The NVIDIA H200 SXM5: The Complete Guide for AI and HPC Infrastructure in 2026

By Joshua Rhines, Founder & CEO, CubCloud AI — March 2026

We are at an inflection point in data center GPU history. NVIDIA announced the Vera Rubin platform at CES 2026 and confirmed it is already in production. Blackwell B200 and B300 systems are shipping to hyperscalers and AI labs. And yet, for the overwhelming majority of organizations making infrastructure decisions right now, the H200 SXM5 remains the most practical, most available, and most ecosystem-mature GPU on the market.

That situation will not last indefinitely — but it is the reality of March 2026, and it is the context in which this guide is written.

Almost every H200 article gets the story wrong in the same way. They lead with a raw FLOPS number, draw a graph showing a big performance jump over the H100, and attribute the improvement to next-generation architecture. The truth is simpler, more interesting, and more important for anyone making infrastructure decisions: the H200 and H100 have the same compute engine. Identical. The same FP8 TFLOPS. The same FP16 throughput. The same Transformer Engine running the same workloads at the same speed.

Every performance advantage the H200 has over the H100 comes from one thing: memory.

Understanding why that matters — and why it matters so profoundly for the workloads driving AI infrastructure decisions right now — is the entire story of the H200.

Full Technical Specifications

The H200 SXM5 is built on the same GH100 die as the H100. Same Hopper architecture, same CUDA programming model, same SXM5-compatible baseboards. The upgrade is entirely in the memory subsystem — 6 stacks of 24 GB HBM3e, compared to the H100's 5 stacks of 16 GB HBM3. This is not marketing language. It is the hardware reality.

Specification	H200 SXM5	H100 SXM5
Architecture	Hopper (GH100)	Hopper (GH100)
Transistors	~80B (TSMC 4N)	~80B (TSMC 4N)
Memory	141 GB HBM3e	80 GB HBM3
Memory Bandwidth	4.8 TB/s	3.35 TB/s

FP8 Tensor Compute	3,958 TFLOPS	3,958 TFLOPS
BF16 Tensor Compute	1,979 TFLOPS	1,979 TFLOPS
FP64 (Tensor Core)*	67 TFLOPS	67 TFLOPS
TDP	700W	700W
NVLink	4th Gen, 900 GB/s	4th Gen, 900 GB/s
PCIe Interface	Gen5	Gen5
MIG Instances	Up to 7	Up to 7
8-GPU HGX FP8	32+ petaFLOPS	32+ petaFLOPS
8-GPU Aggregate HBM	1.1 TB	640 GB

> * FP64 Tensor Core performance per the GH100 architecture whitepaper (33.5 TFLOPS standard FP64 CUDA cores; 66.9 TFLOPS FP64 Tensor Cores). NVIDIA's H100 product page separately states '60 teraflops of FP64 computing for HPC' — a figure derived from 3x the A100's 19.5 TFLOPS Tensor Core baseline. All FP8, BF16, and TF32 figures are with 2:1 structured sparsity.

The key takeaway: every compute metric is identical between the H200 and H100. The H200's differentiation is purely in memory capacity and bandwidth. This is not a limitation — it is a precisely targeted architectural decision that turns out to be exactly what the industry needed.

H200 SXM5 vs H200 NVL: Which Form Factor, and Why

There are two distinct H200 hardware designs, and they are meaningfully different.

The H200 SXM5 is the high-performance cluster GPU. It mounts on a dedicated SXM5 baseboard inside an HGX server, with 4 or 8 GPUs in a fully-connected NVSwitch mesh topology. Each GPU communicates with every other at 900 GB/s simultaneously — the full all-to-all bandwidth that makes large-scale distributed training and multi-GPU inference work efficiently. TDP is 700W per GPU. This is the form factor discussed throughout the rest of this guide, and the one CubCloud operates.

The H200 NVL is the enterprise flexibility GPU. It is a dual-slot PCIe Gen5 card designed for organizations that need H200 performance without a full HGX infrastructure build. Per-GPU TDP is 600W, and multi-GPU connectivity uses NVLink bridges (2-way or 4-way) rather than a full NVSwitch fabric.

Specification	H200 SXM5	H200 NVL (PCIe)
Form factor	SXM5 baseboard	Dual-slot PCIe Gen5
TDP	700W	600W
Memory per GPU	141 GB HBM3e	141 GB HBM3e

Memory bandwidth	4.8 TB/s	4.8 TB/s
Multi-GPU topology	NVSwitch full-mesh (up to 8)	NVLink bridges (2- or 4-way)
4-GPU aggregate memory	564 GB	564 GB
4-GPU aggregate bandwidth	~1.8 TB/s	~1.8 TB/s
Cooling	Air or liquid (server-dependent)	Air-cooled compatible
Software bundle	None included	5-year NVIDIA AI Enterprise
Upgrade path	Drop-in for H100 SXM5	Drop-in for H100 NVL, A100

The decision rule: For distributed training across 8 GPUs, multi-node clusters, or production inference requiring full NVSwitch bandwidth — SXM5. For H200 memory capacity and bandwidth in a flexible PCIe deployment, standard air-cooled rack, or existing server chassis — NVL. For organizations upgrading directly from H100 NVL infrastructure — H200 NVL is the drop-in path.

Why Memory Is the Bottleneck That Actually Matters

A large language model generates text token by token. At each step, the GPU must load the model's weights from memory to compute the next token and maintain a KV (key-value) cache that stores the entire conversation context. For a 70B parameter model like Llama 2 70B, the weights alone consume roughly 140 GB at FP16 precision. Long-context requests add further pressure.

An H100 with 80 GB of HBM3 cannot load a full 70B model at FP16. It requires either quantization to reduce precision, or sharding across multiple GPUs — adding inter-GPU communication overhead, deployment complexity, and cost. An H200 with 141 GB loads the same model natively on a single GPU. The performance difference is not because the H200 computes faster. It is because the H200 stops waiting for data.

This is the defining insight of the Hopper generation: modern LLM inference is memory-bound, not compute-bound. The Tensor Cores can process tokens faster than the memory subsystem can feed them. Widening the memory pipeline unlocks latent compute capacity that already existed in the H100 but could not be fully utilized.

Benchmark Results: What the Numbers Actually Mean

LLM Inference

NVIDIA's debut H200 submission in MLPerf Inference v4.0 delivered up to 45% more throughput than the H100 on Llama 2 70B. In v4.1, that advantage widened to up to 1.5x versus H100 submissions across the full benchmark suite, with software optimizations alone delivering an additional 27% performance improvement on the same H200 hardware between rounds. For larger models, the

advantage grows: on GPT-3 175B, the H200 runs approximately 1.6x faster than the H100 per NVIDIA's published benchmarks. The pattern is consistent — the larger and more memory-bound the inference workload, the greater the H200's advantage over the H100.

One of the most instructive data points from MLPerf testing: the 27% v4.0-to-v4.1 improvement on H200 came from software optimizations alone — same hardware, better algorithms exploiting the memory headroom. The H100, running closer to its memory ceiling, has less room for software-driven improvement.

Multi-GPU Scaling

Independent benchmarks using the vLLM framework on Llama 3.1 8B Instruct show the H200 delivering 9–10% higher throughput than the H100 across 1, 2, 4, and 8 GPU configurations. For models that push memory limits, the advantage grows substantially.

HPC Workloads

For molecular dynamics, climate simulation, computational fluid dynamics, and financial risk modeling, the H200's 43% memory bandwidth increase over the H100 translates directly to runtime. Financial risk models complete approximately 40% faster on H200 than H100. Molecular dynamics simulations finish 30–35% sooner. Across a broad mix of HPC benchmarks, the H200 delivers approximately 1.7x the throughput of the H100, while both represent roughly 2x the performance of the prior-generation A100.

H200 vs H100: The Decision Framework

Given that the H200 carries a per-GPU price premium of approximately 15–20% over the H100 at equivalent configurations, while delivering identical compute specifications, the decision is clean.

Choose H200 when:

- Serving 70B+ parameter models in production at FP16 or BF16 precision
- Running long-context inference where KV cache consumes significant memory
- Operating multi-modal AI systems maintaining text, image, and embedding data simultaneously
- Running HPC simulations that are memory-bandwidth-bound — most serious scientific workloads qualify
- Wanting to avoid tensor-parallel sharding across multiple GPUs for large model inference
- Prioritizing simpler deployment: fewer GPUs, same output

Choose H100 when:

- Models are under 70B parameters and fit comfortably in 80 GB

- Running distributed training where multi-node cluster economics outweigh single-GPU memory gains
- Cost per GPU-hour matters more than performance per GPU for the workload in question
- Existing H100 infrastructure is being expanded rather than replaced

The Software Ecosystem

The H200 inherits the full Hopper software stack — the most mature AI acceleration ecosystem in existence.

Training: PyTorch, TensorFlow, JAX, and MXNet all have full H200 support via CUDA 12.x. NVIDIA's Megatron-LM for distributed training, Microsoft's DeepSpeed, and Meta's FSDP are production-tested on HGX H200 clusters.

Inference: TensorRT-LLM is the primary high-performance runtime for Hopper, delivering the optimized kernels behind the MLPerf results above. vLLM has native H200 support, benefiting directly from the larger memory for longer context windows and larger batch sizes. NVIDIA Triton inference server handles the production serving layer.

Scientific computing: cuBLAS, cuFFT, cuSPARSE, and cuDNN are all optimized for Hopper's memory architecture. GROMACS, AMBER, and NAMD have H200-optimized builds. OpenFOAM and other CFD packages benefit from the bandwidth increase for memory-intensive fluid simulation.

Multi-Instance GPU: The H200 supports up to 7 isolated instances per GPU, each with dedicated compute, memory, and cache resources — critical for multi-tenant inference where a single H200 serves 7 independent workloads with full isolation and no resource contention.

Confidential computing: Full support for NVIDIA's Confidential Computing framework enables encrypted computation for regulated industries where data must remain protected from the infrastructure operator. For healthcare, financial services, and government deployments, this is a production compliance feature.

Real-World Use Cases

Foundation model serving. Organizations running production LLM APIs in 2026 are increasingly serving 70B–405B parameter models. The H200's ability to run Llama 3.1 70B on a single GPU rather than a two-GPU sharded configuration reduces infrastructure cost and latency simultaneously.

Long-context applications. Legal contract analysis, clinical note summarization, codebase comprehension, scientific literature review — applications processing long documents require large KV caches that scale with context length. The H200's 141 GB enables 128K token context windows

without sacrificing batch size.

Fine-tuning and continual learning. The H200's ability to keep larger model states resident on-GPU during fine-tuning enables more aggressive gradient accumulation and larger effective batch sizes — faster convergence and shorter time-to-deployment for custom domain models.

Drug discovery and molecular simulation. Molecular dynamics simulations are among the most memory-intensive scientific workloads. The H200's 141 GB allows larger molecular systems to be simulated on a single GPU, directly reducing screening timelines in hit-identification workflows.

Climate and atmospheric modeling. High-resolution climate models are memory-bandwidth-bound. The H200's 43% bandwidth increase over H100 translates directly to faster simulation cycles and higher resolution models for the same compute budget.

Financial risk computing. Monte Carlo simulations and portfolio risk models see consistent 30–40% runtime improvements on H200 relative to H100 — faster intraday risk reassessment and tighter response to market events.

Pricing and Access in 2026

The H200 SXM5 costs between $30,000 and $40,000 per GPU to purchase outright. Full 8-GPU HGX H200 server systems from major OEMs including Supermicro, Dell, and Aivres typically exceed $300,000 for the board alone, with complete server configurations ranging from $400,000 to $500,000+.

Provider	Per GPU/Hour	Notes
Azure	~$10.60	High-availability SLA, 8-GPU minimum
AWS	~$4.33	1-day minimum billing, 8-GPU minimum
Lambda	~$4.49	On-demand (Q1 2026)
RunPod	~$3.99	Per-second billing
Vast.ai	~$2.43	Lowest published rate
CubCloud	Sovereign pricing	Dedicated, locally governed, Montana

Hyperscalers currently offer H200s only in 8-GPU bundles. For single-GPU or small-cluster access, specialist providers are the practical path. The TCO calculation strongly favors ownership over sustained cloud rental for steady-state workloads — at hyperscaler rates, an 8-GPU H200 cluster costs more to rent for a year than to purchase outright. The break-even against specialist cloud pricing typically falls between 18 and 24 months of continuous operation.

H200 in Context: B200, B300, and Vera Rubin

This is where the picture gets genuinely interesting for anyone planning infrastructure beyond the next six months. NVIDIA's GPU roadmap has compressed dramatically, and understanding where the H200 sits relative to what is shipping and what is coming is essential for making defensible infrastructure decisions in 2026.

The B200: Blackwell's Foundation

The B200 is shipping now. Built on TSMC's 4NP process with a dual-die design at 208 billion transistors, it represents the first fully new GPU architecture since Hopper — and the performance jump is real.

Specification	H200 SXM5	B200 SXM6
Architecture	Hopper	Blackwell
Transistors	~80B (TSMC 4N)	208B (TSMC 4NP, dual-die)
Memory	141 GB HBM3e	192 GB HBM3e*
Memory Bandwidth	4.8 TB/s	~8 TB/s
FP8 Tensor Compute (with sparsity)	3,958 TFLOPS	9,000 TFLOPS
FP4 Tensor Compute (dense)	Not supported	9 PFLOPS
FP64 (Tensor Core)*	~67 TFLOPS	~37-40 TFLOPS
TDP	700W	1,000W
NVLink	4th Gen, 900 GB/s	5th Gen, 1.8 TB/s
DGX system inference vs H100	--	15x (NVIDIA claim, rack-level)

* In HGX B200 server configurations, approximately 180 GB is addressable per GPU; 192 GB reflects total die capacity. † FP64 Tensor Core per GH100 whitepaper (66.9 TFLOPS); NVIDIA product page states '60 TFLOPS FP64.' B200 FP64 is CUDA core.

The B200 introduces FP4 precision — a new low-precision inference format the H200 does not support. At 9,000 TFLOPS FP8 with sparsity (4,500 TFLOPS dense), the B200 delivers approximately 2.3x the FP8 throughput of the H200. The bigger architectural leap is FP4: at 9,000 TFLOPS dense FP4, the B200 doubles effective inference throughput again for workloads that can use 4-bit precision — a precision tier unavailable on Hopper entirely. It also nearly doubles memory capacity to 192 GB and nearly doubles bandwidth to ~8 TB/s. NVIDIA claims 15x the inference performance of H100 for the DGX B200 system, though this is a rack-level comparison for the GPT-MoE-1.8T model that includes format and density improvements rather than a straight GPU-to-GPU figure.

The infrastructure gap from H200 to B200 is significant. The 1,000W TDP typically requires liquid cooling, 800 Gbps networking for full-scale multi-node deployments, and power densities that most existing facilities were not designed for. The software ecosystem, while maturing rapidly, is

approximately 18–24 months behind Hopper in production hardening.

Who should be on B200 today: Organizations with greenfield facilities, confirmed liquid cooling infrastructure, frontier model requirements exceeding 141 GB per GPU, and the engineering capacity to work through a less mature software stack. Hyperscalers deploying at the largest scales where the efficiency improvements translate to meaningful OpEx savings.

The B300 (Blackwell Ultra): More Memory, Less FP64

The B300 arrived in 2025 as NVIDIA's "Blackwell Ultra" — an incremental but meaningful step beyond the B200 targeted at the memory and compute ceiling.

Specification	H200 SXM5	B200 SXM6	B300 (Blackwell Ultra)
Memory	141 GB HBM3e	192 GB HBM3e*	288 GB HBM3e
Memory Bandwidth	4.8 TB/s	~8 TB/s	~8 TB/s
Dense FP4	Not supported	9 PFLOPS	14-15 PFLOPS
FP64 (Tensor Core)*	~67 TFLOPS	~37-40 TFLOPS	~1.25 TFLOPS
TDP	700W	1,000W	1,400W

* In HGX B200 server configurations, approximately 180 GB is addressable per GPU; 192 GB reflects total die capacity. † FP64 Tensor Core figure per GH100 whitepaper (66.9 TFLOPS). NVIDIA product page states '60 TFLOPS FP64.' Non-Tensor FP64 (CUDA cores) is ~33.5 TFLOPS. B200/B300 figures are CUDA core FP64.

The B300's headline upgrade is memory: 288 GB of HBM3e — 50% more than the B200, more than double the H200 — enabling single-GPU inference on models that would require multi-GPU sharding even on B200. Dense FP4 compute increases by approximately 55% over the B200. The B300 NVL72 rack delivers 1.1 exaFLOPS of FP4 compute. TDP rises to 1,400W, requiring robust liquid cooling infrastructure.

There is one critical caveat for HPC users that most coverage buries: the B300 trades nearly all FP64 performance for higher FP4 throughput. Where the H200 delivers ~67 TFLOPS of FP64 Tensor Core compute — sufficient for serious scientific simulation — the B300 delivers approximately 1.25 TFLOPS. That is a greater than 50x reduction. For climate modeling, molecular dynamics, computational fluid dynamics, and any scientific workload requiring double-precision arithmetic, the B300 is a poor fit. The H200 — and even the B200, which retains ~37–40 TFLOPS of FP64 — remains the correct choice for these workloads.

Who should be on B300 today: AI labs and enterprises whose primary workload is inference and fine-tuning of very large language models (300B+ parameters) where FP4 throughput and memory capacity are the dominant constraints. Anyone who does not need double-precision scientific computing.

Vera Rubin: The Architecture on the Horizon

At CES 2026, NVIDIA confirmed that Vera Rubin is in production. Partner availability — meaning organizations can actually receive and deploy systems — is scheduled for the second half of 2026. The specifications represent a generational step beyond everything discussed above.

Specification	H200 SXM5	B300	Vera Rubin
Architecture	Hopper	Blackwell Ultra	Rubin
Process node	TSMC 4N	TSMC 4NP	TSMC 3nm
Transistors	~80B	~208B	336B
Memory	141 GB HBM3e	288 GB HBM3e	288 GB HBM4
Memory Bandwidth	4.8 TB/s	~8 TB/s	22 TB/s
FP4 Compute	Not supported	14-15 PFLOPS	50 PFLOPS (NVFP4)
NVLink	4th Gen, 900 GB/s	5th Gen, 1.8 TB/s	6th Gen, 3.6 TB/s
CPU pairing	None	None	Vera (88-core Arm)
NVL rack FP4	--	1.1 EFLOPS (NVL72)	3.6 EFLOPS (NVL72)
Cooling	Air or liquid	Liquid required	Liquid required (100%)

At GTC 2026 (March 16), NVIDIA formally launched Vera Rubin as a seven-chip platform, adding the Groq 3 LPX inference accelerator to the six chips announced at CES: the Rubin GPU, Vera CPU, NVLink 6 switch, ConnectX-9, BlueField-4 DPU, and Spectrum-6 Ethernet switch. The Groq 3 LPX is purpose-built for low-latency decode, pairing with Rubin GPUs to handle the bandwidth-bound phases of inference that GPUs are not optimized for. This is not a GPU upgrade — it is a full-stack AI factory architecture designed for the agentic AI era, where inference dominates and cost-per-token determines commercial viability.

The memory bandwidth jump is the most striking number: 22 TB/s per GPU, compared to 4.8 TB/s on the H200. That is a 4.6x increase in a single architectural generation. Combined with HBM4's lower latency characteristics and tight CPU-GPU coherence via NVLink-C2C at 1.8 TB/s, Rubin treats CPU and GPU memory as a unified pool — eliminating a class of data movement overhead that currently constrains even the best Hopper deployments.

NVIDIA's claim of 10x lower inference cost versus Blackwell requires careful reading. The company is comparing NVFP4 inference performance at rack level in the NVL72 configuration, which reflects both per-GPU performance and higher GPU density per rack — not a straight GPU-to-GPU comparison. The 3.3x per-rack compute improvement over B300 NVL72 is the more direct figure.

Infrastructure requirements for Vera Rubin are substantial. Liquid cooling is mandatory — there is no air-cooled Rubin configuration. The NVL72 rack draws power at levels requiring purpose-built or

significantly upgraded data center facilities. Organizations planning Rubin deployments need to begin facility planning now; lead times on data center power and cooling infrastructure are measured in months to years. Rubin is already in production at Q1 2026 at NVIDIA, with partner availability H2 2026.

Rubin Ultra arrives in H2 2027 in the NVL576 configuration — 576 GPUs per rack, 15 EFLOPS of FP4, HBM4e memory. A planning consideration, not a 2026 purchase decision.

The H200's Position in This Landscape

With all four platforms now in view, the H200's position in 2026 is clear.

It is not the most powerful GPU. The B200 outpaces it on AI compute. The B300 outpaces it on memory capacity and FP4 throughput. Vera Rubin renders both of those conversations moot at a different scale entirely.

But the H200 is, right now, the most deployable, most software-mature, most infrastructure-compatible high-performance AI GPU available. It runs in existing H100 server chassis without facility modifications. It has meaningful FP64 Tensor Core performance (~67 TFLOPS) for scientific computing — a capability Blackwell's B300 largely sacrificed. It supports the same CUDA stack that two-plus years of production optimization has been built around. And it costs substantially less than either Blackwell variant.

For organizations that need to deploy and operate production AI infrastructure in 2026 — not plan for what might be available in Q3 — the H200 SXM5 remains the right answer. That window will close as Blackwell matures and Rubin becomes available. But the window is open now, and the H200 is exceptional hardware for the workloads that exist today.

The Sovereign Case for Owned H200 Infrastructure

There is a dimension to H200 deployment decisions that pricing comparisons do not capture: data governance.

The U.S. CLOUD Act allows American authorities to compel American cloud providers to disclose data regardless of where it physically resides. No enterprise agreement overrides federal law. For organizations in regulated industries — healthcare, finance, defense, legal, research — this is not a hypothetical risk. H200 hardware deployed in locally-governed, sovereign infrastructure is not equivalent to H200 hours rented from a hyperscaler, even when the raw compute is identical.

At CubCloud, we operate H200 SXM5 hardware alongside H100 SXM5 and RTX PRO 6000 Blackwell Server Edition systems in Missoula, Montana. Data processed on our H200 infrastructure stays in Montana, governed by our clients' policies, accessible only to the organizations that generated it. For the institutions we serve — regional universities running research AI, healthcare systems processing

clinical data, government facilities with strict data jurisdiction requirements — this is the deployment decision that determines whether their AI program is legally defensible.

The Bottom Line

The H200 is not a faster H100. It is an H100 with the bottleneck removed.

Modern AI and HPC workloads are memory-limited. The compute resources in both the H200 and H100 are equivalent. The H200's 141 GB of HBM3e at 4.8 TB/s feeds those compute resources more consistently, handles larger models natively on a single GPU, and enables long-context inference, large-model fine-tuning, and 100B+ parameter serving that the H100 can only address through multi-GPU workarounds.

Compared to what is coming — B200, B300, Vera Rubin — the H200 is a proven generation of hardware, not the frontier. But in March 2026, it is the GPU with the largest production installed base, the deepest software optimization, the most direct upgrade path from existing H100 infrastructure, and the only high-performance Hopper option with serious FP64 capability for scientific computing.

The SXM5 is the right choice for HGX cluster deployments. The NVL PCIe is the right choice for flexible enterprise rack environments. Both deliver the same per-GPU memory capacity and bandwidth. Both run the same software stack. Both are available today.

Run it where you own the infrastructure. Govern the data it processes like the strategic asset it is.

Joshua Rhines is the Founder and CEO of CubCloud AI, a sovereign AI infrastructure company headquartered in Missoula, Montana. CubCloud operates NVIDIA H200 SXM5, H100 SXM5, and RTX PRO 6000 Blackwell Server Edition hardware for regulated industries, research institutions, and sovereign AI applications across the Mountain West.

Key Sources

• NVIDIA Technical Blog — "NVIDIA Hopper Architecture In-Depth" (GH100 whitepaper, FP64 specs); "NVIDIA H200 Tensor Core GPU" product page; "NVIDIA H200 NVL PCIe" (Nov 2024); "Inside the NVIDIA Vera Rubin Platform" (Jan 2026); "MLPerf Inference v4.1" (Sep 2024); "Blackwell MLPerf v5.0" (Jan 2026)

• NVIDIA Newsroom / Developer Blog — Vera Rubin platform official launch, GTC 2026 (Mar 16, 2026); DGX B200 product page; DGX B300 product page; Vera Rubin NVL72 product page

• NVIDIA Developer Forums — "Parallel usage of FP64 and Tensor cores in H100" (33.5 TFLOPS CUDA / 66.9 TFLOPS Tensor Core FP64 confirmation)

• Tom's Hardware — GTC 2026 keynote live blog (Mar 17, 2026); Rubin Ultra tray / Kyber rack reveal (Mar 2026); B300 1,400W TDP report (Dec 2024); NVIDIA Announces Rubin GPUs roadmap (Mar 2025)

• Data Center Knowledge — "GTC 2026: NVIDIA Unveils Vera Rubin AI Platform" (Mar 2026) — Rubin Ultra NVL576 15 EFLOPS, 576 GPU count, H2 2027 availability

• Data Center Dynamics — "Nvidia's Rubin Ultra NVL576 rack expected to be 600kW, coming second half of 2027" (Mar 2026)

• CNBC — Jensen Huang GTC 2026 keynote: $1T orders, Vera Rubin shipping H2 2026 (Mar 16, 2026)

• Awesome Agents — "NVIDIA Vera Rubin Arrives at GTC 2026 With 6 Chips" — confirms seven-chip platform post-update (Mar 2026)

• Lenovo Press — ThinkSystem NVIDIA H200 141GB Product Guide; HGX B200 180GB 1000W Product Guide

• MLCommons — MLPerf Inference v4.0 Llama 2 70B benchmark documentation

• HPCwire — "MLPerf Inference 4.0: H200 up to 45% faster than H100 on Llama 2 70B" (Mar 2024)

• CoreWeave — MLPerf v5.0 H200: 33,000 tokens/sec on Llama 2 70B (2025)

• VideoCardz — NVIDIA Vera Rubin NVL72 Detailed (Jan 2026); Blackwell Ultra GB300 features (Aug 2025)

• Verda — B300 vs B200 Complete Comparison (updated Feb 2026); B200 and B300 Architecture and Software Stack

• Introl — Blackwell Ultra B300 Infrastructure Requirements (Feb 2026); Rubin Full Production (Jan 2026)

• TRG Datacenters — H200 vs H100 Comparison; Blackwell vs Blackwell Ultra B300

• Spheron Network — NVIDIA B200 Complete Guide (FP8 dense/sparse specifications)

• Exxact — Comparing Blackwell vs Hopper (2025)

• Fluence — H200 vs A100: HPC benchmarks (financial risk +40%, molecular dynamics 30-35% vs H100)

• Jarvislabs.ai — NVIDIA H200 Price Guide 2026 (Jan 2026); AWS/Azure pricing confirmation

• Thunder Compute — NVIDIA H200 Price Comparison (Feb 2026); Lambda pricing

• gpu.fm — Cloud GPU Providers Comparison 2026 (Feb 2026) — Lambda $4.49/hr rate

---

All specifications verified against primary sources as of March 2026. Pricing reflects Q1 2026 on-demand rates.