the NVLink-C2C interconnect, giving the CPU and GPU a shared, coherent 900 GB/s memory path instead of a conventional PCIe connection between two separate chips (Source: developer.nvidia.com) (Source: arxiv.org).

What is NVIDIA MGX architecture used for? MGX is an open modular reference architecture that lets server manufacturers combine different GPUs, CPUs, DPUs, storage, and networking into more than 100 validated configurations spanning single-node inference servers to rack-scale AI factories, distinct from HGX's fixed, maximum-density eight-GPU baseboard (Source: amcompute.com).

How does GB200 NVL72 compare to HGX systems? An HGX system caps its non-blocking NVLink domain at eight GPUs per node, requiring slower InfiniBand or Ethernet networking to scale further; GB200 NVL72 extends the non-blocking NVLink domain to 72 GPUs within a single rack, delivering 130 TB/s of aggregate bandwidth versus 7.2 TB/s for an 8-GPU HGX H100 node, at the cost of requiring liquid cooling and roughly 120 kW to 140 kW of rack power (Source: blog.prompt20.com).

Is there a real performance difference between PCIe and SXM GPUs, or is it just marketing? The difference is measurable and substantial: independent specification comparisons show the H100 SXM delivering 68% higher memory bandwidth and roughly double the TDP-driven sustained throughput of the H100 PCIe variant in multi-GPU workloads, driven by NVLink's roughly 7x bandwidth advantage over the PCIe bus, not by any difference in the underlying silicon (Source: cloudgputracker.com) (Source: amcompute.com).

What is a "superchip" versus a standard GPU? A superchip physically combines a Grace CPU and one or more GPUs on one module with coherent, shared memory access via NVLink-C2C at up to 900 GB/s; a standard GPU (PCIe or SXM) pairs with a separate host CPU over conventional, non-coherent PCIe, requiring explicit memory copies between the two (Source: developer.nvidia.com).

Which form factor should a small enterprise AI team start with? For teams running single-GPU or lightly parallel inference and fine-tuning workloads without frequent multi-GPU synchronization, PCIe-based systems or MGX-based single-node servers typically offer lower cost and broader server compatibility; teams that need to train large models across many GPUs simultaneously should evaluate HGX-based or DGX systems, and only

move to NVL72-class rack-scale hardware once model or context size genuinely exceeds what an eight-GPU NVLink domain can serve efficiently (Source: server-parts.eu) (Source: servermall.com).

Why does NVIDIA sell both PCIe and SXM versions of the same GPU instead of just one? Because workload requirements diverge sharply: single-GPU or loosely-coupled tasks gain little from NVLink's 7x bandwidth premium and would simply pay extra for unused interconnect capacity, while tightly-coupled multi-GPU training loses substantial throughput without it, so NVIDIA maintains both a broadly compatible, lower-cost PCIe line and a higher-bandwidth, higher-cost SXM/HGX line to match each workload class (Source: deploybase.ai).

Conclusion

NVIDIA's GPU form factors are best understood as five distinct layers of the same stack rather than five competing product lines. **PCIe and SXM** describe how an individual GPU physically connects to a server, with SXM's HGX baseboard and NVLink fabric delivering roughly 7 times the interconnect bandwidth of PCIe at the cost of higher power draw, higher unit price, and stricter server compatibility requirements. **HGX and DGX** describe who assembles and supports the resulting eight-GPU system, an OEM partner or NVIDIA itself, while sharing identical underlying hardware. **MGX** is an architecturally distinct, open, and more flexible alternative optimized for heterogeneous fleets and rapid deployment rather than maximum single-node density. **Superchips** such as GH200 and GB200 solve a different problem entirely, coherent CPU-GPU memory sharing, rather than GPU-to-GPU scaling. And **NVL72** extends the NVLink fabric itself from a single eight-GPU server out to an entire 72-GPU, 120-kilowatt rack that behaves as one unified accelerator.

The practical decision for any given deployment reduces to matching workload communication patterns to the correct layer: independent, batchable jobs belong on PCIe or MGX-based single-node hardware; tightly coupled multi-GPU training belongs on HGX or DGX; CPU-GPU-coupled workloads with heavy data movement, such as graph analytics, RAG pipelines, or memory-oversubscribed HPC simulations, benefit from a superchip; and only workloads that genuinely exceed an eight-GPU NVLink domain, frontier-scale LLM training and real-time trillion-parameter inference, justify NVL72's power, cooling, and cost overhead. Real deployments at CoreWeave, Microsoft Azure, Oracle Cloud Infrastructure, IBM, and the Jülich Supercomputing Centre's JUPITER system confirm this is no longer a theoretical framework but the operating reality of frontier AI infrastructure as of mid-2026, with NVIDIA's Data Center revenue of $51.2 billion in a single quarter reflecting how directly these architectural choices now drive enterprise technology spending. As NVLink bandwidth continues to outpace PCIe generation over generation and rack-scale designs like NVL72 become the default target for frontier workloads, infrastructure buyers who understand this layered taxonomy, rather than treating "GPU" as a single undifferentiated purchase, will be better positioned to match spend to actual workload requirements.

Tags: nvidia gpu form factors, sxm vs pcie, nvidia nvl72, grace hopper superchip, nvidia mgx, gb200 nvl72, nvidia hgx, nvidia dgx, blackwell architecture, nvlink

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.