Rillor

EXPLAINER

HGX H200, B200 and B300 compared

From HGX H200 to B300, GPU memory per system grows from 1.1 TB to over 2 TB, while dense FP8 holds at the B200 level and native FP64 falls.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- From HGX H200 to HGX B200, GPU memory per eight-GPU system rises from 1,128 GB to 1.4 TB, memory bandwidth per GPU from 4.8 to 7.7 TB/s and dense FP8 per GPU from about 2 to 4.5 PFLOPS. FP4 tensor cores arrive with Blackwell [1, 2, 5, 6, 18].
- HGX B300 adds memory and dense FP4 (108 against 72 PFLOPS per system) and doubles attention throughput by NVIDIA's specification. Memory bandwidth, FP8 and FP16 rates and NVLink bandwidth match HGX B200 [1, 3, 4].
- Native FP64 falls at each step: about 536 TFLOPS on FP64 tensor cores for eight H200s (derived), 296 TFLOPS for HGX B200 and 10 TFLOPS for HGX B300. INT8 falls from 72 to 3 POPS sparse from HGX B200 to HGX B300 [1, 6]. NVIDIA expects HPL to run low on B300 and supports FP64 emulation there [22].
- NVIDIA lists HGX B300 memory as 2.1 TB (270 GB per GPU) [1], its DGX B300 user guide gives 2.3 TB (288 GB per GPU) [14], and its Blackwell Ultra blog notes that HBM capacity "varies by SKU" [5].
- DGX power maximums are 10.2 kW for a standard DGX H200 in 8U, 14.3 kW for DGX B200 and 14.5 to 15.1 kW for DGX B300, both in 10U [13, 14, 16, 17, 18].
- In MLPerf® Inference results from the same submitter and round, eight B200s produced about 2.6 to 3.0 times the tokens per second of eight H200s on a 70-billion-parameter benchmark, with FP4 weights against FP8. Eight B300s produced 1.1 to 1.2 times eight B200s on three language benchmarks (derived) [34, 35].

Background

NVIDIA HGX is a baseboard carrying eight SXM GPUs joined by NVLink switches, which server makers build into complete systems [1, 2]. NVIDIA also sells its own eight-GPU systems under the DGX name [12, 15, 18], and calls HGX and DGX B300 systems "Standardized 8 GPU Blackwell Ultra configurations" [5]. Servers built on the same board differ in the rest of the build: host CPUs, host memory, network cards, storage, power supplies and cooling [14, 19, 20, 21].

NVIDIA announced H200 in November 2023 in four-way and eight-way HGX boards "compatible with both the hardware and software of HGX H100 systems" [8]. The datasheets and technical brief covering HGX B200 and HGX B300 date from September and October 2025 [2, 3, 4]. On 24 October 2026, NVIDIA's HGX page lists HGX B300 and HGX B200 as shipping and no longer shows a table for HGX H200 or H100 [1].

NVIDIA's own documents disagree. B200 memory appears as 180 GB [2] or up to 192 GB [4] per GPU, and B300 system totals as 2.1 TB [1] or 2.3 TB [14]. Throughput tables mix sparse and dense rates [1, 4]. Other figures differ because one document describes a chip and another a board or DGX system [2, 5]. This report ties each figure to its product and tests datasheet gaps against measured results.

Method

What was examined. HGX H200, B200 and B300, with HGX H100 as a reference. For each, per GPU and per eight-GPU system: GPU memory and bandwidth, tensor throughput by precision (dense and sparse), FP64, NVLink, the host link, cluster networking, power, size, weight and operating temperature. For the measured side: MLPerf® Inference results for eight-GPU systems on these boards.

Sources. NVIDIA's HGX page, datasheets, technical brief, blog posts, product pages, announcement and line card [1-11]. NVIDIA's DGX product pages, datasheets and user guides [12-18]. Supermicro and Dell documents for other builds of the same boards [19, 20, 21]. NVIDIA's HPC-Benchmarks notes, cuBLAS documentation and FP64 emulation blog, and a 2026 preprint by Uchino, Ozaki and Imamura [22-25]. MLCommons' benchmark pages, announcements, rules, messaging guidelines and result files for MLPerf® Inference v5.1 and v6.1 [26-35].

How. Each figure is recorded as its source states it, with the product it describes: a chip, an HGX board or a DGX system. Dense rates are half the sparse rates where NVIDIA says so. Figures marked "derived" or "(d)" are this report's arithmetic. For MLPerf®, only Closed division results in the Available category for one node with eight accelerators were used, and only pairs from the same submitter and round were compared directly. Every source was opened on 24 October 2026. The report uses no Rillor data and no prices.

Findings

1. The systems side by side

Eight-GPU system	HGX H100	HGX H200	HGX B200	HGX B300
GPU memory	640 GB [18]	1,128 GB [18]	1.4 TB [1]	2.1 TB [1] or 2.3 TB [14]
Memory bandwidth	26.8 TB/s (d)	38.4 TB/s (d)	62 TB/s [2]	62 TB/s [3]
FP4 tensor, sparse and dense	none [5]	none [5]	144 and 72 PFLOPS [1]	144 and 108 PFLOPS [1]
FP8 tensor, sparse and dense	31.7 and 15.8 PFLOPS (d)	31.7 and 15.8 PFLOPS (d)	72 and 36 PFLOPS [1]	72 and 36 PFLOPS [1]
INT8 tensor, sparse	31.7 POPS (d)	31.7 POPS (d)	72 POPS [1]	3 POPS [1]
FP64, vector and tensor core	272 and 536 TFLOPS (d)	272 and 536 TFLOPS (d)	296 TFLOPS [1]	10 TFLOPS [1]
NVLink per GPU	900 GB/s [7]	900 GB/s [6]	1.8 TB/s [1]	1.8 TB/s [1]
GPU power, configurable maximum	700 W [7]	700 W [6]	1,000 W [2]	1,100 W [3]
Host link per GPU	PCIe Gen5, 128 GB/s [7]	PCIe Gen5, 128 GB/s [6]	PCIe Gen5, 128 GB/s [2]	PCIe Gen6, 256 GB/s [3]

(d): eight times NVIDIA's per-GPU figure, with dense at half the sparse rate NVIDIA lists [6, 7]. NVIDIA marks H200 figures as preliminary [6].

H100 and H200 share their compute rates and 700 W maximum [6, 7]. H200 adds memory, 141 GB against 80 GB at 4.8 against 3.35 TB/s, which NVIDIA calls nearly double the capacity with 1.4 times the bandwidth [6, 7].

From H200 to B200 nearly every row changes. Memory per system rises 1.28 times, bandwidth per GPU about 1.6 times and dense FP8 about 2.3 times, and FP4 appears (derived from [1, 2, 6, 18]). NVLink per GPU doubles [4]. FP64 per GPU is lower on B200 than on H200's FP64 tensor cores, 37 against 67 TFLOPS [2, 6].

From B200 to B300 fewer rows change. Memory per system rises 1.5 times at 270 GB per GPU or 1.6 times at 288 GB, and dense FP4 rises 1.5 times (derived from [1, 3, 14]). Bandwidth, FP8 and NVLink stay level, and INT8 and FP64 fall [1].

2. Memory capacity, bandwidth and why published totals differ

HGX B300: 270 GB or 288 GB per GPU. NVIDIA's HGX page gives HGX B300 2.1 TB [1]. The Blackwell Ultra datasheet lists 270 GB of HBM3E per GPU and 2.1 TB per system and describes "over 2 TB of HBM3E memory" [3], and the DGX B300 product page and February 2026 datasheet also say 2.1 TB [12, 13]. NVIDIA's DGX B300 user guide gives "8 x 288 GB = 2.3 TB total" [14]. Supermicro lists "HGX B300 8-GPU (288GB)" for its HGX B300 servers [19, 20], and Dell's August 2026 guide lists 288 GB B300 GPUs [21].

NVIDIA's blog notes in a footnote: "Blackwell Ultra GPUs contain up to 160 SMs and 288GB HBM3E Memory. Available SM count and HBM capacity varies by SKU" [5]. The GB300 NVL72 column of the datasheet shows a third figure, 279 GB [3]. In MLPerf® Inference v6.1, every B300 system lists its accelerator as "B300-SXM-270GB", including NVIDIA's DGX B300 and the Supermicro SYS-822GS-NB3RT [34], whose product page lists 288 GB [19]. The sources do not say which figure a running system reports to software.

HGX B200: 180 GB or 192 GB per GPU. The Blackwell datasheet lists 180 GB per GPU and 1.4 TB per system for HGX B200 [2], and the DGX B200 user guide gives 1,440 GB [16]. NVIDIA's technical brief lists "Up to 192 GB HBM3e" per GPU for HGX B200 [4], and its blog gives 192 GB as Blackwell's maximum [5]. Dell lists 180 GB B200 GPUs [21], and MLPerf® entries use "B200-SXM-180GB" [34, 35].

Bandwidth. HGX B200 and HGX B300 both list 7.7 TB/s per GPU and 62 TB/s per system [2, 3, 4]. The DGX B200 page lists 64 TB/s [15]. NVIDIA's blog gives 8 TB/s per GPU for Blackwell Ultra [5], the figure in the datasheet's GB300 NVL72 column [3]. Chip maxima and rack-scale parts carry higher figures than the HGX boards.

Host memory. Each DGX build ships with 2 TB of system memory, and DGX B200 and B300 can be configured to 4 TB [13, 15, 17]. Dell relays NVIDIA design guidance that eight-GPU servers need more than 1 TB of host memory with B200 and more than 2 TB with B300 [21].

What memory does for inference. NVIDIA's H200 page shows a 1.9 times gain over H100 on a 70-billion-parameter model by running one H200 at batch size 32 against one H100 at batch size 8 [6]: memory allowed a larger batch at the same compute rate.

3. Tensor throughput by precision, dense and sparse

NVIDIA lists TF32 and lower-precision tensor rates with sparsity [1, 6]. A sparse rate assumes weights pruned to a 2:4 pattern, two zeros in every four contiguous values, which lets sparse tensor cores "complete the same effective calculation in half the time" [10]. NVIDIA introduced the pattern with Ampere. Pruning a trained

network to it can affect accuracy, which NVIDIA's workflow recovers by retraining [10]. Its Blackwell tables state that dense is half the sparse figure for FP8, INT8, FP16, BF16 and TF32, and print FP4 both ways: HGX B300's dense FP4 is 108 of 144 PFLOPS [1, 2, 3].

Eight GPUs, sparse and dense	HGX H200 (d)	HGX B200	HGX B300
FP4	none	144 and 72 PFLOPS	144 and 108 PFLOPS
FP8	31.7 and 15.8 PFLOPS	72 and 36 PFLOPS	72 and 36 PFLOPS
INT8	31.7 and 15.8 POPS	72 and 36 POPS	3 and 1.5 POPS
FP16 and BF16	15.8 and 7.9 PFLOPS	36 and 18 PFLOPS	36 and 18 PFLOPS
TF32	7.9 and 4.0 PFLOPS	18 and 9 PFLOPS	18 and 9 PFLOPS
FP32	536 TFLOPS	600 TFLOPS	600 TFLOPS

Sources: [1, 6]. H200 figures and the HGX B300 dense INT8 figure are derived.

Per GPU, NVIDIA's technical brief prints dense and sparse side by side: FP4 at 9 and 18 PFLOPS for B200 and 14 and 18 for B300, FP8 at 4.5 and 9 for both, and INT8 at 4.5 and 9 petaOPS for B200 against 0.15 and 0.30 for B300 [4]. Per-GPU and per-system figures do not match exactly: HGX B300's 108 PFLOPS of dense FP4 is 13.5 per GPU, and eight times the datasheet's 307 TOPS of sparse INT8 is 2.46 POPS against "3 POPS" on the HGX page (derived from [1, 3]).

Chip figures and board figures. NVIDIA's Blackwell Ultra blog gives chip maxima: FP8 at 5 PFLOPS dense and power up to 1,200 W for Blackwell and 1,400 W for Blackwell Ultra [5], the levels of the NVL72 parts [2, 3]. The HGX boards run at 4.5 PFLOPS dense FP8 and up to 1,000 W and 1,100 W [2, 3].

FP4. NVFP4, introduced with Blackwell, stores 4-bit values with an FP8 scale for each block of 16 values and an FP32 scale per tensor. NVIDIA reports 1% or less accuracy degradation against FP8 on the reasoning model it tested, and a memory footprint about 3.5 times smaller than FP16 and 1.8 times smaller than FP8 [9]. Hopper has no FP4 tensor rate [5].

Attention. NVIDIA rates Blackwell Ultra's attention-layer compute at twice Blackwell's, from doubled special-function-unit throughput for softmax: 10.7 against 5 TeraExponentials per second [4, 5]. NVIDIA aims the gain at reasoning models with long context windows [5]. No source here measures it in an application.

Projected claims. NVIDIA's "up to 15x" inference claim for HGX B200 over Hopper describes one scenario, a 1.8-trillion-parameter mixture-of-experts model at a 32,768-token input, and is labeled "Projected performance, subject to change" [2, 15].

4. FP64 and HPC workloads

Native rates. H200 SXM lists 34 TFLOPS of FP64 and 67 TFLOPS on FP64 tensor cores per GPU [6]. HGX B200 lists 37 TFLOPS for "FP64/FP64 Tensor Core" per GPU and 296 TFLOPS for eight [1, 2]. HGX B300 lists 1.2 TFLOPS per GPU and 10 TFLOPS for eight [1, 3]. For FP64 matrix work the order runs H200, B200, then B300, at about 2% of H200's tensor-core rate (derived).

NVIDIA's guidance. The NVIDIA HPC-Benchmarks container supports DGX B200 and B300 clusters. Its notes say that B300 "has low native FP64 throughput, therefore, the performance of the NVIDIA HPL Benchmark on this hardware is expected to be low", and that HPL "supports FP64 emulation on (G)B300 using the Ozaki-II scheme" [22]. NVIDIA's August 2025 line card leaves the HPC and scientific column blank for HGX B300 and GB300 NVL72 and rates HGX B200 and HGX H200 there [11].

Emulation in cuBLAS. cuBLAS emulates higher-precision matrix multiplication by splitting inputs into lower-precision values and recombining the partial results [23]. FP64 emulation uses fixed-point algorithms based on the Ozaki-I and Ozaki-II schemes, from CUDA 13.0 Update 2, on compute capability 10.x, which includes B300 (sm103) [22, 23]. By default it chooses enough mantissa bits to keep accuracy "equal or better than FP64" and dispatches to native FP64 when the bits needed exceed a library default [23]. Emulation may handle infinite and NaN inputs differently from a standard implementation [23]. NVIDIA's blog describes this first release as FP64 emulation "on Blackwell INT8 Tensor Cores" and reports speedups on a B200 [24].

The INT8 problem on B300. INT8-based emulation runs on the units B300 cut from 4.5 to 0.15 petaOPS dense per GPU [4]. Uchino, Ozaki and Imamura describe that cut as making "reliance on INT8 alone insufficient" and present an Ozaki-II variant on FP8 units [25]. On one B200 at matrix size 16,384, their INT8-based method reached 137 to 138 TFLOPS and their FP8-based method 61 to 65 TFLOPS [25]. Because B200 and B300 share FP8 throughput and memory bandwidth, the authors expect similar FP8-based throughput on a B300 [25]. That is an expectation, and the paper reports no B300 measurement. At matrix size 1,024, both methods ran slower than native FP64 on the B200 [25]. The paper is a preprint.

Emulation covers matrix multiplication [23]. Other FP64 work runs at the native rate.

5. NVLink, PCIe and scale-out networking

NVLink. H100 and H200 use fourth-generation NVLink at 900 GB/s per GPU, through four NVSwitch chips in DGX H100 and H200 [6, 7, 18]. B200 and B300 use fifth-generation NVLink: 18 links per GPU, 1.8 TB/s in total, 900 GB/s each way, double the fourth generation [4]. Each Blackwell HGX board has 14.4 TB/s of total NVLink bandwidth, through two NVLink switch chips in DGX B200 and B300 [1, 14, 16].

Host link. H200 and HGX B200 list PCIe Gen5 at 128 GB/s per GPU [2, 4, 6], and HGX B300 PCIe Gen6 at 256 GB/s [3, 4]. NVIDIA's Blackwell Ultra blog lists PCIe Gen6 for Blackwell too [5], which differs from the HGX B200 datasheet and brief [2, 4]. In DGX B300 the ConnectX-8 cards sit in the GPU tray and connect directly to the GPUs over PCIe Gen6 [14].

Cluster networking. DGX H200 and DGX B200 each carry eight ConnectX-7 adapters at up to 400 Gb/s [15, 16, 17, 18]. DGX B300 carries eight ConnectX-8 adapters at up to 800 Gb/s [12, 14]. NVIDIA's HGX page lists networking bandwidth of 1.6 TB/s for HGX B300 and 0.8 TB/s for HGX B200 [1]. Eight 800 Gb/s ports carry 0.8 TB/s each way, so 1.6 TB/s counts both directions, and eight 400 Gb/s ports give 0.8 TB/s on the same basis (derived).

Builds differ here. NVIDIA's line card pairs HGX B300 and HGX B200 with Quantum-X800 InfiniBand and ConnectX-8 SuperNICs, and HGX H200 with Quantum-2 and ConnectX-7 [11], while DGX B200 ships with ConnectX-7 [15]. Dell's XE9780 puts eight ConnectX-8 ports on its B300 configuration and leaves four PCIe slots free, against up to twelve with B200 [21].

6. Power, size, weight and inlet temperature

DGX system	DGX H200	DGX B200	DGX B300
Power	10.2 kW max, standard configuration. Up to 14.3 kW with the custom thermal solution [17]	14.3 kW max [15, 16]	About 14 kW [12]. 14.5 kW busbar, 15.1 kW PSU [13]. 14.5 kW, 15 kW system max [14]
Power supplies	6 x 3.3 kW, 4+2 [18]	6 x 3.3 kW, 5+1 [16]	12, N+N, listed as 3.2 kW and as 3.3 kW [14]
Height and weight	8U, 130.45 kg [17, 18]	10U, 142.4 kg [16]	10U, 168 kg with PSUs, 123 kg busbar version [14]
Operating temperature	5 to 30 °C [17, 18]	10 to 35 °C [15, 16]	10 to 35 °C [13, 14]
Host CPUs	2 x Xeon Platinum 8480C [17, 18]	2 x Xeon Platinum 8570 [15, 16]	2 x Xeon 6776P [12, 14]

DGX B300's power figures. The product page says "~14 kW" [12]. The February 2026 datasheet says "14.5 kW (Busbar)" and "15.1 kW (PSU)" [13]. The user guide says 14.5 kW in its physical specifications and "15 kW system max." in its power table, for the AC and the 54 V DC versions [14]. The same guide lists "12 x 3.2 kW" supplies in

one table and "3.3 kW @ 200-240 V" per supply in another, and requires at least six supplies [14]. A rack plan built on 14 kW and one built on 15.1 kW differ by 1.1 kW per system (derived).

AC and busbar versions. DGX B300 comes with twelve power supplies for N+N redundancy or as a busbar version for a busbar-equipped rack, with power shelves not included [14]. NVIDIA lists the two versions among its enhancements over earlier DGX systems and describes DGX B300 as deployable in NVIDIA MGX racks for the first time [12, 14].

When supplies fail. In DGX B200, losing one of six supplies has no effect. With three or four left, maximum GPU power drops from 1,000 W to 800 W, and with two the system shuts down [16].

Temperature and power per GPU. DGX H200 runs at 5 to 30 °C, and DGX B200 and B300 at 10 to 35 °C [13, 15, 16, 17, 18]. The maximums come to about 1.28 kW per GPU for DGX H200, 1.79 kW for DGX B200 and 1.81 kW for DGX B300 at 14.5 kW, covering CPUs, memory, network cards, storage and fans as well (derived).

Other builds of the same boards. Supermicro puts HGX B300 in an 8U air-cooled server at 119 kg and in a 4U server with direct-to-chip cold plates at 84.6 kg, sold with a complete rack and onsite service [19, 20]. Dell's XE9780 is a 10U air-cooled server of 159.65 kg with B300, with twelve 3,200 W supplies [21]. All three list 10 to 35 °C [19, 20, 21].

Measured power. In MLPerf® Inference v5.1, a Lenovo ThinkSystem SR680a V3 with eight B200 GPUs at a 1,000 W GPU power setting averaged 10,434.7 W at the wall in the server scenario of the 70-billion-parameter benchmark and 10,401.0 W offline [35]. MLCommons measures average AC power for the whole system at the wall and does not measure or validate TDP settings or power supply ratings [26]. A rated maximum and a measured average answer different questions.

7. Reading measured results in MLPerf®

How the benchmark works. MLPerf® Inference: Datacenter measures how fast systems process inputs with a trained model, and its Closed division requires the reference model so hardware and software can be compared [26, 28]. In the server and interactive scenarios queries arrive at random intervals under latency limits, and in the offline scenario all at once [33]. For the 70-billion-parameter question-answering benchmark, the limits are 2,000 ms to the first token and 200 ms per output token in the conversational category and 450 ms and 40 ms in the interactive category, with an accuracy gate of 99.9% of the FP32 reference [33]. MLCommons notes that published results are sometimes modified or invalidated [26].

The benchmarks used here. A 70-billion-parameter question-answering model [33]. A 671-billion-parameter mixture-of-experts reasoning model with 37 billion parameters active per token [30]. A 117-billion-parameter mixture-of-experts model with 5.1 billion active per token, scored on math, science and code [31, 33]. MLCommons published v5.1 results on 9 September 2025 and v6.1 results on 16 September 2026 [27, 29].

B200 against H200, v5.1. Closed division, Available, eight accelerators in one node, 70-billion-parameter benchmark, tokens per second [35]:

Result ID	Submitter	System	Weights	Server	Offline	Interactive
5.1-0073	Nebius	Nebius B200 (8x B200-SXM-180GB)	FP4	101,611.0	101,246.0	59,622.7
5.1-0075	Nebius	Nebius H200 (8x H200-SXM-141GB)	FP8	34,029.4	34,812.1	23,079.9
5.1-0069	NVIDIA	NVIDIA DGX B200 (8x B200-SXM-180GB)	FP4	99,123.0	101,527.0	59,545.4
5.1-0026	Dell	Dell PowerEdge XE9680 (8x H200-SXM-141GB)	FP8	33,244.3	35,316.7	21,915.6

From the same submitter, eight B200s produced 2.99 times the tokens per second of eight H200s in the server scenario, 2.91 times offline and 2.58 times in the interactive scenario (derived). NVIDIA's DGX B200 against Dell's H200 server, two submitters, gives 2.98, 2.87 and 2.72 times (derived). Every B200 row uses FP4 weights and every H200 row FP8 [35]. H200 has no FP4 tensor rate [5]. MLCommons' working group chairs credit FP4 in place of FP8, new accelerators and software for a 5.58 times rise in median per-accelerator performance on this benchmark over six rounds [28].

B300 against B200, v6.1. Closed division, Available, eight accelerators in one node, FP4 weights on every row, tokens per second [34]:

Result ID	Submitter	System	70B server	70B offline	671B server	671B offline	117B server	117B offline
6.1-0021	CoreWeave	CoreWeave B200 SXM	102,398	102,703	56,121.3	58,913.2	90,204.7	91,487.4
6.1-0022	CoreWeave	CoreWeave B300 SXM	114,576	115,530	no entry	no entry	111,247	112,840
6.1-0077	Nebius	Nebius B200 n1	no entry	no entry	55,869.0	56,929.9	89,856.3	88,491.0
6.1-0078	Nebius	Nebius B300 n1	no entry	no entry	63,909.3	69,655.1	108,124	108,529
6.1-0071	NVIDIA	NVIDIA DGX B300	106,949	109,277	59,867.8	70,132.3	100,630	106,637

Same-submitter ratios, B300 over B200 (derived): CoreWeave 1.12 times on the 70B benchmark in both scenarios and 1.23 times on the 117B benchmark. Nebius 1.14 times in the server scenario and 1.22 times offline on the 671B benchmark, and 1.20 and 1.23 times on the 117B benchmark. Nebius lists identical software for its two systems, and CoreWeave's B300 entry lists one more software component than its B200 entry [34]. Against datasheet gains of 1.5 times for dense FP4 and 2 times for attention [1], the measured gain is 1.1 to 1.2 times.

Spread across builds. Thirteen single-node, eight-GPU B300 systems reported the 70B benchmark's server scenario in v6.1, from 96,180.6 tokens per second (6.1-0036, Dell XE9780) to 114,832 (6.1-0045, GigaComputing G894-AD3-AAX7) [34]. Server design and software moved results by up to about 19% on the same GPU (derived).

Gaps. v6.1 has no single-node eight-H200 SXM result on these benchmarks in the Closed, Available category, so no same-round comparison of H200 with B300 exists, and none of the B300 or H200 results above carries a power measurement [34, 35]. MLCommons' guidelines allow comparisons only between compatible versions of the same benchmark and scenario, and require derived metrics to be labeled [32].

MLPerf® notes: MLPerf® Inference: Datacenter v5.1 and v6.1, Closed division, Available category, benchmarks llama2-70b-99.9, deepseek-r1 and gpt-oss-120b, Server, Offline and Interactive scenarios as shown. Result IDs 5.1-0026, 5.1-0061, 5.1-0069, 5.1-0073, 5.1-0075, 6.1-0021, 6.1-0022, 6.1-0036, 6.1-0045, 6.1-0071, 6.1-0077 and 6.1-0078. Retrieved 24 October 2026 from the MLCommons results repositories [34, 35]. Ratios and percentages are this report's arithmetic and are not MLPerf® metrics. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

8. Which specifications matter for which workloads

No source ranks the systems for a buyer, and this report does not.

Workload trait	Specification that limits it	H200	B200	B300
Large weights, long context, many users	GPU memory per system	1,128 GB	1.4 TB	2.1 or 2.3 TB
Token generation bound by memory reads	Memory bandwidth per GPU	4.8 TB/s	7.7 TB/s	7.7 TB/s
FP4 inference	Dense FP4 per system	none	72 PFLOPS	108 PFLOPS

Workload trait	Specification that limits it	H200	B200	B300
FP8 or BF16 training	Dense FP8 and BF16 per system	15.8 and 7.9 PFLOPS (d)	36 and 18 PFLOPS	36 and 18 PFLOPS
Double-precision simulation	FP64 per system	536 TFLOPS tensor (d)	296 TFLOPS	10 TFLOPS
INT8-quantized inference	Dense INT8 per GPU	about 2 POPS (d)	4.5 POPS	0.15 POPS

Sources: [1, 4, 6, 14, 18].

B300's datasheet gains over B200 sit in memory capacity, FP4 math, attention and networking, and a workload bound by memory bandwidth or FP8 math has the same datasheet ceiling on both [1, 3]. H200 and B200 keep native FP64 and INT8 rates that B300 gives up [1, 22]. Measured gains, about three times from H200 to B200 and 1.1 to 1.2 times from B200 to B300, differ from datasheet ratios [34, 35].

Implications

For GPU infrastructure leads and ML platform engineers. Chip tables and NVL72 racks list higher FP8, bandwidth and power figures than the HGX boards, and DGX builds differ from other makers' HGX servers [2, 3, 5, 14, 21]. Two questions settle many of the gaps in this report: which B300 part a system carries, 270 GB or 288 GB, and which network card, since DGX B200 ships ConnectX-7 while NVIDIA's line card pairs HGX B200 with ConnectX-8 [11, 14, 15]. For throughput, MLPerf® results from the same submitter and round at a stated precision give a closer view than datasheet ratios [32, 34].

For HPC and scientific computing teams. Native FP64 on HGX B300 is 10 TFLOPS per system, against 296 TFLOPS on HGX B200 and about 536 TFLOPS on FP64 tensor cores for eight H200s (derived) [1, 6]. On B300, large FP64 matrix products depend on emulation, and NVIDIA expects low HPL results [22, 23, 25]. Emulated results vary with matrix size and library version [23, 25].

For facility and site planners. A 10U DGX B200 or B300 is rated at 14.3 to 15.1 kW maximum at 10 to 35 °C, against 10.2 kW for an 8U DGX H200 at 5 to 30 °C [13, 14, 16, 17, 18]. DGX B300's busbar version needs power shelves in the rack, and liquid-cooled HGX B300 builds exist in 4U [14, 20]. A DGX B300 planning figure needs its source document named beside it [12, 13, 14].

For procurement and finance teams. Servers on the same HGX board differ in CPU, host memory, network cards, chassis, cooling and power supplies [19, 20, 21]. Rillor's GPU systems service specifies, sources and delivers HGX H100, H200, B200 and B300 systems, new or secondary, and each quote states the exact build, its condition and the tests run before delivery.

Limits of this analysis

- Every specification comes from manufacturer documents and gives peak rates. NVIDIA marks H200 figures as preliminary [6].
- HGX H200 has no current NVIDIA HGX table. Its eight-GPU figures are derived from per-GPU figures, the 2023 announcement and DGX H200 documents [6, 8, 17, 18].
- The sources do not say which memory figure, 270 GB or 288 GB, a given HGX B300 reports to software, beyond NVIDIA's note that capacity varies by SKU [5].
- The MLPerf® comparisons cover inference on three language benchmarks. MLPerf® Training results were not examined, so training comparisons rest on specifications only.
- Same-submitter pairs reduce differences in software and server design without removing them. NVIDIA's own DGX B300 entry lists Xeon Platinum 8570 host processors, while DGX B300 documents list Xeon 6776P [14, 34].
- No measured FP64 emulation result on a B300 and no HPL result on an HGX B300 appear in the sources. The FP8-based emulation figure for B300 is its authors' expectation [25].
- The only measured system power for these GPUs in the results examined comes from two Lenovo B200 servers in v5.1, and only one of them, the ThinkSystem SR680a V3, ran the 70-billion-parameter benchmark used here [35].
- The report covers NVIDIA boards only and contains no prices, lead times or availability statements.

Sources

1. NVIDIA HGX Platform. NVIDIA, viewed 24 October 2026. https://www.nvidia.com/en-us/data-center/hgx/
2. Datasheet for NVIDIA Blackwell Architecture. NVIDIA, October 2025. https://dam-cdn.nvd.orangelogic.com/AssetLink/y441155802qub41q118b2852i557jem5.pdf
3. Datasheet for NVIDIA Blackwell Ultra. NVIDIA, October 2025. https://dam-cdn.nvd.orangelogic.com/AssetLink/1k0p832eq8r5ca0u5383ie5o4tp3bst1.pdf
4. NVIDIA Blackwell Architecture Technical Brief, version 2.1. NVIDIA, September 2025. https://dam-cdn.nvd.orangelogic.com/AssetLink/gl2l4l4812s5fw0p614s6i8bv6mi3vx5.pdf
5. Inside NVIDIA Blackwell Ultra: The Chip Powering the AI Factory Era. NVIDIA Technical Blog, 22 August 2025, updated 24 September 2025. https://developer.nvidia.com/blog/inside-nvidia-blackwell-ultra-the-chip-powering-the-ai-factory-era/
6. NVIDIA H200 GPU. NVIDIA, viewed 24 October 2026. https://www.nvidia.com/en-us/data-center/h200/
7. NVIDIA H100 GPU. NVIDIA, viewed 24 October 2026. https://www.nvidia.com/en-us/data-center/h100/
8. NVIDIA Supercharges Hopper, the World's Leading AI Computing Platform. NVIDIA Newsroom, 13 November 2023. https://nvidianews.nvidia.com/news/nvidia-supercharges-hopper-the-worlds-leading-ai-computing-platform
9. Introducing NVFP4 for Efficient and Accurate Low-Precision Inference. NVIDIA Technical Blog, 24 June 2025. https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/
10. Accelerating Inference with Sparsity Using the NVIDIA Ampere Architecture and NVIDIA TensorRT. NVIDIA Technical Blog, 20 July 2021. https://developer.nvidia.com/blog/accelerating-inference-with-sparsity-using-ampere-and-tensorrt/
11. NVIDIA Data Center Platform Line Card. NVIDIA, August 2025. https://docs.nvidia.com/data-center-gpu/line-card.pdf
12. NVIDIA DGX B300. NVIDIA, viewed 24 October 2026. https://www.nvidia.com/en-us/data-center/dgx-b300/
13. NVIDIA DGX B300: AI Performance From Training to Inference (datasheet). NVIDIA, February 2026. https://dam-cdn.nvd.orangelogic.com/AssetLink/625w0j7hnw6f07ui7fo63211mti2v08o.pdf
14. Introduction to NVIDIA DGX B300 System (NVIDIA DGX B300 User Guide). NVIDIA, viewed 24 October 2026. https://docs.nvidia.com/dgx/dgxb300-user-guide/introduction-to-dgxb300.html
15. NVIDIA DGX B200. NVIDIA, viewed 24 October 2026. https://www.nvidia.com/en-us/data-center/dgx-b200/
16. Introduction to NVIDIA DGX B200 Systems (NVIDIA DGX B200 User Guide). NVIDIA, viewed 24 October 2026. https://docs.nvidia.com/dgx/dgxb200-user-guide/introduction-to-dgxb200.html
17. NVIDIA DGX H200 datasheet. NVIDIA, June 2024. https://dam-cdn.nvd.orangelogic.com/AssetLink/2kadwg5l260j3f07hlngs58an1r5ga2k.pdf

18. Introduction to NVIDIA DGX H100/H200 Systems (NVIDIA DGX H100/H200 User Guide). NVIDIA, viewed 24 October 2026. https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html

19. SYS-822GS-NB3RT: DP Intel air cooled GPU server with NVIDIA HGX B300 8-GPU. Super Micro Computer, viewed 24 October 2026. https://www.supermicro.com/en/products/system/gpu/8u/sys-822gs-nb3rt

20. SYS-422GS-NB3RT-LCC: DP Intel 4U Liquid-Cooled System with NVIDIA HGX B300 8-GPU. Super Micro Computer, viewed 24 October 2026. https://www.supermicro.com/en/products/system/gpu/4u/sys-422gs-nb3rt-lcc

21. Dell PowerEdge XE9780 Technical Guide, Rev. A03. Dell Technologies, August 2026. https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9780-technical-guide.pdf

22. NVIDIA HPC-Benchmarks (container, tag 26.02.01). NVIDIA NGC Catalog, viewed 24 October 2026. https://catalog.ngc.nvidia.com/orgs/nvidia/containers/hpc-benchmarks

23. cuBLAS documentation, section 1.5, Floating Point Emulation. NVIDIA, viewed 24 October 2026. https://docs.nvidia.com/cuda/cublas/index.html

24. Unlocking Tensor Core Performance with Floating Point Emulation in cuBLAS. NVIDIA Technical Blog, 24 October 2025. https://developer.nvidia.com/blog/unlocking-tensor-core-performance-with-floating-point-emulation-in-cublas/

25. Uchino, Y., Ozaki, K. and Imamura, T. Double-Precision Matrix Multiplication Emulation via Ozaki-II Scheme with FP8 Quantization. arXiv:2603.10634v2, 6 April 2026. https://arxiv.org/abs/2603.10634

26. Benchmark MLPerf Inference: Datacenter. MLCommons, viewed 24 October 2026. https://mlcommons.org/benchmarks/inference-datacenter/

27. MLCommons Sets Participation Record with New MLPerf Inference v6.1 Benchmark Results. MLCommons, 16 September 2026. https://mlcommons.org/2026/09/mlperf-inference-v6-1-results/

28. Where the Industry Is Investing: A Look at MLPerf Inference v6.1. MLCommons, 17 September 2026. https://mlcommons.org/2026/09/chairs-mlperf-inference-v6-1/

29. MLCommons Releases New MLPerf Inference v5.1 Benchmark Results. MLCommons, 9 September 2025. https://mlcommons.org/2025/09/mlperf-inference-v5-1-results/

30. DeepSeek Reasoning for MLPerf Inference v5.1. MLCommons, 9 September 2025. https://mlcommons.org/2025/09/deepseek-inference-5-1/

31. A new GPT-OSS benchmark and DeepSeek R1 updates for latency-optimized reasoning. MLCommons, 24 March 2026. https://mlcommons.org/2026/03/mlperf-inference-gpt-oss/

32. MLPerf® Results Messaging Guidelines. MLCommons, viewed 24 October 2026. https://github.com/mlcommons/policies/blob/master/MLPerf_Results_Messaging_Guidelines.adoc

33. MLPerf Inference Rules. MLCommons, viewed 24 October 2026. https://github.com/mlcommons/inference_policies/blob/master/inference_rules.adoc

34. MLPerf Inference v6.1 results, datacenter closed division (summary.xlsx and summary.json). MLCommons, published 16 September 2026. https://github.com/mlcommons/inference_results_v6.1/raw/refs/heads/main/summary.xlsx

35. MLPerf Inference v5.1 results (summary_results.json). MLCommons, published 9 September 2025. https://raw.githubusercontent.com/mlcommons/inference_results_v5.1/main/summary_results.json

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. HGX H200, B200 and B300 compared. 7 November 2026.
https://rillor.com/insights/hgx-h200-b200-b300-compared

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

NVIDIA, AMD, Ceph, Kubernetes and related product names are trademarks of their owners and are used only to identify products. No affiliation or endorsement is implied.

Work with Rillor

Send the workload, system, quantity, delivery window and destination. Contact us at rillor.com/contact?type=gpu-systems or contact@rillor.com. More on GPU systems: rillor.com/compute/gpu-systems