Blackwell vs Hopper: NVIDIA GPU Architecture Comparison 2026

Published July 10, 2026 36 min read

Executive Summary

NVIDIA's **Hopper** and **Blackwell** architectures anchor two successive generations of data center graphics processing units (GPUs) that have defined large-scale artificial intelligence (AI) infrastructure since 2022. **Hopper**, announced March 22, 2022, and led by the **H100 Tensor Core GPU**, packs 80 billion transistors on a custom TSMC 4N process and delivers up to 3,958 teraFLOPS (TFLOPS) of FP8 performance with 80GB of HBM3 memory at 3.35 terabytes per second (TB/s), detailed further in Table 1 below. **Blackwell**, unveiled March 18, 2024 at NVIDIA's GTC conference and led by the **B200 GPU** and **GB200 Grace Blackwell Superchip**, packs 208 billion transistors across two reticle-limited dies on a custom TSMC 4NP process, connected by a 10 TB/second chip-to-chip link (detailed in the Blackwell Architecture section below). Blackwell GPUs pair up to 192GB of HBM3e memory (versus 80GB on H100 SXM) with a new 4-bit floating point (FP4) precision mode that roughly doubles peak inference throughput over FP8 on the same silicon (detailed further in the Blackwell Architecture section below, with figures corroborated independently by GPU integrator Exxact) (Source: exxactcorp.com).

The two architectures are not simply "old versus new": Hopper remains in high-volume production and continues to receive software-driven performance gains, while Blackwell targets the largest generative AI training and inference clusters. On MLPerf Inference v5.0, an industry-standard benchmark suite operated by MLCommons, an eight-GPU **B200** system delivered 3x higher throughput than an eight-GPU **H200** system on the Llama 2 70B benchmark, and NVIDIA's rack-scale **GB200 NVL72** system delivered up to 3.4x higher per-GPU performance than an eight-GPU H200 system on the newly introduced Llama 3.1 405B benchmark (Source: developer.nvidia.com). At the system level, the liquid-cooled GB200 NVL72 rack, which links 72 Blackwell GPUs and 36 Grace CPUs over a 130 TB/s NVLink domain, delivers up to 30x faster real-time large language model (LLM) inference than the same number of H100 GPUs and up to 25x lower cost and energy consumption per token (detailed in the Data Analysis and Evidence section below).

These gains come with a substantially higher power budget: a single Blackwell GPU in the DGX B200 draws up to 1,000 watts (W), compared with up to 700W for an H100 SXM module, and a full DGX B200 chassis draws up to 14.3 kilowatts (kW) (Source: docs.nvidia.com), a gap independently corroborated by cloud provider Modal, which likewise reports 700W for H100 versus 1,000W for B200 (Source: modal.com). Cloud rental prices reflect this generational split: as of mid-2026, H100 GPUs rent for roughly $2 to $5 per GPU-hour depending on provider and form factor, while B200 GPUs

typically range from about $3.75 to $8 per GPU-hour (Source: coreweave.com) (Source: getdeploying.com). NVIDIA's own financial results underscore how quickly Blackwell has been absorbed into the market: Data Center revenue, which includes both Hopper and Blackwell product lines, reached a record $75.2 billion in the first quarter of fiscal 2027 (the quarter ended April 26, 2026), up 92% year over year, following full-year fiscal 2026 Data Center revenue of $193.7 billion (Source: nvidianews.nvidia.com) (detailed further in the Data Analysis and Evidence section below).

For buyers, the practical answer to "Blackwell vs Hopper" depends on workload and deployment stage. Organizations training frontier-scale models above roughly 100 billion parameters, or running latency-sensitive inference on trillion-parameter mixture-of-experts (MoE) models, gain the most from Blackwell's larger memory pool, FP4 support, and NVLink 5 bandwidth. Organizations running established, well-optimized inference pipelines on models in the tens of billions of parameters, or operating under power, cooling, or capital constraints, often find Hopper GPUs, whose per-GPU throughput has itself grown substantially over the past year through software optimization alone (detailed in the Hopper Architecture section below), remain a cost-effective and widely available choice. The remainder of this report details architecture, specifications, benchmarks, deployments, and total cost of ownership considerations behind that choice.

Introduction and Background

NVIDIA's data center GPU roadmap advances in roughly two-year cycles, each named after a scientist or mathematician. **Hopper**, named for computer scientist Grace Hopper, succeeded the 2020-era Ampere architecture and was announced at NVIDIA's GTC conference on March 22, 2022 (Source: nvidianews.nvidia.com). Its flagship product, the H100 Tensor Core GPU, entered full production in September 2022 and became the workhorse behind the initial wave of large language model (LLM) training runs, including systems used by OpenAI, Meta, and numerous cloud providers (Source: nvidianews.nvidia.com). **Blackwell**, named for mathematician and statistician David Blackwell, was announced two years later on March 18, 2024, and began broad availability through 2024 and 2025, as detailed in the Blackwell Architecture section below.

This comparison matters because, as of July 2026, both architectures remain commercially active and available for purchase or rental. Hopper-generation H100 and H200 GPUs are widely deployed and continue to receive throughput-boosting software updates from NVIDIA, while Blackwell-generation B100, B200, GB200, and their "Ultra" and NVL72 rack-scale variants have become the default choice for new large-scale AI infrastructure buildouts. Understanding the concrete differences, in transistor count, memory, precision support, interconnect bandwidth, power draw, and price, helps organizations size infrastructure decisions correctly rather than defaulting to "buy the newest chip."

The stakes are unusually large. NVIDIA's Data Center segment, which sells both architectures, generated record revenue of $75.2 billion in a single quarter (Q1 fiscal 2027, ended April 26, 2026), up 92% from a year earlier, and full fiscal year 2026 Data Center revenue reached $193.7 billion, up 68% year over year (detailed further in the Data Analysis and Evidence section below). That scale reflects a broader shift NVIDIA CEO Jensen Huang has described as "AI factories," data centers whose primary output is tokens rather than traditional compute cycles (Source: nvidianews.nvidia.com). Choosing the wrong architecture for a given workload, over-provisioning Blackwell for light inference workloads, or under-provisioning Hopper for frontier training, translates directly into wasted capital or missed performance targets at a scale of tens of millions of dollars per cluster.

This report examines Hopper and Blackwell side by side: their core silicon and memory architecture, their real-world adoption, their strengths and limitations, a feature-by-feature comparison matrix, benchmark results from MLCommons' MLPerf suite, quantitative market and pricing data, named real-world deployments, and forward-looking implications including the transition to NVIDIA's next-generation Vera Rubin platform.

Both architectures ship in multiple product forms rather than as a single GPU. Hopper spans the original H100 (available in SXM and PCIe form factors), the memory-enhanced H200, and the multi-GPU H100 NVL card designed for large language model inference in power-constrained environments, with cloud GPU provider Voltage Park noting the H100 is "well-suited for fine-tuning models in the 7B-70B parameter range" (Source: voltagepark.com). Blackwell spans the air-cooled, drop-in-compatible B100; the higher-throughput B200; the two-GPU-plus-CPU GB200 Grace Blackwell Superchip; and rack-scale systems such as the GB200 NVL72 and its Blackwell Ultra-based successor, the GB300 NVL72.Readers evaluating a specific purchase or rental decision should therefore treat "Hopper" and "Blackwell" as architecture families rather than single, interchangeable products, since specifications, power draw, and pricing vary considerably within each family.

Hopper Architecture

Capabilities

The **Hopper architecture** is built around fourth-generation Tensor Cores and a first-generation **Transformer Engine** that automatically manages FP8 and FP16 mixed-precision computation for transformer-based neural networks (Source: nvidianews.nvidia.com). The flagship **H100 GPU** is fabricated with 80 billion transistors on a custom TSMC 4N process and was, at launch, described by NVIDIA as capable of nearly 5 terabytes per second (TB/s) of external connectivity (Source: nvidianews.nvidia.com). The H100 SXM variant delivers 67 TFLOPS of FP32, 989 TFLOPS of TF32 Tensor Core

throughput (with sparsity), and 3,958 TFLOPS of FP8 Tensor Core throughput, alongside 80GB of HBM3 memory at 3.35 TB/s bandwidth (see Table 1 below). Fourth-generation **NVLink** provides 900GB/s of GPU-to-GPU bandwidth, and an external NVLink Switch can connect up to 256 H100 GPUs at 9x the bandwidth of the prior generation via NVIDIA Quantum-2 InfiniBand networking (Source: nvidianews.nvidia.com). Hopper also introduced second-generation secure Multi-Instance GPU (MIG) partitioning, confidential computing for protecting models and data in use, and DPX instructions that accelerate dynamic-programming algorithms used in genomics and route optimization by up to 40x versus CPUs (Source: nvidianews.nvidia.com). A mid-generation refresh, the **H200 GPU**, retains the Hopper compute architecture but adds 141GB of HBM3e memory at 4.8 TB/s, nearly double the memory capacity of the original H100 and 1.4x its bandwidth (Source: nvidia.com).

Adoption

Hopper reached full production in September 2022 and has since been adopted by every major cloud provider, including Alibaba Cloud, Amazon Web Services (AWS), Baidu AI Cloud, Google Cloud, Microsoft Azure, Oracle Cloud, and Tencent Cloud, alongside system makers such as Dell, HPE, Lenovo, and Supermicro (Source: nvidianews.nvidia.com). Three years after launch, Hopper GPUs remain in active large-scale use: xAI's Colossus supercomputer in Memphis, Tennessee combines roughly 150,000 H100 and 50,000 H200 GPUs (alongside newer GB200 GPUs) in what the company describes as the world's largest single, coherent AI training cluster (Source: introl.com), and xAI's own site states the initial cluster reached 200,000 H100 GPUs in a single interconnected system (Source: x.ai) (Source: x.ai).

Strengths and Limitations

Hopper's principal strength is maturity: three-plus years of software optimization on TensorRT-LLM have increased H100 inference throughput on the Llama 2 70B benchmark by up to 1.5x over the past year alone, and cumulative improvement on the GPT-J benchmark since its introduction has reached 2.9x in the offline scenario and 3.8x in the server scenario, all on unchanged silicon, according to NVIDIA's MLPerf Inference v5.0 benchmark disclosures (Source: developer.nvidia.com). Hopper GPUs are also widely available secondhand and across cloud marketplaces, with mature driver, framework, and container support. Its principal limitations are the absence of native FP4 precision, a smaller maximum memory pool (80GB to 141GB per GPU versus up to 192GB or more for Blackwell), and a lower-bandwidth NVLink domain (900GB/s per GPU versus 1.8TB/s for Blackwell), all of which constrain how large a model can fit on, or be efficiently sharded across, a fixed number of GPUs (Source: nvidia.com).

Blackwell Architecture

Capabilities

The **Blackwell architecture** packs 208 billion transistors across two reticle-limited dies, manufactured on a custom-built TSMC 4NP process and joined by a 10 TB/s chip-to-chip interconnect that NVIDIA presents as a single, unified GPU (Source: nvidia.com). Its second-generation **Transformer Engine** adds micro-tensor scaling and new dynamic range management, enabling native 4-bit floating point (FP4) inference that doubles peak throughput relative to FP8 on the same hardware while meeting benchmark accuracy requirements (Source: nvidianews.nvidia.com) (Source: developer.nvidia.com). The single **B200 GPU** offers 192GB of HBM3e memory, and NVIDIA's DGX B200 system, which houses eight Blackwell GPUs, delivers 144 petaFLOPS (PFLOPS) of sparse FP4 Tensor Core performance and 72 PFLOPS of FP8 Tensor Core performance, alongside 1,440GB of total GPU memory at 64 TB/s aggregate HBM3e bandwidth (Source: nvidia.com). Fifth-generation **NVLink** delivers 1.8 TB/s of bidirectional bandwidth per GPU, scaling to as many as 576 GPUs in a single domain, a substantial jump from Hopper's 900GB/s NVLink 4 (Source: nvidianews.nvidia.com). Blackwell also introduces a dedicated decompression engine for data analytics, a Reliability, Availability, and Serviceability (RAS) engine for predictive fault management, and the industry's first TEE-I/O-capable confidential computing implementation, which secures data in transit across NVLink as well as at rest (Source: nvidia.com).

The rack-scale **GB200 NVL72** connects 36 Grace CPUs and 72 Blackwell GPUs in a liquid-cooled design, forming a single 72-GPU NVLink domain with 130 TB/s of GPU-to-GPU bandwidth, 13.4 terabytes (TB) of aggregate HBM3e memory at 576 TB/s, and up to 1,440 PFLOPS of sparse FP4 Tensor Core throughput (Source: nvidia.com) (Source: nvidia.com). NVIDIA states that a GB200 NVL72 rack acts as a single massive GPU with 1.4 exaflops of AI performance and 30TB of fast memory (Source: nvidianews.nvidia.com). A subsequent refresh, **GB300 NVL72**, integrates 72 "Blackwell Ultra" GPUs and delivers up to a 50x overall increase in AI factory output versus Hopper-based platforms, according to NVIDIA (Source: nvidia.com). Separately, NVIDIA also offers the air-cooled **B100** GPU as a drop-in-compatible option for existing HGX H100 server designs, operating at the same 700W per-GPU thermal design power (TDP) as H100 to ease the upgrade path for data centers not yet equipped for Blackwell's higher power density (Source: introl.com).

Adoption

Blackwell has been adopted across every major cloud provider and server manufacturer. At GTC 2025, Microsoft announced general availability of the Azure ND GB200 V6 virtual machine series, accelerated by GB200 NVL72 and NVIDIA Quantum InfiniBand networking, alongside existing Azure virtual machines built on H100 and H200 GPUs (Source: crn.com). NVIDIA's fourth-quarter fiscal 2026 results describe a "multiyear, multigenerational" partnership with Meta covering large-scale deployment of Blackwell and Rubin-generation GPUs, alongside continued rollout of GB300 NVL72 systems by cloud providers including Microsoft, CoreWeave, and Oracle Cloud Infrastructure for low-latency agentic AI workloads (Source: nvidianews.nvidia.com). Executives from Meta, Microsoft, and OpenAI publicly endorsed the platform at its March 2024 unveiling: Meta's Mark Zuckerberg said the company looked forward to using Blackwell "to help train our open-source Llama models," Microsoft's Satya Nadella cited bringing "the GB200 Grace Blackwell processor to our datacenters globally," and OpenAI's Sam Altman said "Blackwell offers massive performance leaps" (Source: nvidianews.nvidia.com) (Source: nvidianews.nvidia.com) (Source: nvidianews.nvidia.com).

Blackwell Ultra and the HGX Platform

NVIDIA has since extended Blackwell with an enhanced variant, **Blackwell Ultra**, packaged in the **HGX B300** platform. Compared with the original eight-GPU HGX B200, the eight-GPU HGX B300 keeps the same 1.8TB/s NVLink 5 per-GPU bandwidth and 14.4TB/s total NVLink bandwidth, but raises sparse FP4 Tensor Core throughput from 72 to 108 PFLOPS dense (144 PFLOPS shared sparse ceiling) and doubles total networking bandwidth to 1.6TB/s, while NVIDIA states its attention-mechanism performance is 2x that of the original Blackwell generation (Source: nvidia.com) (Source: nvidia.com). Total onboard memory across the eight-GPU HGX B300 baseboard rises to 2.1TB, versus 1.4TB for HGX B200 (Source: nvidia.com). NVIDIA has already begun previewing Blackwell's eventual successor, the Rubin-generation HGX Rubin NVL8, which it says will deliver 400 PFLOPS of NVFP4 inference compute, 176 TB/s of memory bandwidth (3x more than HGX B200), and 10x more token factory throughput than HGX B200 (Source: nvidia.com), underscoring how quickly the Blackwell generation itself will be superseded.

Strengths and Limitations

Blackwell's core strengths are memory capacity, interconnect bandwidth, and native low-precision inference, all of which matter disproportionately for the largest generative AI models. Its FP4 support alone doubles peak throughput per GPU relative to FP8 on comparable silicon (as detailed in the Blackwell Architecture section above), and NVLink 5's 1.8TB/s per-GPU bandwidth materially reduces the communication bottleneck that constrains distributed training and inference of trillion-parameter mixture-of-experts models (Source: nvidia.com). The trade-offs are power density, cooling requirements, and price. A DGX B200 chassis draws up to 14.3kW (see Table 1 above), and full GB200 NVL72 and GB300 NVL72 racks require liquid cooling rather than the air cooling sufficient for most Hopper deployments (Source: nvidia.com), a retrofit cost some data center operators must absorb before they can take advantage of the architecture at all.

Feature Comparison

Table 1 below summarizes the core specifications of the principal Hopper and Blackwell data center GPUs discussed in this report, drawn from NVIDIA's own product pages and datasheets.

SPECIFICATION	H100 SXM (HOPPER)	H200 SXM (HOPPER)	B200 (BLACKWELL, PER GPU IN DGX B200)	GB200 NVL72 (BLACKWELL, PER RACK)
Transistor count	80 billion (Source: nvidianews.nvidia.com)	80 billion (same die as H100)	208 billion per GPU package (2 dies) (Source: nvidia.com)	208 billion x 72 GPUs
Process node	TSMC 4N (Source: nvidianews.nvidia.com)	TSMC 4N	TSMC 4NP (Source: nvidia.com)	TSMC 4NP
GPU memory	80GB HBM3 (Source: nvidia.com) (Source: exxactcorp.com)	141GB HBM3e (Source: nvidia.com)	192GB HBM3e (1,440GB / 8 GPUs, DGX B200) (Source: nvidia.com)	13.4TB total (72 GPUs), consistent with the per-GPU figures above scaled across the rack
Memory bandwidth	3.35 TB/s (Source: nvidia.com)	4.8 TB/s (Source: nvidia.com)	8 GPU aggregate: 64 TB/s (Source: nvidia.com)	576 TB/s aggregate (see Blackwell Architecture section above)
FP8 Tensor Core (sparse)	3,958 TFLOPS (Source: nvidia.com)	3,958 TFLOPS (Source: nvidia.com)	72 PFLOPS / 8 = 9,000 TFLOPS per GPU (DGX B200) (Source: nvidia.com)	720 PFLOPS FP8/FP6 across 72 GPUs (see Blackwell Architecture section above)
FP4 Tensor Core (sparse)	Not supported natively	Not supported natively	144 PFLOPS / 8 = 18,000 TFLOPS per GPU (DGX B200) (Source: nvidia.com)	1,440 PFLOPS NVFP4 across 72 GPUs (see Blackwell Architecture section above)
NVLink generation and bandwidth	4th gen, 900GB/s per GPU (Source: nvidia.com)	4th gen, 900GB/s per GPU (Source: nvidia.com)	5th gen, 1.8TB/s per GPU (see Blackwell Architecture section above)	5th gen, 130TB/s domain-wide (see Blackwell Architecture section above)
Max thermal design power (TDP)	Up to 700W (configurable) (Source: nvidia.com) (Source: modal.com)	Same 700W-class SXM module	Up to 1,000W per GPU (DGX B200) (Source: docs.nvidia.com)	~14.3kW per DGX B200 node; rack-level power is substantially higher for liquid-cooled NVL72
Launch context	Announced March 22, 2022 (Source: nvidianews.nvidia.com); full production September 2022 (Source: nvidianews.nvidia.com)	Built on Hopper architecture	Announced March 18, 2024 (see Blackwell Architecture section above)	Announced March 18, 2024

As the table shows, the generational leap is concentrated in three areas: nearly triple the memory bandwidth at the GPU level, double the NVLink bandwidth per GPU, and the addition of a native FP4 precision mode absent from Hopper entirely. Raw FP8 throughput per GPU also roughly doubles from Hopper to Blackwell (3,958 TFLOPS versus approximately 9,000 TFLOPS on the DGX B200's per-GPU FP8 figure), while FP4 support on Blackwell effectively doubles throughput again for workloads that tolerate the lower precision. These architectural gains, however, come paired with a substantially higher per-GPU power draw, rising from 700W to 1,000W, which cascades into rack-level cooling and power-delivery requirements that many existing data centers cannot support without retrofit.

Table 2 below compares representative cloud rental pricing for H100 and B200 GPUs across several providers, illustrating how the architecture gap translates into operating cost.

PROVIDER	H100 (PER GPU-HOUR)	B200 (PER GPU-HOUR)	NOTES
CoreWeave	$4.76 (HGX H100, on-demand) (Source: coreweave.com)	Not listed on classic pricing page at time of access	CoreWeave's newer pricing page separately lists H100 on-demand at $49.24/hour for an 8-GPU node equivalent and B200 NVL rates up to roughly $10.50/hour
Lambda	Not directly captured this session	$6.99 (B200 SXM, on-demand) (Source: getdeploying.com)	Low stock at time of access
Runpod	Not directly captured this session	$5.89 to $5.98 (B200 SXM, secure/community cloud) (Source: getdeploying.com)	Two tiers with different SLAs
Market average (GetDeploying, 24 providers tracked)	Not separately averaged this session	$5.95/hour average; $2.69/hour lowest spot price (Source: getdeploying.com)	Aggregated across on-demand and spot listings
Outright purchase (research estimate, GPU-cloud broker Modal)	$25,000 to $30,000 per GPU (Source: modal.com)	$40,000 or more per GPU (Source: modal.com)	Capital purchase price, distinct from cloud rental; H200 falls between the two at an estimated $30,000 to $40,000 per GPU (Source: modal.com)

The pattern across providers is consistent: B200 rentals command a premium of roughly 25% to 75% over comparable H100 rentals on a per-GPU-hour basis, though the gap narrows or reverses in some markets as B200 supply increases and H100 demand shifts toward lighter, latency-tolerant inference workloads that do not require the newest silicon. Outright purchase prices follow a similar pattern at a much larger scale: GPU cloud broker Modal estimated H100 GPUs cost roughly $25,000 to $30,000 each, H200 GPUs roughly $30,000 to $40,000, and B200 GPUs $40,000 or more, noting that "these chips are quite pricy to buy outright, not to mention that availability is incredibly constrained" for individual buyers (Source: modal.com).

Performance and Benchmarks

Benchmark evidence for the Hopper-versus-Blackwell comparison comes primarily from **MLPerf**, an independently governed benchmark suite operated by MLCommons, a nonprofit engineering consortium. On **MLPerf Inference v5.0**, published in 2025, NVIDIA submitted results for both the Blackwell and Hopper architectures across every data center benchmark (Source: developer.nvidia.com). On the newly introduced Llama 3.1 405B benchmark, a 405-billion-parameter dense large language model, the rack-scale GB200 NVL72 delivered up to 3.4x higher per-GPU performance in the server scenario and 2.8x in the offline scenario compared to an eight-GPU H200 system, and up to a 30x increase in absolute system-level throughput, reflecting both higher per-GPU performance and 9x more GPUs connected in a single NVLink domain (Source: developer.nvidia.com).

On the Llama 2 70B Interactive benchmark, which imposes stricter latency requirements than the standard Llama 2 70B test, an eight-GPU B200 system achieved 3.1x higher throughput than an eight-GPU H200 submission (Source: developer.nvidia.com). Across three other established benchmarks, eight Blackwell GPUs outperformed eight H200 GPUs by 3x (server) and 2.8x (offline) on standard Llama 2 70B, 2.1x on the Mixtral 8x7B mixture-of-experts model, and 1.6x on Stable Diffusion XL image generation, according to NVIDIA's published results retrieved from MLCommons (Source: developer.nvidia.com).

Importantly, Hopper did not stand still during this period. The same MLPerf v5.0 round showed H100 throughput on Llama 2 70B increasing by up to 1.5x over the prior year purely from software optimizations, including GEMM (general matrix multiply) and attention kernel improvements, advanced kernel fusion, and chunked prefill techniques in TensorRT-LLM (Source: developer.nvidia.com). On **MLPerf Training v5.0**, which introduced a new pretraining benchmark based on Llama 3.1 405B, MLCommons recorded 201 total performance results from 20 submitting organizations, including first-time submissions using two distinct NVIDIA Blackwell processor configurations (GB200 and B200-SXM-180GB) alongside continuing submissions on Hopper-class hardware (Source: mlcommons.org) (Source: mlcommons.org).

Independent third-party benchmarking corroborates the general magnitude of these gains outside NVIDIA's own submissions. Cloud infrastructure provider Civo, in an independent technical comparison, described Blackwell's B200 as bringing "a chiplet design, doubled memory capacity, next-level precision support, and massive bandwidth gains" relative to H100 (Source: civo.com), and an independent early-access benchmark by machine learning tooling vendor Lightly reported the B200 running up to 57% faster than the H100 on real-world computer vision pretraining and large language model inference workloads, rather than synthetic tests (Source: lightly.ai). Separately, GPU server integrator Exxact reviewed verified performance data across both platforms and reported that Blackwell-generation B300 and B200 systems achieve "up to 11 to 15 times faster LLM throughput per GPU" compared with the Hopper generation, while noting that H100 and H200 "remain strong in FP64 and FP8 compute for HPC and training workloads" (Source: exxactcorp.com) (Source: exxactcorp.com). These independently gathered figures are directionally consistent with, though larger and smaller in magnitude at different points than, NVIDIA's own MLPerf submissions, a spread that is typical between vendor-optimized benchmark submissions and third-party estimates before workload-specific tuning is applied.

Data Analysis and Evidence

Quantifying the Hopper-to-Blackwell transition requires triangulating NVIDIA's financial disclosures, MLCommons' independently audited benchmark results, and market pricing data, since no single source captures the full picture.

On the financial side, NVIDIA's Data Center segment, which includes both architectures, posted record quarterly revenue of $75.2 billion in the first quarter of fiscal 2027 (ended April 26, 2026), an increase of 92% from the same quarter a year earlier and 21% sequentially (Source: nvidianews.nvidia.com). Within that figure, Data Center compute revenue, the subset most directly tied to GPU sales rather than networking equipment, reached $60.4 billion, up 77% year over year (Source: nvidianews.nvidia.com). Full fiscal year 2026 Data Center revenue reached $193.7 billion, up 68% from fiscal 2025 (Source: nvidianews.nvidia.com), and total company revenue for fiscal 2026 was $215.9 billion, up 65% from the prior year (Source: nvidianews.nvidia.com). NVIDIA CEO Jensen Huang characterized the trend by stating "Grace Blackwell with NVLink is the king of inference today, delivering an order-of-magnitude lower cost per token" (Source: nvidianews.nvidia.com). It is worth noting that these figures reflect combined Hopper and Blackwell shipments, and NVIDIA has not publicly disaggregated revenue by architecture; the growth trend is therefore informative about overall AI infrastructure demand rather than a precise Blackwell-only figure.

On the benchmark side, MLPerf Inference v5.0 recorded Blackwell-based systems (both GB200 NVL72 and DGX B200) delivering between 1.6x and 3.4x higher per-GPU throughput than Hopper-based H200 systems, depending on the specific model and scenario tested, as detailed in the Performance and Benchmarks section above. At the platform level, NVIDIA's own marketing claims for GB200 NVL72 cite up to 30x faster real-time LLM inference and 25x lower cost and energy consumption compared to an equivalent number of H100 GPUs (Source: nvidianews.nvidia.com), and NVIDIA's energy-efficiency comparisons state that GB200 delivers 25x more performance at the same power draw as H100 air-cooled infrastructure (detailed further in the Frequently Asked Questions section below). These platform-level multiples are substantially larger than the per-GPU MLPerf multiples because they compound higher per-GPU throughput with denser GPU packing per rack (72 GPUs in an NVL72 versus 8 in a typical HGX server) and should not be read as equivalent to a single-GPU comparison.

On the pricing side, aggregator GetDeploying, which tracks 24 cloud providers for B200 rentals, reported an average price of $5.95 per GPU-hour with a low of $2.69 per GPU-hour on spot pricing as of its most recent update (Source: getdeploying.com), while CoreWeave's classic pricing tier lists HGX H100 nodes at $4.76 per GPU-hour on demand (Source: coreweave.com). Reddit discussion in the r/BetterOffline community, in a thread discussing GPU depreciation, characterized rapid value decline in older GPU generations as newer architectures ship, framing it as a recurring capital risk for buyers who purchase rather than rent GPU capacity (Source: reddit.com). Such community sentiment should be read as informal commentary rather than an audited financial metric, but it reflects a genuine planning consideration: rapid architectural cadence compresses the useful economic life of any single GPU generation, favoring rental or short-depreciation-cycle strategies over long-term capital purchases for organizations without guaranteed multi-year utilization.

Finally, on the policy dimension that shapes global GPU availability, U.S. export control decisions materially affect which architecture is available in which market. In December 2025, the Trump administration announced it would permit export of NVIDIA's H200 chips (Hopper generation) to China subject to a 25% fee, while explicitly declining to permit export of Blackwell-generation chips, which the administration and industry analysts described as meaningfully faster (Source: reuters.com). The nonpartisan Institute for Progress estimated that Blackwell chips in use by U.S. firms were roughly 1.5 times faster than H200 for training and five times faster for inference, according to Reuters' reporting on the think tank's analysis (Source: reuters.com), while NVIDIA's own research suggested Blackwell servers sped up AI models by up to tenfold for some tasks relative to H200, per separate Reuters reporting (Source: reuters.com). NVIDIA has stated it is not assuming any Data Center compute revenue from China in its forward guidance, underscoring how export policy, not just architecture, now shapes addressable market size for both product lines (Source: nvidianews.nvidia.com).

Total cost of ownership (TCO) calculations further complicate a simple hourly-rate comparison. NVIDIA's own framing of GB200 NVL72 economics claims up to 25x lower total cost of ownership and 25x less energy consumption for real-time inference compared with prior-generation Hopper infrastructure, a figure CoreWeave cited directly when announcing general availability of its GB200 NVL72 instances (Source: investors.coreweave.com). Because Blackwell racks pack far more compute per unit of data center floor space and per watt than Hopper racks, once a workload is large enough to fully utilize a dense GB200 NVL72 or GB300 NVL72 deployment, the effective cost per token or per training step can fall well below Hopper's, even though the sticker price of a single Blackwell GPU or GPU-hour is higher. Conversely, workloads too small to saturate a 72-GPU NVLink domain, or that only intermittently need GPU capacity, often cannot realize these TCO gains and are better served by right-sized Hopper deployments or fractional cloud rentals. This is precisely why Reddit's r/BetterOffline community discussion of GPU depreciation frames the calculus as a capital risk rather than a pure performance question: an organization that buys Hopper hardware outright, rather than renting, bears the full cost of an architecture that can lose the majority of its resale value within a few years of a faster successor's release (Source: reddit.com).

Cloud GPU marketplace Spheron illustrates how sensitive this cost-per-token math is to precision mode. At on-demand pricing, Spheron measured B200 FP8 inference as roughly 51% more expensive per token than H100, while B200 FP4 inference measured about 26% cheaper per token than H100 at the same on-demand rate, because FP4 approximately doubles throughput on the same GPU for workloads that tolerate the lower precision (Source: spheron.network). Spheron's own B200 pricing listed $2.12 per hour at spot and $6.03 per hour on-demand, alongside H100 pricing of roughly $2.00 to $2.50 per hour, underscoring that the B200's economic advantage over H100 depends heavily on whether an inference workload can run in FP4 rather than FP8 (Source: spheron.network). Spheron's spec sheet also corroborates the B200's 192GB HBM3e memory, 8.0 TB/s memory bandwidth (roughly 2.4 times the H100's 3.35 TB/s), and 1,000W thermal design power against the H100's 700W (Source: spheron.network).

Case Studies and Real-World Examples

xAI's Colossus Supercomputer (Hopper, with Blackwell Expansion)

xAI's Colossus facility in Memphis, Tennessee, illustrates both the scale Hopper GPUs can reach and the transition path toward Blackwell. According to xAI's own published figures, the initial cluster combined to reach 200,000 H100 GPUs in a single interconnected system, built in 122 days and then doubled in a further 92 days (Source: x.ai) (Source: x.ai). A more recent third-party account describes the cluster as now comprising approximately 150,000 H100 GPUs, 50,000 H200 GPUs, and 30,000 GB200 (Blackwell) GPUs, making it, by that description, the world's largest single, coherent AI training cluster as of the account's publication (Source: introl.com). The facility demonstrates that Hopper-generation GPUs remain viable at the largest scale even as Blackwell GPUs are incorporated incrementally, rather than requiring a wholesale architectural replacement.

Microsoft Azure's GB200 NVL72 Rollout

Microsoft Azure represents a direct hyperscaler case of Blackwell adoption layered onto existing Hopper infrastructure. At GTC 2025, Microsoft announced the general availability of the Azure ND GB200 V6 virtual machine series, powered by GB200 NVL72 and NVIDIA Quantum InfiniBand networking, explicitly positioned alongside Azure's existing H200- and H100-based virtual machines rather than replacing them (Source: crn.com). Microsoft framed the addition as supporting "the next wave of complex AI tasks like planning, reasoning, and adapting in real-time," language that maps directly onto the agentic and reasoning workloads Blackwell's larger memory pool and FP4 inference mode are designed to accelerate (Source: crn.com). Separately, Microsoft's Satya Nadella described bringing "the GB200 Grace Blackwell processor to our datacenters globally" as building on the company's "long-standing history of optimizing NVIDIA GPUs for our cloud" (Source: nvidianews.nvidia.com).

Meta's Multigenerational NVIDIA Partnership

Meta's relationship with NVIDIA spans both architectures and illustrates how large AI labs plan multi-year hardware roadmaps rather than committing to a single generation. At Blackwell's March 2024 unveiling, Meta CEO Mark Zuckerberg said the company was "looking forward to using NVIDIA's Blackwell to help train our open-source Llama models and build the next generation of Meta AI and consumer products" (Source: nvidianews.nvidia.com). By NVIDIA's fourth-quarter fiscal 2026 results (cited in the Adoption section above), the relationship had expanded into what NVIDIA described as a "multiyear, multigenerational strategic partnership with Meta spanning on-premises, cloud and AI infrastructure, including the large-scale deployment of NVIDIA CPUs, networking and millions of NVIDIA Blackwell and Rubin GPUs," demonstrating a planned transition that spans Hopper-era Llama training through Blackwell and into NVIDIA's next-generation Rubin platform.

CoreWeave's First-to-Market GB200 NVL72 Deployment

Specialized cloud provider CoreWeave illustrates how quickly a Hopper-focused infrastructure provider pivoted to Blackwell once it became available. In February 2025, CoreWeave announced it was the first cloud provider to make GB200 NVL72-based instances generally available, offering rack-level NVLink connectivity and NVIDIA Quantum-2 InfiniBand networking at 400 gigabits per second (Gb/s) per GPU across clusters of up to 110,000 GPUs (Source: investors.coreweave.com). CoreWeave cited NVIDIA's own performance claims for the platform, up to 30x faster real-time LLM inference, up to 25x lower total cost of ownership and energy consumption, and up to 4x faster LLM training compared to previous-generation infrastructure (Source: investors.coreweave.com). Notably, CoreWeave's rollout followed directly from its earlier position as one of the first cloud providers to deploy H200 GPUs (Hopper generation) to train large language models in August 2024, illustrating a provider moving through both architectures within roughly six months of each other rather than skipping a generation (Source: investors.coreweave.com). CoreWeave subsequently became the first hyperscale cloud provider to deploy the follow-on GB300 NVL72 platform in July 2025, a cadence that illustrates how compressed the replacement cycle has become even for infrastructure specialists (Source: investors.coreweave.com).

(Hypothetical Example) A Mid-Sized AI Startup Choosing Between H100 and B200 for Fine-Tuning

Consider a hypothetical AI startup fine-tuning a 70-billion-parameter open-source model for a customer support application, evaluating whether to rent H100 or B200 capacity. Based on the pricing and performance data gathered in this report, an eight-GPU H100 cluster renting at approximately $4.76 per GPU-hour (Source: coreweave.com) would cost roughly $38 per hour for the full node, while a comparable B200 configuration at the market average of $5.95 per GPU-hour (Source: getdeploying.com) would cost approximately $47.60 per hour, a 25% premium. Given MLPerf's measured 3x throughput advantage for B200 over H200 (a closer comparator than H100) on the standard Llama 2 70B benchmark in server mode (see Performance and Benchmarks above), the startup's effective cost per token would likely favor B200 despite its higher hourly rate, illustrating why raw hourly price comparisons alone can mislead procurement decisions that should instead weigh cost per unit of completed work.

Implications and Future Directions

The Hopper-to-Blackwell transition previews a broader pattern in NVIDIA's roadmap: architectural generations increasingly optimize specifically for inference at massive scale rather than general-purpose acceleration, and each generation compounds gains from lower-precision arithmetic (FP8 to FP4), higher-bandwidth interconnect, and denser rack-scale integration rather than from transistor count alone. NVIDIA has already signaled the next step in this progression. At its fourth-quarter fiscal 2026 results, the company unveiled the **NVIDIA Vera Rubin** platform, comprising six new chips designed to deliver up to a 10x reduction in inference token cost compared with Blackwell, with AWS, Google Cloud, Microsoft Azure, and Oracle Cloud Infrastructure named among the first cloud providers expected to deploy Vera Rubin-based instances (Source: nvidianews.nvidia.com). This means organizations planning infrastructure purchases in the second half of 2026 and beyond should treat Blackwell not as the endpoint of the roadmap but as a mid-cycle generation that will itself face the same "should we upgrade" calculus Hopper faces today.

Independent benchmarking firm SemiAnalysis has already published InferenceX results showing Blackwell Ultra (an enhanced Blackwell variant, distinct from the original B200/GB200) delivering up to 50x better performance and 35x lower cost for agentic AI workloads compared with the Hopper platform, according to NVIDIA's citation of that research (Source: nvidianews.nvidia.com). This suggests the effective performance gap between Hopper and the newest Blackwell variants is wider than the original B200-versus-H100 comparison implies, and organizations benchmarking today should specify precisely which Blackwell SKU (original B200/GB200, or the newer Blackwell Ultra/GB300) they are evaluating against which Hopper SKU (H100 or H200), since the terms are frequently used loosely in vendor and press materials.

Software and framework maturity will also continue to narrow the gap between the two architectures over time, as it has throughout Hopper's lifecycle. NVIDIA's own MLPerf disclosures, discussed above, show that GEMM and attention kernel optimizations, kernel fusion, and chunked prefill techniques added up to 1.5x additional Llama 2 70B throughput on unchanged H100 silicon within a single year, and there is no structural reason equivalent optimization work will not continue to extract additional throughput from Blackwell GPUs over their own multi-year deployment lifecycle. Organizations benchmarking a purchase decision today should therefore treat any single-point-in-time benchmark comparison, including the MLPerf figures cited throughout this report, as a snapshot rather than a permanent ceiling, and should weight vendor software roadmap commitments (TensorRT-LLM, NVIDIA Dynamo, and similar inference-serving stacks) alongside raw hardware specifications when making a multi-year infrastructure commitment.

Power and cooling infrastructure will likely remain the binding constraint on how quickly organizations can adopt Blackwell regardless of its performance advantages. With per-GPU power draw rising from 700W (H100) to 1,000W (B200) (Source: docs.nvidia.com) and full rack-scale Blackwell systems requiring liquid cooling (discussed further in the FAQ below), data centers built for air-cooled Hopper deployments cannot simply

swap in Blackwell hardware without capital investment in cooling infrastructure. This dynamic partly explains why NVIDIA continues to offer the air-cooled, drop-in-compatible B100 as a bridge product for exactly this scenario (Source: introl.com), and why H100 and H200 GPUs remain in heavy use at even the largest AI labs rather than being immediately retired.

Finally, geopolitical export policy is likely to keep shaping which architecture is available in which region for the foreseeable future. With Blackwell exports to China still restricted as of December 2025, while H200 exports were newly permitted subject to a 25% fee (Source: reuters.com), Hopper-generation hardware is likely to remain the most advanced NVIDIA architecture legally available in some markets well after Blackwell becomes the default in the United States and allied markets, a divergence that will shape competitive dynamics between regionally isolated AI ecosystems.

Frequently Asked Questions (FAQs)

Is Blackwell faster than Hopper? Yes, on nearly every published benchmark. MLPerf Inference v5.0 results show Blackwell-based systems delivering between 1.6x and 3.4x higher per-GPU throughput than Hopper-based H200 systems depending on the model and scenario, and up to 30x higher throughput at the full rack-scale system level when comparing GB200 NVL72 against an equivalent number of H100 GPUs (see Performance and Benchmarks above).

How does the B200 compare to the H100 in specs? The B200 offers 192GB of HBM3e memory versus the H100's 80GB of HBM3, roughly 2.4 times the capacity, alongside native FP4 precision support that the H100 lacks entirely and double the NVLink bandwidth per GPU (1.8TB/s versus 900GB/s), as summarized in Table 1 above.

How does Blackwell compare to the H200? The H200 shares Hopper's core compute architecture but adds a larger, faster memory pool (141GB HBM3e at 4.8TB/s) than the original H100 (see Hopper Architecture above). Against B200, the H200 still trails significantly on interconnect bandwidth and lacks native FP4 support, which is why MLPerf shows B200 outperforming H200 by roughly 1.6x to 3.1x across most inference benchmarks at the eight-GPU system level (see Performance and Benchmarks above).

How much power does Blackwell use compared to Hopper? A single Blackwell GPU in the DGX B200 draws up to 1,000W, compared with up to 700W for an H100 SXM module, and a full DGX B200 chassis draws up to 14.3kW, as detailed in Table 1 above. NVIDIA states that GB200 NVL72 delivers 25x more performance than H100 air-cooled infrastructure at the same power draw, meaning the additional wattage is more than offset by performance per watt at the rack level (see Data Analysis and Evidence above).

How does the B100 compare to the H100? The B100 is a Blackwell-architecture GPU designed as a drop-in replacement for existing HGX H100 server infrastructure, operating at the same 700W per-GPU thermal design power as H100 to allow data centers without upgraded cooling to adopt some Blackwell capabilities without a full facility retrofit (Source: introl.com).

What is NVIDIA's GPU roadmap after Blackwell? NVIDIA has announced the Vera Rubin platform as Blackwell's successor, comprising six new chips targeting up to a 10x reduction in inference token cost compared with Blackwell, with major cloud providers expected to be early adopters (see Implications and Future Directions above).

Should I buy Hopper or Blackwell GPUs today? The answer depends on workload size and infrastructure readiness. Organizations training or serving models above roughly 100 billion parameters, or requiring FP4 inference efficiency and NVLink 5 bandwidth, generally benefit most from Blackwell. Organizations with existing air-cooled Hopper infrastructure, workloads in the tens-of-billions-of-parameters range, or budget constraints often find H100 or H200 GPUs, whose throughput continues to improve through software updates, a more cost-effective choice, particularly at current cloud rental price gaps of roughly 25% to 75% between the two generations (Source: coreweave.com) (Source: getdeploying.com). GPU marketplace Spheron frames the decision similarly: it recommends staying on H100 when "your models fit in 80 GB" and "training is the primary workload," since "the B200's FP4 advantage is inference-specific," and recommends upgrading once an organization is "running 70B+ models or paying $2K+/month in H100 inference costs" (Source: spheron.network) (Source: spheron.network).

Can I run Blackwell GPUs in an existing Hopper-generation data center? In some cases, yes. NVIDIA's air-cooled B100 GPU is designed as a drop-in replacement for existing HGX H100 server designs, operating at the same 700W per-GPU thermal design power as H100, specifically so operators without liquid-cooling infrastructure can adopt some Blackwell capabilities without a full facility retrofit (Source: introl.com). Rack-scale systems like GB200 NVL72 and GB300 NVL72, however, require liquid cooling and cannot be retrofitted into a standard air-cooled Hopper facility without significant capital investment (Source: nvidia.com).

Is it cheaper to rent H100s or B200s in the cloud? As of mid-2026, H100 rentals are generally cheaper on a per-GPU-hour basis, with representative on-demand pricing around $4.76 per GPU-hour for an HGX H100 node (Source: coreweave.com), compared with a market average around $5.95 per GPU-hour for B200 across 24 tracked providers (Source: getdeploying.com). However, because B200 typically completes a given

inference or training workload faster, its effective cost per completed unit of work, tokens generated or training steps completed, can be lower despite the higher hourly rate, particularly for large language models above 70 billion parameters where MLPerf shows the largest Blackwell speedups (see Performance and Benchmarks above).

Conclusion

Hopper and Blackwell represent two generations of NVIDIA data center GPU built for different moments in the AI infrastructure buildout. Hopper, anchored by the H100 and H200, established the hardware baseline for the first wave of large-scale generative AI training and remains a mature, widely available, and increasingly cost-optimized option, with throughput on established benchmarks still improving through software alone years after launch. Blackwell, anchored by the B100, B200, and rack-scale GB200 and GB300 NVL72 systems, roughly doubles memory capacity and NVLink bandwidth per GPU, introduces native FP4 precision, and delivers benchmark-verified throughput gains ranging from roughly 1.6x to 3.4x per GPU and up to 30x at the full rack-scale system level relative to Hopper-class hardware, at the cost of substantially higher power density and a cooling infrastructure requirement that not every data center can meet today.

Neither architecture is a universal answer. The choice between them turns on model scale, latency requirements, existing power and cooling infrastructure, and whether an organization can capture Blackwell's efficiency gains at sufficient utilization to justify its price premium and infrastructure retrofit costs. As NVIDIA's Data Center revenue trajectory and its already-announced Vera Rubin successor both indicate, this generational cadence shows no sign of slowing, and organizations evaluating Blackwell versus Hopper today should frame the decision not as a final architecture choice but as one stage in a recurring, multi-year infrastructure planning cycle.

Tags: blackwell vs hopper, nvidia blackwell, nvidia hopper, b200 vs h100, h100 vs b200, gpu architecture, nvidia h200, gb200 nvl72, ai data center gpu

---