NVIDIA Data Center GPU Comparison: H100 vs B200 vs A100

Published July 10, 2026 44 min read

Executive Summary

NVIDIA's data center GPU lineup spans four active generations as of July 2026: the Ampere-based **A100**, the Hopper-based **H100** and **H200**, and the Blackwell-based **B200** and rack-scale **GB200 NVL72**, with the Blackwell Ultra **GB300 NVL72** now shipping into hyperscaler fleets and the next-generation **Rubin** platform announced for second-half 2026 availability (Source: nvidianews.nvidia.com). The **H100** remains the industry's workhorse, offering 80GB or 94GB of memory, up to 3,958 teraFLOPS of FP8 Tensor Core throughput, and up to 700W of configurable thermal design power (TDP) in its SXM form factor (Source: www.nvidia.com). The **H200** keeps the same Hopper compute engine but nearly doubles memory to 141GB of HBM3e at 4.8 terabytes per second (TB/s) of bandwidth, making it the preferred choice for memory-bound large language model (LLM) inference on Hopper-class hardware (Source: www.nvidia.com). The **B200**, NVIDIA's flagship Blackwell-architecture accelerator, jumps to 180GB of HBM3e at roughly 8 TB/s bandwidth, a dual-die 208-billion-transistor design, and native FP4 precision that delivers 20 petaFLOPS per GPU (Source: www.runpod.io).

For the highest-throughput training and inference at scale, NVIDIA's answer is not a single chip but the **GB200 NVL72**, a liquid-cooled rack that unites 36 Grace CPUs and 72 Blackwell GPUs into a single NVLink domain delivering 13.4 terabytes of HBM3e at 576 TB/s and 1,440 petaFLOPS of sparse NVFP4 compute (Source: www.nvidia.com). In MLPerf Training v5.0, GB200 NVL72 delivered up to 2.6x more performance per GPU than Hopper, including a 2.2x speedup on the Llama 3.1 405B pretraining benchmark at 512-GPU scale (Source: developer.nvidia.com). The newer GB300 NVL72, built on the Blackwell Ultra architecture, adds up to 288GB of HBM3e per GPU and posted 45% higher DeepSeek-R1 inference throughput than GB200 NVL72 in MLPerf Inference v5.1 (Source: blogs.nvidia.com).

Cloud pricing as of July 2026 reflects this generational ladder. CoreWeave lists an 8-GPU HGX H100 node at $49.24 per hour on demand (about $6.16 per GPU-hour), an HGX H200 node at $50.44 per hour, and an HGX B200 node at $68.80 per hour (about $8.60 per GPU-hour), while GB200 NVL72 capacity runs $42.00 per GPU-hour on demand (see *Table 1* below). Lambda prices HGX B200 clusters from $8.87 to $9.86 per GPU-hour depending on cluster size, versus $5.54 to $6.16 per GPU-hour for H100 clusters (see *Table 2* below). A100 instances remain the budget option, with per-GPU on-demand pricing near $21.60 per hour on major clouds and aggregator getdeploying.com showing a market range as low as $0.13 per

hour across 68 providers (Source: getdeploying.com). Independent benchmarking firm Lightly AI reported up to 57% higher training throughput on B200 versus H100 for computer vision workloads, alongside potentially 6 to 30 times lower self-hosted operating cost per GPU-hour depending on utilization (Source: www.reddit.com).

NVIDIA's Data Center segment posted record revenue of $75.2 billion in the first quarter of fiscal 2027 (the quarter ended April 26, 2026), up 92% year over year, underscoring how central this GPU lineup is to the company's business (detailed in the Data Analysis section below). Real-world deployments illustrate the scale at stake: xAI's Colossus cluster in Memphis grew from 100,000 H100 GPUs built in 122 days to a mixed fleet of 150,000 H100, 50,000 H200, and 30,000 GB200 GPUs by December 2025 (detailed in the case studies section below), while Microsoft Azure brought online the first production GB300 NVL72 cluster with more than 4,600 Blackwell Ultra GPUs for OpenAI (Source: blogs.nvidia.com). For teams choosing among these GPUs today, the practical decision hinges on memory footprint versus model size, NVLink domain size versus cluster topology, and cost-per-token versus raw hourly price, all of which this report quantifies across the A100, H100, H200, B200, GB200 NVL72, and the emerging Rubin roadmap.

Introduction and Background

Choosing among NVIDIA's data center GPUs has become one of the highest-stakes procurement decisions in enterprise computing. A single miscalculation, provisioning **H100** capacity for a workload that needs the **H200**'s extra memory, or paying **B200** premiums for a job that an **A100** cluster could handle just as cheaply, can shift a training budget by millions of dollars. This report compares the five most commonly deployed NVIDIA data center accelerators as of July 2026: the **A100** (Ampere architecture, 2020), the **H100** and **H200** (Hopper architecture, 2022 and 2023), and the **B200** alongside the rack-scale **GB200 NVL72** (Blackwell architecture, 2024 to 2025), with reference to the newer **GB300 NVL72** (Blackwell Ultra) and the announced **Rubin** successor platform.

Each generational jump has followed a consistent pattern: more transistors, wider and faster High Bandwidth Memory (HBM), and a new lower-precision numeric format that roughly doubles throughput for the same silicon area. The **A100** introduced structural sparsity and the TF32 (Tensor Float 32) format, delivering up to 20x higher performance than its Volta-generation predecessor with zero code changes (Source: www.nvidia.com). The **H100** added a dedicated Transformer Engine with FP8 precision, delivering up to 4x faster GPT-3 (175B parameter) training than A100 clusters (Source: www.nvidia.com). The **H200** kept the same Hopper compute die (GH100) but became the first NVIDIA GPU to ship with HBM3e memory, addressing the memory-capacity bottleneck that constrained H100 for the largest LLMs (Source: www.runpod.io). The **B200** represents a more fundamental break: a dual-die Blackwell package with a second-generation Transformer Engine and native FP4 support, described by NVIDIA as "the largest GPU ever built with 2.5x the transistors of Hopper GPUs" (Source: techcommunity.microsoft.com).

Understanding these GPUs in isolation, however, understates how NVIDIA now sells compute. Since the Blackwell generation, the company's flagship product is arguably not a chip but a rack: the **GB200 NVL72** couples 36 Grace CPUs with 72 Blackwell GPUs over fifth-generation NVLink to create what NVIDIA describes as "a single, massive GPU" with a 130 TB/s NVLink domain (Source: www.nvidia.com). This report treats the individual GPU (A100, H100, H200, B200) and the rack-scale system (GB200 NVL72, GB300 NVL72) as complementary units of comparison, since real-world buying decisions increasingly happen at the rack or pod level rather than per-card. All specifications below are current as of July 2026 and cite NVIDIA's own datasheets alongside independent cloud pricing, MLPerf benchmark results, and named enterprise deployments.

This comparison matters beyond hardware specification sheets because the underlying business is now one of the largest in computing. NVIDIA's founder and chief executive officer (CEO) Jensen Huang described the current moment as "the largest infrastructure expansion in human history," accelerating "at extraordinary speed" as enterprises move from experimenting with generative AI to running agentic AI systems in production (Source: nvidianews.nvidia.com). Every generation covered in this report, from the four-year-old A100 to the newly announced Rubin platform, remains commercially relevant simultaneously: cloud providers rent all of them concurrently, and enterprise buyers routinely mix generations within a single fleet to match hardware cost to workload requirements rather than always chasing the newest silicon.

NVIDIA A100: The Established Baseline

Capabilities

Launched on the Ampere architecture, the **A100** Tensor Core GPU remains in production and in wide cloud availability as of mid-2026, primarily as a lower-cost option for workloads that do not require Hopper- or Blackwell-class throughput. The 80GB SXM4 variant delivers 9.7 teraFLOPS of FP64, 19.5 teraFLOPS of FP64 Tensor Core throughput, and up to 624 teraFLOPS of FP16 Tensor Core performance with sparsity, alongside 80GB of HBM2e memory at 2,039 gigabytes per second (GB/s) of bandwidth (Source: www.nvidia.com). Multi-Instance GPU (MIG) technology allows a single A100 to be partitioned into up to seven isolated GPU instances, each with 10GB of dedicated memory, letting operators right-size capacity for inference workloads that do not need a full GPU (Source: www.nvidia.com). The SXM4 form factor draws up to 400W TDP as standard, with a custom

thermal solution SKU supporting up to 500W, while the PCIe variant is capped at 300W (Source: www.nvidia.com). NVIDIA positions A100 as capable of up to 20x higher performance than the prior Volta generation, a jump the company attributes to third-generation Tensor Cores combined with structural sparsity support that doubles effective throughput on pruned models (Source: www.nvidia.com).

Adoption

The A100 was the default training GPU for the generative AI boom of 2022 to 2023 and still underpins a large share of installed base capacity. Cloud pricing aggregator getdeploying.com tracks A100 availability across more than 36 providers, with hourly rates spanning $0.13 to $5.04 depending on memory configuration (40GB or 80GB) and provider tier (Source: getdeploying.com). Lambda prices the 80GB SXM variant at $2.79 per hour and the 40GB variant at $1.99 per hour on its self-serve cloud (Source: lambda.ai), while other major clouds price single-GPU A100 capacity nearer $21.60 per hour on demand and roughly $9.65 per hour on spot markets (see *Table 2* in the Data Analysis section below). This wide pricing spread signals that A100 has become a commodity tier, useful for fine-tuning, smaller-model training, and inference workloads where the H100's or B200's extra throughput and memory would be underutilized.

Strengths and Limitations

The A100's core strength is cost efficiency for workloads under roughly 40 billion parameters or for inference tasks that fit within 40GB to 80GB of memory, especially when MIG partitioning is used to maximize utilization. Its principal limitation is the lack of FP8 or FP4 Tensor Core support and the absence of a dedicated Transformer Engine, meaning it cannot match H100-class or B200-class throughput on the mixed-precision techniques modern LLM training and inference pipelines rely on. NVIDIA's own comparison shows H100 delivering up to 4x faster GPT-3 175B training and up to 30x higher inference performance versus A100 on the largest models, illustrating how far the performance gap has widened over two generations (Source: www.nvidia.com). A100 also lacks the DPX (Dynamic Programming X) instructions NVIDIA introduced with Hopper, which accelerate dynamic-programming algorithms such as Smith-Waterman genome sequencing by up to 7x on H100 relative to A100 and by up to 40x relative to CPU-only systems, meaning HPC (high-performance computing) shops running bioinformatics or route-optimization workloads see a disproportionately larger gain from upgrading past A100 than pure LLM-training buyers do.

NVIDIA H100: The Hopper Workhorse

Capabilities

The **H100** Tensor Core GPU, built on the Hopper architecture, is available in SXM and PCIe form factors with materially different specifications. The SXM variant offers 80GB of HBM3 at 3.35 TB/s bandwidth, 900GB/s of fourth-generation NVLink interconnect, and up to 700W of configurable TDP, while a higher-memory 94GB PCIe variant (H100 NVL) offers 3.9 TB/s bandwidth at 350 to 400W TDP (Source: www.nvidia.com). Peak FP8 Tensor Core throughput reaches 3,958 teraFLOPS on the SXM variant, with INT8 Tensor Core performance at 3,958 TOPS (tera operations per second) (Source: www.nvidia.com). For HPC workloads, H100 delivers 67 teraFLOPS of FP64 Tensor Core throughput, triple the double-precision performance of the prior generation, as shown in *Table 1* below. H100 systems scale to 4 or 8 GPUs in NVIDIA HGX partner servers, or 8 GPUs in the NVIDIA DGX H100 appliance, the same 8-GPU node configuration AWS and CoreWeave both build their H100 cloud instances around (see the Adoption discussion below).

Adoption

H100 is the most widely deployed high-end AI training GPU in production today. Amazon Web Services (AWS) offers it through EC2 P5 instances, with the p5.48xlarge configuration providing 8 H100 GPUs and 640GB of aggregate HBM3 memory, scaling within EC2 UltraClusters to as many as 20,000 H100 or H200 GPUs delivering 20 exaflops of aggregate compute (Source: aws.amazon.com). CoreWeave prices an 8-GPU HGX H100 node at $49.24 per hour on demand, or roughly $6.16 per GPU-hour, with a spot rate of around $19.71 per hour (see *Table 2*). Meta built two flagship 24,576-GPU H100 clusters, one using a 400G RoCEv2 (RDMA over Converged Ethernet) backend on Arista switches and the other using 400G NDR InfiniBand on NVIDIA Quantum-2 switches, to train the Llama 3 model family (Source: www.glennklockwood.com). xAI's Colossus cluster in Memphis was built around 100,000 H100 GPUs deployed in just 122 days, at the time the fastest large-scale AI data center build publicly documented (see the xAI Colossus case study below).

Strengths and Limitations

H100's core strength is broad software and cloud ecosystem maturity: it remains the reference target for most training frameworks, container images, and MLPerf submissions, and it enjoys the deepest secondary and spot market liquidity of any current-generation NVIDIA data center GPU. Its principal limitation, relative to the H200 and B200, is memory capacity. At 80GB, the H100 SXM cannot hold the largest open-weight models (70B-plus parameters) without tensor parallelism across multiple GPUs, and it lacks the FP4 precision that gives Blackwell its inference throughput advantage. One Reddit user summarized the practical trade-off after testing both generations: "H100 jump was amazing for our inference and training jobs. 2.3x multiplier while the price difference was <2x per hr," referring to the H100-versus-A100 jump but illustrating the recurring pattern that each generation's price premium has historically been smaller than its throughput gain (Source: www.reddit.com). Other commenters on the same thread raised early Blackwell supply concerns, with one noting rumors of "inherent flaws in TSMC's Blackwell packaging process" causing "significant delays in large-scale production," a reminder that generational transitions carry manufacturing risk in addition to the pricing and performance trade-offs quantified elsewhere in this report (Source: www.reddit.com).

NVIDIA H200: The Memory-Optimized Hopper Refresh

Capabilities

The **H200** is built on the same Hopper compute die as H100 but pairs it with 141GB of HBM3e memory at 4.8 TB/s bandwidth, described by NVIDIA as "nearly double the capacity of the NVIDIA H100 GPU... with 1.4X more memory bandwidth" (Source: www.nvidia.com). Compute throughput is identical to H100 at the Tensor Core level, 3,958 teraFLOPS of FP8 on the SXM variant, since the two share the same GH100 silicon; the improvement is purely in memory subsystem capacity and bandwidth (Source: www.nvidia.com). NVIDIA reports the H200 delivers 1.9x faster Llama 2 70B inference and 1.6x faster GPT-3 175B inference than H100 as a result of this larger, faster memory pool, plus up to 110x faster time-to-results than CPU-only systems on HPC workloads (Source: www.nvidia.com). AWS lists P5e and P5en instances built on 8 H200 GPUs with up to 1,128GB of aggregate HBM3e memory per instance (Source: aws.amazon.com). NVIDIA's H200 datasheet also confirms up to seven MIG (Multi-Instance GPU) partitions of 18GB each on the SXM variant, letting operators subdivide a single H200 for multiple smaller inference workloads rather than dedicating the full 141GB to one job (Source: www.nvidia.com).

Adoption

CoreWeave prices HGX H200 capacity at $50.44 per hour on demand for an 8-GPU node, marginally above H100's $49.24 per hour, with a spot rate around $20.93 per hour (see *Table 2*). Getdeploying.com's market data shows H200 spanning $1.00 to $13.78 per hour across more than 28 providers (Source: getdeploying.com). xAI's Colossus cluster incorporated 50,000 H200 GPUs alongside its H100 and GB200 fleet as of its December 2025 configuration update, reflecting how operators mix generations within a single facility to match memory needs to specific model families (see the case study below).

Strengths and Limitations

H200's strength is straightforward: it is a near drop-in replacement for H100 hardware and software stacks that immediately relieves memory pressure, useful for serving 70B-to-100B-parameter models on a single GPU or reducing tensor-parallel sharding overhead. Because it shares H100's compute engine, it offers no FP4 support and no improvement in raw FP8 or BF16 (Brain Floating Point 16) compute throughput, meaning workloads that are compute-bound rather than memory-bound see limited benefit from upgrading. Runpod's comparison guide summarizes the trade-off plainly: "H200 is the right choice when 141 GB VRAM is sufficient and Hopper-level compute meets the requirement. B200 is justified for workloads that need maximum throughput or models exceeding 141 GB at full precision" (Source: www.runpod.io).

NVIDIA B200 and GB200 NVL72: The Blackwell Generation

Capabilities

The individual **B200** GPU packages two reticle-limit dies into a single Blackwell package with 208 billion transistors, 180GB of HBM3e memory, and approximately 8 TB/s of memory bandwidth (as introduced above). Per-GPU compute reaches 20 petaFLOPS of sparse FP4 Tensor Core throughput and 10 petaFLOPS of FP8/FP6, according to NVIDIA's Blackwell datasheet (Source: www.primeline-solutions.com). NVLink bandwidth doubles to 1.8 TB/s bidirectional per GPU under fifth-generation NVLink, versus 900GB/s on Hopper (see *Table 1* below). Note a discrepancy worth flagging for

buyers: independent guides and MLCommons list the shipping B200 SXM configuration as 180GB usable HBM3e (Source: mlcommons.org), while Microsoft's Azure announcement describes the same-generation Blackwell GPU memory as "192GB," reflecting the difference between total physical HBM3e capacity and the usable capacity after reserving headroom (Source: techcommunity.microsoft.com); buyers should confirm the exact usable figure with their hardware vendor rather than assume parity across sources. NVIDIA's Blackwell architecture datasheet describes the platform's broader feature set beyond raw throughput, noting that "the incorporation of the second-generation Transformer Engine, alongside the faster and wider NVIDIA NVLink interconnect, propels the data center into a new era," and that a new decompression engine paired with Spark RAPIDS libraries "deliver unparalleled database performance to fuel data analytics applications," extending Blackwell's relevance beyond pure LLM workloads into data analytics and confidential-computing use cases (Source: www.primeline-solutions.com). The same datasheet frames NVIDIA's rack-scale ambitions plainly, describing GB200 NVL72 as "powering the new era of computing" and listing the HGX B200 baseboard as the second of two "key offerings" alongside the full rack, underscoring that NVIDIA now sells Blackwell primarily as two standardized system tiers (8-GPU baseboard or 72-GPU rack) rather than as loose individual GPUs (Source: www.primeline-solutions.com).

The **DGX B200** server packages 8 Blackwell GPUs with 1,440GB of total GPU memory at 64 TB/s of aggregate HBM3e bandwidth, 14.4 TB/s of aggregate NVLink bandwidth, and roughly 14.3 kilowatts (kW) of maximum system power draw (Source: www.nvidia.com). NVIDIA states DGX B200 delivers 3x the training performance and 15x the inference performance of the previous-generation DGX H100 (Source: www.nvidia.com). The system pairs its 8 GPUs with 2 Intel Xeon Platinum 8570 processors totaling 112 CPU cores, 2TB of system memory configurable to 4TB, and four OSFP networking ports serving eight single-port ConnectX-7 SuperNICs at up to 400 gigabits per second (Gb/s) each (Source: www.nvidia.com). NVIDIA ships DGX B200 with a three-year Enterprise Business-Standard support contract covering both hardware and software, reflecting the appliance-style, fully supported positioning NVIDIA uses to differentiate DGX from bare HGX baseboards sold to server partners (Source: www.nvidia.com). Scaling further, the **GB200 NVL72** rack connects 36 Grace CPUs and 72 Blackwell GPUs into a single 72-GPU NVLink domain with 13.4 terabytes (TB) of HBM3e at 576 TB/s aggregate bandwidth, 130 TB/s of NVLink bandwidth, and 1,440 petaFLOPS of sparse NVFP4 compute (Source: www.nvidia.com). A single GB200 Grace Blackwell Superchip, pairing one Grace CPU with two Blackwell GPUs, delivers 40 petaFLOPS of sparse NVFP4 and 372GB of HBM3e at 16 TB/s (Source: www.nvidia.com). NVIDIA's Blackwell datasheet describes the GB200 NVL72 as connecting its 36 Grace CPUs and 72 Blackwell GPUs "in an NVIDIA NVLink-connected, liquid-cooled, rack-scale design" that acts, in the company's words, "as a single, massive GPU" for trillion-parameter inference workloads (Source: www.primeline-solutions.com). The same source describes the GB200 Superchip building block as "connecting two high-performance NVIDIA Blackwell GPUs and an NVIDIA Grace CPU with the NVLink-C2C interconnect," the chip-to-chip link that keeps Grace's LPDDR5X system memory within microsecond reach of both attached GPUs (Source: www.primeline-solutions.com).

Adoption

Microsoft Azure's ND GB200 v6 virtual machine series, generally available as of mid-2025, launched with a 4,000-GPU GB200 cluster and demonstrated 860,000 tokens per second of Llama 70B throughput per rack, a 9x increase over the previous-generation ND H100 v5 (Source: techcommunity.microsoft.com). CoreWeave prices GB200 NVL72 capacity at $42.00 per GPU-hour on demand (see *Table 1* above), and Lambda's B200 SXM6 on-demand instances run $6.69 to $6.99 per GPU-hour depending on cluster configuration (Source: lambda.ai). Runpod offers standalone B200 instances at $4.99 per hour on demand, dropping to $4.24 per hour on a one-year commitment (Source: www.runpod.io). xAI's Colossus 2 expansion in Memphis is targeting at least 110,000 GB200 GPUs, with roughly 30,000 already operational as of the cluster's December 2025 snapshot (see the case study below).

Strengths and Limitations

Blackwell's core advantage is its rack-scale NVLink domain: because GB200 NVL72 treats 72 GPUs as a single addressable accelerator, it eliminates much of the cross-node communication overhead that limits scaling efficiency on H100 or A100 clusters built from smaller 8-GPU NVLink islands. In MLPerf Training v5.0, this translated into a 2.2x speedup on Llama 3.1 405B pretraining (121.09 minutes versus 269.12 minutes on Hopper at the same 512-GPU scale) and up to 2.6x higher performance per GPU overall (Source: developer.nvidia.com). The trade-offs are cost, power density, and supply. Blackwell hardware commands roughly a 1.4x to 1.9x hourly price premium over Hopper hardware across the cloud providers surveyed in this report, and DGX B200's 14.3kW power draw per 10-rack-unit chassis requires liquid cooling infrastructure many existing data centers lack (Source: www.nvidia.com). NVIDIA's own HGX B200 baseboard datasheet lists per-GPU thermal design power as configurable up to 1,000W, roughly 40% higher than H100 SXM's 700W ceiling, which is the physical reason Blackwell deployments lean so heavily on liquid cooling even outside the rack-scale NVL72 form factor (Source: images.nvidia.com). The same baseboard datasheet lists the fully populated 8-GPU HGX B200 assembly at 32 kilograms (kg), a physical weight increase over prior HGX generations driven by the denser cooling hardware Blackwell's higher power envelope requires (Source: images.nvidia.com). The datasheet formally specifies the HGX B200 form factor as "8x NVIDIA Blackwell GPUs" on a single baseboard, the same 8-GPU building block CoreWeave, Lambda, and other clouds resell as HGX B200 instances at the per-GPU rates quoted

throughout this report (Source: images.nvidia.com). One Reddit thread on r/BetterOffline captured the resale-value risk buyers face amid rapid generational turnover, observing that "H100 GPUs have already lost 85% of their value" as Blackwell supply ramped, with commenters speculating the same fate awaits B200 as B300 and Rubin arrive (Source: www.reddit.com). Not every commenter agreed resale price tells the full story: the self-identified CEO of GPU cloud platform Thunder Compute pushed back that "resale value and demand are largely decoupled in this industry," noting that "H100 capacity is completely sold out across most of the industry and demand for these chips has only continued to climb," even as new data center build-outs shift toward Blackwell's different power and cooling requirements rather than expanding H100 fleets (Source: www.reddit.com).

Feature Comparison

Table 1 below summarizes the core specifications of the five GPUs (and the GB200 NVL72 rack system) covered in this report, drawn directly from NVIDIA's published datasheets and cross-checked against independent guides where NVIDIA's own pages did not list a figure.

SPEC	A100 (80GB SXM4)	H100 (SXM)	H200 (SXM)	B200 (SXM)	GB200 NVL72 (PER RACK)
Architecture	Ampere	Hopper	Hopper	Blackwell	Blackwell
GPU memory	80GB HBM2e	80GB HBM3	141GB HBM3e	180GB HBM3e	13.4TB HBM3e (72 GPUs)
Memory bandwidth	2,039 GB/s	3.35 TB/s	4.8 TB/s	~8 TB/s	576 TB/s aggregate
FP8/FP6 Tensor Core	not supported	3,958 TFLOPS	3,958 TFLOPS	10 PFLOPS	720 PFLOPS (72 GPUs)
FP4 Tensor Core	not supported	not supported	not supported	20 PFLOPS	1,440 PFLOPS (72 GPUs)
FP64 Tensor Core	19.5 TFLOPS	67 TFLOPS	67 TFLOPS	40 TFLOPS	2,880 TFLOPS (72 GPUs)
NVLink bandwidth	600 GB/s	900 GB/s	900 GB/s	1.8 TB/s	130 TB/s domain-wide
Max TDP	400W (SXM)	700W (SXM)	700W (SXM)	~1,000W (HGX config) (Source: images.nvidia.com)	n/a (rack-level, liquid-cooled)
Cloud price (8-GPU node, on demand)	~$21.60/GPU-hr (single-GPU) (Source: www.coreweave.com)	$49.24/hr (~$6.16/GPU-hr) (Source: www.coreweave.com)	$50.44/hr (~$6.31/GPU-hr) (Source: www.coreweave.com)	$68.80/hr (~$8.60/GPU-hr) (Source: www.coreweave.com)	$42.00/GPU-hr (Source: www.coreweave.com)

Reading the table left to right shows the consistent trajectory: each generation roughly doubles memory bandwidth and adds a new lower-precision numeric format, while raw hourly pricing rises more slowly than raw throughput, meaning performance per dollar has generally improved with each generation despite higher list prices. The GB200 NVL72 row is a rack aggregate rather than a per-GPU figure for memory and FP4/FP8 columns, so it is not directly comparable cell-by-cell with the single-GPU columns; it is included because most large training and inference deployments today provision at the rack level rather than the individual card level.

Performance and Benchmarks

Independent benchmark data from MLCommons' MLPerf suite provides the most rigorously comparable throughput figures across generations, since all submissions use the same reference model, dataset, and quality threshold. In MLPerf Training v5.0, published in June 2025, NVIDIA's Blackwell-based GB200 NVL72 system delivered the fastest time-to-train across all seven benchmarks in the suite, spanning LLM pretraining, LLM fine-tuning, text-to-image generation, recommender systems, graph neural networks, natural language processing, and object detection (Source: developer.nvidia.com). On the newly introduced Llama 3.1 405B pretraining benchmark, Blackwell trained the model in 20.8 minutes at scale, versus 269.12 minutes for Hopper at the comparable 512-GPU submission (Source: developer.nvidia.com). The Llama 2 70B LoRA (Low-Rank Adaptation) fine-tuning benchmark improved 2.10x from the prior v4.1 round, and the Stable Diffusion text-to-image benchmark improved 2.28x, both outpacing historical Moore's Law-style expectations for six-month hardware cycles (Source: mlcommons.org). This round drew 201 performance results from 20 submitting organizations, including AMD, CoreWeave, Dell Technologies, Google Cloud, Lambda, and NVIDIA itself, testing hardware that included NVIDIA's Blackwell GB200 and B200-SXM-180GB configurations alongside AMD's Instinct MI300X and MI325X and Google's TPU-trillium (Source: mlcommons.org). MLPerf Training working group co-chair Hiwot Kassa characterized the broader trend behind these results: "AI workloads are scaling up, systems are scaling up to run them, and hardware innovation continues to boost performance for key scenarios," adding that "the increased proliferation, and competition, in AI-optimized systems is enabling the broader community to scale up their own infrastructure," with cloud service providers increasingly "democratizing access to training large models" that were previously the province of a handful of well-capitalized labs (Source: mlcommons.org). The number of multi-node system submissions increased more than 1.8x compared with the prior v4.1 round, a further sign that rack-scale systems like GB200 NVL72 are becoming the default unit of benchmark submission rather than single 8-GPU servers (Source: mlcommons.org).

On the inference side, MLPerf Inference v5.1 introduced a DeepSeek-R1 reasoning benchmark on which the newer GB300 NVL72 (Blackwell Ultra) delivered 45% higher throughput in the offline scenario than the GB200 NVL72 (Blackwell) system it superseded (Source: blogs.nvidia.com). Blackwell Ultra's architectural gains stem from 1.5x more NVFP4 compute and 2x more attention-layer acceleration than standard Blackwell, plus up to 288GB of HBM3e memory per GPU (Source: blogs.nvidia.com). NVIDIA also reported that disaggregated serving, splitting the context (prompt processing) and generation (token output) phases of inference across different GPU pools, delivered a 47% increase in performance per GPU on the Llama 3.1 405B Interactive benchmark for GB200 NVL72 compared with running the same workload on a single 8-GPU DGX B200 server using traditional, non-disaggregated serving (Source: blogs.nvidia.com). NVIDIA frames inference throughput as a direct economic lever, noting that higher throughput "increasing revenue, driving down total cost of ownership (TCO) and enhancing the system's overall productivity," which is why per-GPU inference records carry as much weight in NVIDIA's benchmark messaging as raw training speedups (Source: blogs.nvidia.com).

Third-party benchmarking outside MLCommons' formal suite tells a broadly consistent story with more workload-specific nuance. Machine learning infrastructure vendor Lightly AI ran independent computer vision training benchmarks comparing early-access B200 hardware against H100 in a European data center, reporting "up to 57% higher training throughput with the B200 compared to the H100 on the specific CV tasks we tested," alongside an estimate that self-hosted B200 capacity could be 6 to 30 times cheaper per GPU-hour than typical cloud H100 pricing, depending heavily on utilization, energy costs, and amortization assumptions (Source: www.reddit.com). The team specified that "all tests were conducted on our own hardware cluster hosted at GreenMountain, a data center running on 100% renewable energy," a detail worth noting since self-hosted, renewable-powered infrastructure changes the cost baseline compared with the on-demand cloud pricing quoted elsewhere in this report (Source: www.reddit.com). NVIDIA's own H100-to-B200 comparison suggests a more conservative but still substantial gain, stating B200's FP4-enabled Tensor Cores deliver "approximately 4x the training throughput of the H100 on transformer models," with FP4 inference roughly 3x faster than H100's FP8 path (Source: www.runpod.io). The gap between the 57% figure (a specific computer vision workload) and the 2.2x to 4x figures (LLM training and inference workloads) illustrates that Blackwell's architectural gains are workload-dependent, with the largest advantages appearing on transformer-based LLM training and inference rather than convolutional vision tasks.

Not every independent observer reads NVIDIA's official 2.2x MLPerf gain as impressive on its own terms. In a Reddit thread discussing the original MLPerf Training v4.1 result that preceded the 2.6x figure reported in v5.0, one commenter noted that because B200 is "roughly 2 H100 sized dies on a CoWoS-L package," a 2.2x gain represents "about a 10% improvement over the H100 in terms of die space," a more skeptical framing than NVIDIA's headline multiplier implies (Source: www.reddit.com). Another commenter in the same thread pushed back, arguing that multi-chip module (MCM) packaging "inherently decreases perf/area due to the interface, and scaling is never linear to begin with," so "getting over 2x is indicative of significant architecture improvements under the hood" rather than simply doubling the silicon (Source: www.reddit.com). A third commenter added a transistor-count data point worth noting for buyers comparing generations purely by silicon budget: B200 carries roughly 205 billion transistors, about a 25% increase over the 80 billion in the H200's underlying die, meaning MLPerf's headline multiplier already outpaces the raw transistor-count increase and is not simply explained by NVIDIA adding more silicon (Source: www.reddit.com). This exchange is a useful reminder that headline MLPerf multipliers should be read alongside die area, transistor count, and power draw, all of which this report's Feature Comparison table above provides, rather than treated as a pure architecture-efficiency metric in isolation.

Data Analysis and Evidence

NVIDIA's own financial disclosures provide the clearest quantitative signal of how demand has shifted across this GPU lineup. In the first quarter of fiscal 2027 (the quarter ended April 26, 2026), NVIDIA reported record total revenue of $81.6 billion, up 85% from a year earlier, with Data Center segment revenue reaching a record $75.2 billion, up 92% year over year and up 21% sequentially from the prior quarter (Source: nvidianews.nvidia.com). Under NVIDIA's disclosed sub-segment breakout, Data Center compute revenue (which includes GPU sales) reached $60.4 billion for the quarter, up 77% year over year, while Data Center networking revenue reached $14.8 billion, up 199% year over year, reflecting the growing share of revenue tied to NVLink, InfiniBand, and Ethernet fabric that ships alongside rack-scale systems like GB200 NVL72 (Source: nvidianews.nvidia.com). GAAP gross margin for the quarter was 74.9%, indicating the company retains substantial pricing power even as generational competition intensifies (Source: nvidianews.nvidia.com). On the earnings call, NVIDIA CFO Colette Kress told analysts that hyperscalers made up more than half of all Data Center revenue, reaching $38 billion and increasing 12% quarter-over-quarter, while the remaining $37 billion came from a newly disclosed segment NVIDIA calls ACIE (AI Clouds, Industrial and Enterprise), whose revenue "more than tripled year over year" (Source: www.cnbc.com). Kress also disclosed a data point directly relevant to GPU buyers: "the price of renting an H100 has risen 20% year to date, and A100 cloud pricing is up nearly 15%," adding that customers are generating profitable revenue "beyond the depreciable life of their GPUs," a signal that older-generation capacity is tightening rather than becoming obsolete as newer Blackwell hardware ships (Source: www.cnbc.com). NVIDIA CEO Jensen Huang described the quarter as "extraordinary," telling analysts that "demand has gone parabolic" because "agentic AI has arrived" (Source: www.cnbc.com). Huang further previewed the Rubin transition covered in the Implications section below, stating on the call that Nvidia expects Vera Rubin to be "even more successful than Grace Blackwell" and that the Vera Rubin system, comprising 72 Rubin GPUs and 36 Vera CPUs across 1.3 million components, delivers 10 times more performance per watt than its Blackwell predecessor (Source: www.cnbc.com). NVIDIA's own 10-Q filing acknowledged a competitive dynamic buyers should weigh against any single-vendor GPU roadmap: "some of our customers are developing their own ASICs and other products, including designs optimized for certain workloads that may not require all of the features and functionality our data center systems provide," referring to the custom AI chips hyperscalers like Google are increasingly deploying alongside, or instead of, NVIDIA GPUs (Source: www.cnbc.com).

AWS's reserved-capacity "Capacity Blocks for ML" pricing offers a cleaner apples-to-apples view of Blackwell-generation costs than on-demand rates alone. AWS lists an 8-GPU p6-b200.48xlarge instance at an effective $98.84 per hour total, or $12.355 per GPU-hour, in US East (Ohio), and an 8-GPU p6-b300.48xlarge (Blackwell Ultra) instance at $112.32 per hour total, or $14.04 per GPU-hour, in US West (Oregon), both materially above the on-demand HGX B200 rates quoted by CoreWeave and Lambda above because Capacity Blocks guarantee dedicated reserved access rather than shared on-demand availability (Source: aws.amazon.com). At the rack scale, AWS prices a reserved 72-GPU GB200 UltraServer at $761.904 per hour total, or $10.582 per GPU-hour, broadly consistent with CoreWeave's $42.00 GB200 NVL72 on-demand rate once dedicated-capacity and networking guarantees are factored in (Source: aws.amazon.com). For comparison, AWS's reserved 8-GPU H100 capacity (p5.48xlarge) runs $41.528 per hour total, or $5.191 per GPU-hour, in most US regions, confirming that reserved Blackwell capacity commands roughly a 2x to 2.4x premium per GPU-hour over reserved Hopper capacity on the same cloud (Source: aws.amazon.com). The Blackwell Ultra premium widens further still: AWS's reserved 8-GPU B300 instance (p6-b300.48xlarge) prices at $112.32 per hour total, or $14.04 per GPU-hour, in US West (Oregon), roughly 13% above the equivalent B200 reserved rate and nearly 3x the reserved H100 rate (Source: aws.amazon.com). AWS also still lists legacy P4 capacity for buyers with older workloads: an 8-GPU p4d.24xlarge instance built on A100 GPUs reserves at $11.80 per hour total, or $1.475 per GPU-hour, in US East (N. Virginia), roughly 3.5x cheaper per GPU-hour than reserved H100 capacity and illustrating how far down the price curve A100 has settled four years after launch (Source: aws.amazon.com).

Cloud rental pricing data assembled from multiple providers for this report (July 2026) is summarized in *Table 2* below, illustrating both the absolute price spread and the price-per-generation step. The data was gathered directly from each provider's public pricing page rather than from secondary aggregation, except where noted.

PROVIDER	A100 (PER GPU-HR)	H100 (PER GPU-HR, 8-GPU NODE)	H200 (PER GPU-HR, 8-GPU NODE)	B200 (PER GPU-HR)
CoreWeave (on demand, see *Table 1* for sourcing)	$21.60 (single GPU)	$6.16 ($49.24/8)	$6.31 ($50.44/8)	$8.60 ($68.80/8)
Lambda (self-serve on demand) (Source: lambda.ai)	$2.79 (80GB)	$3.99	not listed standalone	$6.69 to $6.99
Runpod (on demand, see B200 discussion above)	not directly listed above	$2.69 (implied, per B200 guide)	not listed	$4.99
Market range (68 providers, getdeploying.com) (Source: getdeploying.com)	$0.13 to $5.04	$0.64 to $14.90	$1.00 to $13.78	$2.69 to $16.11

This data shows two consistent patterns. First, small independent clouds like Runpod consistently underprice large hyperscale-oriented providers like CoreWeave, sometimes by more than 2x for the same GPU generation, reflecting differences in SLA (service-level agreement) guarantees, network fabric quality, and contract minimums rather than hardware differences. Second, the generational price premium compresses as workload throughput scales: B200 costs roughly 1.4x more per GPU-hour than H100 on CoreWeave's on-demand pricing, but delivers 2.2x to 4x more throughput on LLM training benchmarks, meaning effective cost-per-token or cost-per-training-step can favor the newer generation despite its higher sticker price. Spot and reserved pricing further complicates comparisons: CoreWeave's HGX B200 spot rate of roughly $34.11 per hour is about half its on-demand rate, and Lambda's HGX B200 cluster pricing drops from $9.86 to $8.87 per GPU-hour as commitment size scales from 16 to 256-plus GPUs (Source: lambda.ai). Lambda's HGX H100 cluster pricing follows the same volume curve, falling from $6.16 per GPU-hour at 16 GPUs to $5.54 per GPU-hour at 256 GPUs, confirming that committed cluster scale, not just GPU generation, is a material lever on effective price (Source: lambda.ai). CoreWeave's A100 spot price of roughly $9.65 per hour, versus $21.60 on demand (see *Table 1*), further shows that spot markets can roughly halve GPU costs across every generation surveyed in this report, not just Blackwell. The market also already prices in the next Blackwell refresh: getdeploying.com lists the newer B300 GPU, based on Blackwell Ultra with up to 288GB of HBM3e, at $2.50 to $18.00 per hour across 13-plus providers, a similar spread to B200 despite its additional memory (Source: getdeploying.com).

Case Studies and Real-World Examples

xAI Colossus: Fastest Large-Scale Build and Multi-Generation Fleet

xAI's Colossus supercomputer in Memphis, Tennessee is the clearest public example of how quickly NVIDIA GPU deployments can scale and how operators mix generations within one facility. The initial build deployed 100,000 H100 GPUs in a former Electrolux appliance factory in just 122 days, using HGX servers with eight GPUs each in Supermicro liquid-cooled racks housing 64 GPUs per rack across 1,500 total racks (Source: introl.com). xAI then doubled the cluster to 200,000 GPUs in 92 additional days (Source: introl.com). By December 2025, the operational fleet comprised 150,000 H100, 50,000 H200, and 30,000 GB200 GPUs, described as "the largest fully operational, single-coherent AI training cluster in the world," with xAI targeting expansion to 1 million GPUs (Source: introl.com). The cluster draws approximately 250 megawatts (MW) of power, up from an initial 150MW configuration, supplied by a mix of 35 gas turbines capable of 420MW and 208 Tesla Megapack battery systems, and achieves total memory bandwidth of 194 petabytes per second (PB/s) with more than 1 exabyte of storage (Source: introl.com). The Colossus 2 expansion at an adjacent site targets at least 110,000 additional GB200 GPUs (Source: introl.com).

Meta's Llama 3 Training Clusters: Dual-Fabric H100 Deployment

Meta built two 24,576-GPU H100 clusters to train the Llama 3 model family, using different network fabrics for each: one cluster runs a 400 gigabit (G) RoCEv2 backend on Arista 7800 switches, while the other runs 400G NDR (Next Data Rate) InfiniBand on NVIDIA Quantum-2 switches (Source: www.glennklockwood.com). Each cluster comprises 3,072 Grand Teton compute nodes across 1,536 racks, with each rack-level "pod" containing 3,072 GPUs as a nonblocking fabric domain (Source: www.glennklockwood.com). This dual-fabric design let Meta directly compare Ethernet-based RoCEv2 against InfiniBand at production LLM training scale, and documentation of the design notes that Meta deliberately over-provisioned RoCEv2

injection bandwidth "to work around the congestion that arises from RoCEv2's poor handling of it" at the network's aggregation layer (Source: www.glennklockwood.com). Each node in these clusters pairs one HGX baseboard of 8 H100 GPUs with full NVLink interconnectivity, and two such nodes fit in a single rack, which is why each 24,576-GPU cluster spans exactly 1,536 racks (Source: www.glennklockwood.com).

Microsoft Azure GB200 NVL72: General Availability and Throughput Gains

Microsoft Azure's ND GB200 v6 virtual machine series reached general availability powered by NVIDIA GB200 NVL72, with Azure describing itself as one of the first cloud service providers to launch a 4,000-GPU GB200 Grace Blackwell powered supercomputing cluster for training state-of-the-art models. Each rack-scale GB200 NVL72 system delivers up to 1.4 exaFLOPS of FP4 Tensor Core throughput, 13.5TB of shared high-bandwidth memory, and 130 TB/s of cross-sectional NVLink bandwidth (Source: techcommunity.microsoft.com). NVIDIA's Ian Buck, Vice President of Hyperscale and HPC, described the collaboration's significance: "The NVIDIA GB200 NVL72, with its unparalleled performance and connectivity, tackles the most complex AI workloads, enabling businesses to innovate faster and more securely" (Source: techcommunity.microsoft.com).

Microsoft Azure GB300 NVL72 for OpenAI: First Production Blackwell Ultra Cluster

Extending the Azure-NVIDIA partnership, Microsoft announced what NVIDIA called "the industry's first supercomputing-scale production cluster of NVIDIA GB300 NVL72 systems, purpose-built for OpenAI's most demanding AI inference workloads," comprising over 4,600 (specifically 4,608) Blackwell Ultra GPUs connected via NVIDIA Quantum-X800 InfiniBand (Source: blogs.nvidia.com). Each rack provides 37TB of fast memory and 1.44 exaflops of FP4 Tensor Core performance per virtual machine (VM), with the fifth-generation NVLink Switch fabric delivering 130 TB/s of all-to-all bandwidth within the rack and NVIDIA Quantum-X800 providing 800 gigabits per second (Gb/s) of scale-out bandwidth per GPU across the full 4,608-GPU cluster (Source: blogs.nvidia.com). Microsoft Azure AI Infrastructure corporate vice president Nidhi Chappell said the achievement "reflects Microsoft Azure and NVIDIA's shared commitment to optimize all parts of the modern AI data center," adding that the collaboration "helps ensure customers like OpenAI can deploy next-generation infrastructure at unprecedented scale and speed" (Source: blogs.nvidia.com). NVIDIA noted that as Azure scales toward its stated goal of deploying hundreds of thousands of Blackwell Ultra GPUs, this 4,608-GPU cluster represents an early, not final, milestone in that build-out (Source: blogs.nvidia.com).

Cohere and Anthropic: Cloud-Native H100 Adoption

Beyond the largest hyperscaler-built superclusters, AWS states that P5, P5e, and P5en instances "help you accelerate your time to solution by up to 4x compared to previous-generation GPU-based EC2 instances, and reduce cost to train ML models by up to 40%" (Source: aws.amazon.com), a claim borne out in the customer testimonials below illustrating H100 adoption among AI model developers renting rather than owning capacity. Anthropic co-founder Tom Brown stated the company was "using Amazon EC2 P4 instances extensively today, and we are excited about the launch of P5 instances," expecting "substantial price-performance benefits over P4d instances" at the scale required for next-generation LLMs, adding that Anthropic needs to "distribute them efficiently across large clusters of GPUs" to develop and train its foundational models (Source: aws.amazon.com). Cohere CEO Aidan Gomez similarly noted that "NVIDIA H100-powered Amazon EC2 P5 instances will unleash the ability of businesses to create, grow, and scale faster with its computing power combined with Cohere's state-of-the-art LLM and generative AI capabilities" (Source: aws.amazon.com). Insurance and risk-analytics firm AON offers a non-frontier-lab example of the same instance family: Global Head of Life Solutions Van Beach reported that "the ability to use a single H100 GPU instance (p5.4xlarge) means we're not only saving time but also optimizing our computational resources," cutting actuarial simulation workloads that "used to take days" down to hours. These testimonials illustrate that even well-funded frontier labs frequently rent H100 capacity through hyperscaler cloud instances rather than building proprietary data centers, a contrast to xAI's and Meta's owned-infrastructure approach documented above, and that H100-class compute has diffused well beyond AI-native companies into mainstream financial services use cases. AWS notes these P5 family instances are deployed inside the same EC2 UltraClusters discussed above, delivering up to 20 exaflops of aggregate compute, roughly the same order of magnitude as xAI's and Meta's self-built superclusters documented in the case studies below.

Implications and Future Directions

The near-term trajectory for NVIDIA's data center GPU lineup is already public. NVIDIA announced the **Rubin** platform on January 5, 2026, as the successor to Blackwell, comprising six new chips: the Vera CPU, Rubin GPU, sixth-generation NVLink 6 Switch, ConnectX-9 SuperNIC, BlueField-4 DPU (data processing unit), and Spectrum-6 Ethernet switch (Source: nvidianews.nvidia.com). The Rubin GPU is specified to deliver 50 petaFLOPS of NVFP4 compute per GPU for inference, with each GPU offering 3.6 TB/s of NVLink bandwidth and the full Vera Rubin NVL72 rack providing 260 TB/s, which NVIDIA describes as "more bandwidth than the entire internet" (Source: nvidianews.nvidia.com). NVIDIA states the Rubin platform trains

mixture-of-experts (MoE) models with 4x fewer GPUs than Blackwell and cuts inference cost per token by up to 10x (Source: nvidianews.nvidia.com). Rubin-based products are expected to reach partners in the second half of 2026, with AWS, Google Cloud, Microsoft, and Oracle Cloud Infrastructure (OCI) among the first cloud providers slated to deploy Vera Rubin instances, alongside NVIDIA Cloud Partners CoreWeave, Lambda, Nebius, and Nscale. Microsoft has already committed to the platform at the infrastructure level, with NVIDIA indicating that Microsoft's next-generation Fairwater AI superfactories will scale to "hundreds of thousands" of Vera Rubin Superchips, an order of magnitude beyond the 4,608-GPU GB300 NVL72 cluster Microsoft and NVIDIA delivered for OpenAI in the Blackwell generation (Source: nvidianews.nvidia.com). Frontier labs have framed the significance of successive hardware generations in terms of model capability rather than raw specifications: intelligence scales with available compute, so each new GPU generation directly expands what model developers can attempt next.

For buyers evaluating the current lineup, several practical implications follow from the data above. First, memory capacity, not raw compute, is often the binding constraint for LLM inference and fine-tuning; the jump from H100's 80GB to H200's 141GB or B200's 180GB can eliminate multi-GPU sharding for models in the 70B-to-180B parameter range, which often matters more for total cost of ownership than a generation's raw FLOPS improvement. Second, NVLink domain size increasingly determines training efficiency at scale, since GB200 NVL72's 72-GPU NVLink domain measurably reduced Llama 3.1 405B pretraining time in MLPerf submissions relative to Hopper's smaller 8-GPU NVLink islands (Source: developer.nvidia.com). Third, rapid depreciation risk is real and growing: as one community observer noted regarding used H100 valuations amid Blackwell's ramp, hardware that commanded premium pricing eighteen months prior has already seen substantial resale value erosion, a dynamic likely to repeat as B300 and Rubin displace B200 over the next 12 to 18 months (Source: www.reddit.com). Buyers with multi-year commitments should weight this depreciation curve against the near-term throughput advantages of the newest hardware.

Export Controls Reshape Regional Availability

GPU choice is no longer purely a technical or budgetary decision: US export policy now directly determines which chips are legally available in which regions, and that policy has shifted materially within the past year. On December 8, 2025, US President Donald Trump announced that the United States would allow NVIDIA's H200 processors, described in the announcement as the company's "second-best artificial intelligence chips," to be exported to China subject to a 25% fee collected as an import tax, reversing years of tightening controls under both the Biden and earlier Trump administrations (Source: www.reuters.com). Blackwell-generation chips remain excluded from the arrangement: Trump wrote that "NVIDIA's U.S. Customers are already moving forward with their incredible, highly advanced Blackwell chips, and soon, Rubin, neither of which are part of this deal," meaning Chinese buyers remain capped at Hopper-generation H200 hardware even after the policy easing (Source: www.reuters.com). According to a report by the think tank Institute for Progress cited in the same coverage, the H200 is almost six times as powerful as the H20, the downgraded chip NVIDIA had been selling into China, while Blackwell chips now in use by US firms are roughly 1.5 times faster than H200 for training and five times faster for inference, illustrating the performance gap export policy currently preserves between US and Chinese AI infrastructure (Source: www.reuters.com). The decision drew immediate political pushback: Republican Representative John Moolenaar, who chairs the House China Select Committee, warned that "China will rip off its technology, mass-produce it themselves and seek to end Nvidia as a competitor," while several Democratic senators called the move a "colossal economic and national security failure" (Source: www.reuters.com). For multinational buyers, this means the GPU comparison in this report applies fully only within US-aligned markets; procurement teams sourcing capacity for China-based operations must plan around H200-or-below availability and the associated 25% fee rather than assuming access to B200 or GB200 NVL72-class hardware. Even the H200 approval carries conditions: the White House said exports would occur only "under conditions that allow for continued strong National Security," and administration officials described the move as a deliberate compromise between denying China any advanced US chips (which officials feared would benefit Huawei's competing AI chip business) and granting access to Blackwell-class hardware, which Trump declined to allow (Source: www.reuters.com).

Frequently Asked Questions (FAQs)

What is the difference between H100, H200, and B200? The H100 and H200 share the same Hopper compute die but differ in memory: H100 has 80GB of HBM3 while H200 has 141GB of HBM3e at 4.8 TB/s bandwidth (see the H100 and H200 sections above). The B200 is an entirely different Blackwell-architecture dual-die GPU with 180GB of HBM3e, native FP4 precision, and roughly 4x the training throughput of H100 on transformer models (see the B200 and GB200 NVL72 section above).

What are the NVIDIA GB200 specs? A GB200 NVL72 rack connects 36 Grace CPUs and 72 Blackwell GPUs, delivering 13.4TB of HBM3e memory at 576 TB/s aggregate bandwidth, 1,440 petaFLOPS of sparse NVFP4 compute, and 130 TB/s of NVLink bandwidth across the full rack (see *Table 1* and the B200/GB200 NVL72 section above).

How does NVIDIA A100 compare to H100? H100 delivers up to 4x faster GPT-3 175B training and up to 30x higher inference performance on the largest models compared with A100 (see the A100 and H100 sections above), driven by fourth-generation Tensor Cores, a dedicated Transformer Engine with FP8 precision, and roughly 65% higher memory bandwidth.

What is the difference between Blackwell and Hopper? Blackwell introduces a dual-die package with 2.5x the transistors of Hopper, a second-generation Transformer Engine, native FP4 precision, and fifth-generation NVLink at double the per-GPU bandwidth of Hopper's fourth-generation NVLink (see the Introduction and Background and B200/GB200 sections above).

What is the difference between B200 and GB200? B200 is the individual GPU (180GB HBM3e, up to 20 petaFLOPS FP4); GB200 refers to the Grace Blackwell Superchip that pairs one Grace CPU with two Blackwell GPUs, or to the GB200 NVL72 rack that scales this to 36 CPUs and 72 GPUs in a single NVLink domain (Source: www.nvidia.com).

What is the best NVIDIA GPU for LLM training? For the largest frontier-scale training runs, GB200 NVL72 currently delivers the best measured throughput, posting 2.2x to 2.6x higher per-GPU performance than Hopper in MLPerf Training v5.0 on Llama 3.1 405B pretraining (see the Performance and Benchmarks section above). For teams with tighter budgets or models under roughly 70B parameters, H100 or H200 clusters remain cost-competitive given their deeper cloud availability and lower spot pricing.

What is next after Blackwell in NVIDIA's roadmap? NVIDIA's Rubin platform, announced January 5, 2026, succeeds Blackwell with a Rubin GPU delivering 50 petaFLOPS of NVFP4 inference compute per GPU and sixth-generation NVLink; Rubin-based products are expected from partners in the second half of 2026 (Source: nvidianews.nvidia.com).

How much does NVIDIA data center GPU cloud rental cost per hour in 2026? Prices vary widely by provider and commitment length: A100 ranges from roughly $0.13 to $21.60 per GPU-hour, H100 from about $2.69 to $14.90, H200 from about $1.00 to $13.78, and B200 from about $2.69 to $16.11 per GPU-hour (see *Table 2* above), with large hyperscale-grade providers like CoreWeave typically priced above small independent clouds like Runpod for the same silicon.

Is the NVIDIA A100 still worth buying or renting in 2026? Yes, for cost-constrained workloads. The A100's chief advantage four generations later is its market depth and low floor price, with getdeploying.com tracking A100 availability across more than 36 providers at rates starting near $0.13 per hour (see the A100 Adoption section above), making it a defensible choice for fine-tuning, smaller-model training, and inference workloads that do not require FP8 or FP4 Tensor Cores.

Conclusion

Across the five NVIDIA data center GPUs and the GB200 NVL72 rack system compared in this report, the practical decision framework is consistent: match memory capacity to model size first, then weigh NVLink domain size and cluster topology, and only then optimize for hourly price. The A100 remains a defensible choice for inference and smaller training jobs at $0.13 to $5.04 per hour across the wider market, while H100 continues to offer the deepest ecosystem maturity and spot-market liquidity at roughly $6 per GPU-hour on major clouds. H200's memory upgrade is worth the modest price premium for any workload constrained by 80GB, and B200's roughly 2x to 4x throughput gains on LLM workloads increasingly justify its 1.4x to 1.9x price premium over Hopper, especially for teams that can fully utilize its FP4 precision path. GB200 NVL72 and the newer GB300 NVL72 represent the current performance ceiling for both training and inference at rack scale, backed by independently verified MLPerf results and multiple hyperscaler production deployments documented above. With NVIDIA's Rubin platform confirmed for second-half 2026 partner availability and promising up to 10x lower inference cost per token than Blackwell, buyers evaluating multi-year GPU commitments in mid-2026 should weight near-term Blackwell throughput advantages against the reality that today's flagship hardware, much like the A100 and H100 before it, will face a compressed depreciation curve once the next generation ships at scale. NVIDIA's Data Center segment revenue of $75.2 billion in a single fiscal quarter confirms that demand across this entire generational spread, not just the newest chip, remains strong enough to support the aggressive annual product cadence NVIDIA has now committed to, which in turn means buyers should plan procurement and depreciation schedules around a roughly twelve-month major architecture refresh rather than the two-to-three-year cycle common earlier in the last decade.

Tags: nvidia data center gpu comparison, h100 vs h200 vs b200, nvidia gb200 nvl72, nvidia a100 vs h100, blackwell vs hopper, nvidia b200, nvidia h200 specs, gpu for llm training

---