

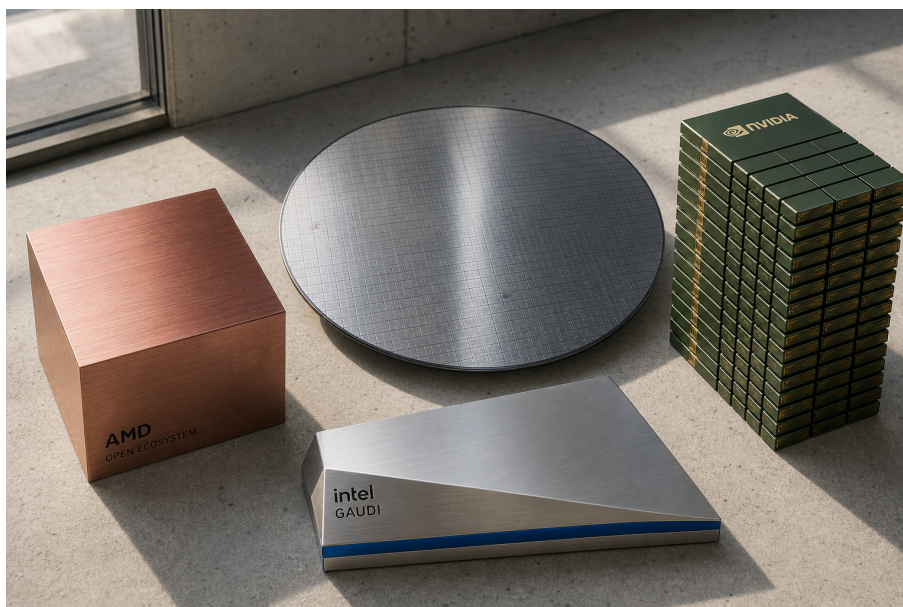


NVIDIA vs AMD vs Intel vs Cerebras AI Accelerator Comparison

Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-07-19 · ai accelerator comparison · nvidia vs amd vs intel · cerebras wse-3 · gpu vs wafer scale engine · mlperf benchmark 2026 · ai chip performance per watt · nvidia h100 b200 · amd instinct mi300x · intel gaudi 3

Original article: <https://gpusmith.com/articles/en/ai-accelerator-comparison-nvidia-amd-intel-cerebras>



In this article

1. Executive Summary
 2. Introduction and Background
 3. NVIDIA: The Data Center Standard
 4. Capabilities
 5. Adoption
 6. Strengths and Limitations
 7. AMD: Instinct and the Open Ecosystem Challenge
 8. Capabilities
 9. Adoption
 10. Strengths and Limitations
 11. Intel: Gaudi and the Price-Performance Contender
 12. Capabilities
 13. Adoption
 14. Strengths and Limitations
 15. Cerebras: Wafer-Scale Computing
 16. Capabilities
 17. Adoption
 18. Strengths and Limitations
 19. Feature Comparison
 20. Performance and Benchmarks
 21. Data Analysis and Evidence
 22. Case Studies and Real-World Examples
 23. xAI's Colossus and the NVIDIA Hopper-to-Blackwell Buildout
 24. Meta and AMD's 6-Gigawatt Instinct Partnership
 25. IBM Cloud and Intel Gaudi 3 for Regulated-Industry Inference
 26. OpenAI, Cerebras, and the Wafer-Scale Inference Bet
 27. Implications and Future Directions
 28. Frequently Asked Questions (FAQs)
 29. Conclusion
-

Executive Summary

The AI accelerator market in 2026 remains dominated by **NVIDIA**, which reported **\$62.3 billion** in quarterly Data Center revenue for its fiscal fourth quarter, up 75% year over year, and **\$193.7 billion** for the full fiscal year, up 68% (nvidianews.nvidia.com) (nvidianews.nvidia.com). Total fiscal 2026 revenue reached **\$215.9 billion**, up 65% year over year (nvidianews.nvidia.com). Independent market research from Mordor Intelligence estimates NVIDIA retained roughly **80%** of global AI training revenue in 2024, with the broader AI accelerator market worth **\$174.69 billion** in 2026 and projected to reach **\$518.12 billion** by 2031, a **24.30%** compound annual growth rate (CAGR) (mordorintelligence.com) (mordorintelligence.com). NVIDIA's flagship

data center platform, the **GB200 NVL72**, links 72 Blackwell Tensor Core GPUs and 36 Grace CPUs into a single rack delivering **30x** faster real-time trillion-parameter large language model (LLM) inference and 25 times more performance per watt than air-cooled Hopper H100 infrastructure ([nvidia.com](https://www.nvidia.com)) ([nvidia.com](https://www.nvidia.com)).

AMD is the most direct GPU-based challenger, reporting Q1 2026 Data Center segment revenue of **\$5.8 billion**, up 57% year over year, driven by EPYC central processing units (CPUs) and Instinct graphics processing units (GPUs) (datacenterdynamics.com). Its **Instinct MI355X** accelerator ships with 288 gigabytes (GB) of HBM3E memory and 8 terabytes per second (TB/s) of bandwidth, exceeding the memory capacity of NVIDIA's B200 (amd.com), and in February 2026 AMD signed a five-year, up to 6-gigawatt (GW) supply partnership with Meta built around custom MI450-based silicon (ir.amd.com). Independent benchmarking by Artificial Analysis found the MI300X achieved **25 to 35%** higher peak system throughput than comparable **NVIDIA H100 and H200 systems** under high-concurrency inference loads, though NVIDIA retained a latency edge at low concurrency (artificialanalysis.ai).

Intel's Gaudi 3 accelerator, priced by industry analysts at roughly **\$16,000** per unit in volume against **\$30,000-plus for an NVIDIA H100**, targets price-conscious inference buyers with 128GB of HBM2e memory and 33% more input/output (I/O) connectivity per accelerator than the H100 (tweaktown.com) (intel.com). Intel's Data Center and AI segment grew to **\$5.1 billion** in Q1 2026 as part of a broader beat that CEO Lip-Bu Tan described as a "sixth consecutive quarter of revenue above our expectations" (intc.com), though official MLPerf Inference v6.0 submissions in April 2026 did not include Gaudi 3 results (spheron.network).

Cerebras Systems takes the most architecturally distinct approach: its third-generation **Wafer-Scale Engine (WSE-3)** is a single 46,225 square millimeter chip carrying 4 trillion transistors and 900,000 AI-optimized cores, delivering 125 petaflops of compute and claimed to offer 19 times more transistors and 28 times more compute than a single **NVIDIA B200** (cerebras.ai) (cerebras.ai). Cerebras completed the largest initial public offering (IPO) of 2026 in May, pricing at a valuation of up to **\$48.8 billion**, up from a **\$23 billion** private round three months earlier (cnbc.com) (cnbc.com), buoyed by an OpenAI inference deal worth more than **\$10 billion** covering 750 megawatts of deployed capacity (cnbc.com) (cnbc.com).

On raw power efficiency, Cerebras's own analysis claims a 2.2 times improvement in performance per watt over a comparable NVIDIA DGX B200 system (cerebras.ai), while an independent, peer-reviewed arXiv comparison finds wafer-scale integration offers genuine advantages in performance per watt and memory scalability but concedes cost-effectiveness at commercial scale remains unproven (arxiv.org). There is no single "best" accelerator for every workload as of July 2026. NVIDIA leads on software maturity, ecosystem breadth, and raw training scale, evidenced by MLPerf Training v6.0 results in which it was "the only platform to submit on every test" (developer.nvidia.com); AMD offers the strongest open-source, memory-per-dollar alternative for inference; Intel competes on acquisition price and standard Ethernet networking; and Cerebras wins outright on single-request inference latency and token throughput for models that fit its architecture. This report examines each vendor's capabilities, adoption, and limitations in detail, presents third-party benchmark and financial evidence, and profiles four real-world deployments to help technical buyers match accelerator choice to workload.

Introduction and Background

Artificial intelligence (AI) accelerators, purpose-built processors designed to speed up the matrix multiplication and tensor operations underlying neural network training and inference, have become the single largest driver of data center capital spending as of mid-2026. What began as a market almost entirely defined by general-purpose graphics processing units (GPUs) has splintered into competing architectural philosophies: NVIDIA's CUDA-centric GPU platform, AMD's open ROCm software stack running on Instinct GPUs, Intel's Ethernet-native Gaudi accelerators, and Cerebras's radically different wafer-scale approach that replaces racks of discrete chips with a single dinner-plate-sized piece of silicon.

The stakes are enormous. NVIDIA alone generated **\$215.9 billion** in total revenue for fiscal 2026, up 65% from the prior year, with Data Center products accounting for the overwhelming majority of that growth (nvidianews.nvidia.com). Mordor Intelligence's industry analysis puts the entire AI accelerator market at **\$174.69 billion** in 2026, up from **\$140.55 billion** in 2025, expanding at a **24.30% CAGR** through 2031 as hyperscale cloud providers, sovereign AI initiatives, and enterprise buyers race to secure compute capacity (mordorintelligence.com). Within that spending, GPUs still account for the majority of revenue, an estimated 59.20% in 2025, while ASICs are catching up at a projected 27.15% CAGR through 2030 (mordorintelligence.com). Cloud and colocation data centers absorb roughly three-quarters of total deployment (74.30% of 2025 spending), while hyperscale cloud service providers alone controlled 52.40% of overall AI accelerator spending in the same year (mordorintelligence.com).

This comparison focuses on four contenders that between them account for the overwhelming majority of merchant (non-captive) AI accelerator silicon shipped for training and inference workloads: **NVIDIA** (Hopper H100/H200 and Blackwell B200/GB200), **AMD** (Instinct MI300X, MI325X, and MI350 series), **Intel** (Gaudi 2 and Gaudi 3), and **Cerebras** (the CS-3 system built around the WSE-3 wafer-scale chip). It deliberately excludes hyperscaler-only, non-merchant silicon such as Google's Tensor Processing Units (TPUs), Amazon's Trainium and Inferentia chips, and Microsoft's Maia, which are not generally available for third-party purchase or lease in the same way. It also notes, where relevant, custom application-specific integrated circuit (ASIC) competition from Broadcom, whose custom AI chip pipeline has driven speculation of a **\$60 billion to \$90 billion** opportunity by 2027 according to Mordor Intelligence's industry tracking, and whose market capitalization CNBC reported at over **\$1.6 trillion** as investor enthusiasm for its custom XPU pipeline grew (mordorintelligence.com) (cnbc.com).

Each vendor is assessed on architecture and capabilities, real-world adoption evidenced by named customer deployments, and documented strengths and limitations, before the report turns to a head-to-head feature matrix, third-party benchmark data (including MLPerf Training v6.0 and Inference v6.0 results published by MLCommons in 2026), financial and market-share evidence, and four detailed case studies. All figures are anchored "as of" their publication or filing date and are drawn from vendor technical documentation, U.S. Securities and Exchange Commission (SEC) filings, peer-reviewed and preprint research, and independent benchmarking organizations, with discrepancies between vendor-reported and independently measured figures flagged explicitly throughout.

NVIDIA: The Data Center Standard

Capabilities

NVIDIA's current data center lineup spans the Hopper-generation **H100** and **H200** GPUs and the newer Blackwell-generation **B200** and rack-scale **GB200 NVL72**. The H100 delivers up to 3,958 teraFLOPS (TFLOPS) of FP8 Tensor Core performance with sparsity, 80GB of HBM3 memory, and 3.35TB/s of memory bandwidth, with a configurable thermal design power (TDP) of up to 700 watts (W) (nvidia.com). NVIDIA states the H100's fourth-generation Tensor Cores and Transformer Engine deliver up to 4 times faster training than the prior generation for GPT-3 175B parameter models, alongside up to 30 times higher inference performance on the largest models compared with the previous Ampere generation. The GB200 NVL72 connects 72 Blackwell GPUs and 36 Grace CPUs in a single liquid-cooled rack using fifth-generation NVLink, providing 130TB/s of low-latency GPU-to-GPU communication across the rack, 13.4TB of pooled HBM3E memory, and 576TB/s of aggregate memory bandwidth (nvidia.com) (nvidia.com). NVIDIA states GB200 NVL72 delivers 4 times faster training and 30 times faster real-time trillion-parameter LLM inference than an equivalent H100 cluster, alongside a claimed 10 times uplift for mixture-of-experts (MoE) architectures. The successor **GB300 NVL72** (Blackwell Ultra) integrates 72 Blackwell Ultra GPUs and 36 Arm-based Grace CPUs and targets up to a 50 times increase in AI factory output versus Hopper-based platforms (nvidia.com).

Adoption

NVIDIA's ecosystem advantage rests on the maturity of its CUDA software stack, which remains the default target for most machine learning research and production frameworks. On the infrastructure side, the clearest evidence of scale is xAI's **Colossus** supercomputer, built with Supermicro and connecting 100,000 NVIDIA Hopper Tensor Core GPUs over NVIDIA Spectrum-X Ethernet networking, described by Supermicro as one of the largest liquid-cooled AI clusters in the world (supermicro.com). In September 2025, NVIDIA said it would commit \$100 billion to support OpenAI's deployment of at least 10 gigawatts of NVIDIA systems, a project CEO Jensen Huang called "a giant project" that will equate to between 4 million and 5 million GPUs (cnbc.com). NVIDIA's own MLPerf Training v6.0 results, published in June 2026, report the platform was "the only platform to submit on every test," including new DeepSeek-V3 671B mixture-of-experts and GPT-OSS-20B benchmarks (developer.nvidia.com), with cloud partners scaling training runs up to 8,192 Blackwell GPUs working in unison across production hyperscale fleets (developer.nvidia.com). NVIDIA also disclosed a new Rubin platform in its fiscal 2026 results, designed to deliver up to a 10 times reduction in inference token cost compared with Blackwell (nvidianews.nvidia.com).

Strengths and Limitations

NVIDIA's principal strength is ecosystem depth: CUDA, cuDNN, TensorRT-LLM, and the broader NVIDIA AI Enterprise software suite give developers a well-trodden path from research to production that competitors still struggle to match. NVIDIA's own data shows the cost-per-token economics improving generationally: the company states H100 delivers inference at approximately \$0.09 per million tokens for a GPT-OSS-120B workload under vLLM, while B200 cuts that to roughly \$0.02 per million tokens under TensorRT-LLM, a 4.5

times improvement, according to SemiAnalysis InferenceX benchmarks NVIDIA cites as of April 2026 ([nvidia.com](#)) ([nvidia.com](#)). The chief limitations are cost, power draw, and supply constraints tied to geopolitics: the April 2025 U.S. export licensing requirement on H20 products forced NVIDIA to take a **\$4.5 billion** charge in a single quarter for excess China-bound inventory, after H20 sales of \$4.6 billion prior to the new requirement and an inability to ship a further \$2.5 billion of H20 revenue in the same quarter ([sec.gov](#)). That exposure has only grown: TrendForce reporting in July 2026 cites Bernstein Research projecting NVIDIA's share of the Chinese AI semiconductor market will fall to around 8% in 2026, down from roughly 40% a year earlier, as Chinese buyers shift toward domestic suppliers such as Huawei ([trendforce.com](#)).

AMD: Instinct and the Open Ecosystem Challenge

Capabilities

AMD's Instinct line is built on the CDNA architecture, with the **MI300X** (launched December 2023) offering 192GB of HBM3 memory, 5.3TB/s of peak memory bandwidth, 750W typical board power, 153 billion transistors, and 19,456 stream processors across 304 compute units ([amd.com](#)) ([amd.com](#)). The newer **MI350 series**, built on 4th Gen CDNA architecture, pushes memory to 288GB of HBM3E with 8TB/s of bandwidth on the flagship MI355X, and an 8-GPU MI350 platform aggregates 2.3TB of total HBM3E memory with 64TB/s of peak theoretical bandwidth ([amd.com](#)) ([amd.com](#)). AMD's own comparisons claim the MI355X delivers up to 2.2 times the AI performance of a comparable B200 SXM5 on select precision formats, and 1.6 times the memory capacity ([amd.com](#)). For buyers who do not need flagship specifications, AMD also offers the PCIe-based MI350P, positioned around lower cost, with 128 GPU compute units, 144GB of HBM3E memory, and up to 4TB/s of peak theoretical memory bandwidth. Software runs on the open-source **ROCm** stack, which AMD positions as a no-licensing-fee alternative to CUDA.

Adoption

AMD's Instinct fleet has moved from niche deployments to hyperscaler-scale commitments in the past two years. At its 2024 Advancing AI event, AMD disclosed that MI300X was "serving all live traffic on Llama 405B" for Meta and had already been deployed by cloud, OEM, and ODM partners serving OpenAI's ChatGPT and Hugging Face workloads ([ir.amd.com](#)), while Databricks reported over 50% performance gains on Llama and its own proprietary models using MI300X's larger memory ([ir.amd.com](#)). AMD CEO Lisa Su used the same event to project "the data center AI accelerator market growing to \$500 billion by 2028" ([ir.amd.com](#)). That relationship deepened materially in February 2026, when AMD and Meta announced a five-year, multi-generation partnership to deploy up to 6 gigawatts of AMD Instinct GPUs, with the first tranche built on a custom MI450-based GPU and AMD's sixth-generation "Venice" EPYC CPUs running on the Helios rack-scale architecture ([ir.amd.com](#)) ([ir.amd.com](#)). OpenAI separately committed to deploying six gigawatts of AMD GPUs across multiple hardware generations starting in the second half of 2026, a deal that came with a warrant for

up to 160 million shares of AMD stock ([cnbc.com](https://www.cnbc.com)). AMD's own case studies also point to cloud infrastructure specialist TensorWave, which the company says built an AMD Instinct GPU cloud "delivering reliable, resilient AI infrastructure" with meaningful performance and cost advantages versus alternatives ([amd.com](https://www.amd.com)).

Strengths and Limitations

AMD's core advantage is memory capacity per dollar: the MI300X's 192GB and the MI355X's 288GB of HBM comfortably exceed the H100's 80GB and even the H200's 141GB, letting a single accelerator host larger models with less multi-GPU partitioning overhead. Independent benchmarking from Artificial Analysis, conducted with AMD's cooperation but using the firm's own published methodology, found the MI300X achieved 25 to 35% higher peak system output throughput than comparable NVIDIA H100 and H200 systems at high concurrency on DeepSeek R1 and Llama 4 Maverick workloads, while NVIDIA retained an edge in per-query latency at low concurrency (artificialanalysis.ai) (artificialanalysis.ai). The persistent limitation remains software maturity. Developer sentiment on Reddit's r/MachineLearning community, while not a benchmark, is illustrative of the gap AMD is closing: one widely upvoted comment noted that "properly written and optimized ROCm code is just as fast as Cuda, right below whatever the maximum number of tflops of your GPU is," but cautioned that most published kernels, including flash-attention implementations, are written and tuned for CUDA first ([reddit.com](https://www.reddit.com)). Another commenter in the same thread argued the software gap reflects hiring and investment, not architecture, writing that AMD "is not serious about hiring talent" relative to NVIDIA, OpenAI, and Meta ([reddit.com](https://www.reddit.com)). AMD has moved to close this gap through acquisitions, including the **\$4.9 billion** purchase of ZT Systems in August 2024 to build systems-integration expertise and the **\$665 million** acquisition of Silo AI the same month for multilingual model-development capabilities, per Mordor Intelligence's tracking of both deals (mordorintelligence.com).

Intel: Gaudi and the Price-Performance Contender

Capabilities

Intel's **Gaudi 3** accelerator, generally available since mid-2024, is built for open, Ethernet-native scaling rather than NVIDIA's proprietary NVLink/NVSwitch interconnects. Intel states Gaudi 3 offers 33% more I/O connectivity per accelerator than an H100 ([intel.com](https://www.intel.com)) and delivers 2 times the FP8 AI compute, 4 times the BF16 (16-bit floating point) compute, and 2 times the network bandwidth of the prior-generation Gaudi 2 accelerator ([intel.com](https://www.intel.com)). The accelerator ships in mezzanine, PCIe, and OAM (open accelerator module) form factors with a TDP that scales from 450W up to 900W for water-cooled OAM configurations ([tweaktown.com](https://www.tweaktown.com)). Intel markets Gaudi's avoidance of proprietary interconnect lock-in as a core selling point for buyers who want to reuse existing Ethernet infrastructure.

Adoption

Intel's most concrete enterprise deployment is with **IBM Cloud**, the first cloud service provider to offer Gaudi 3 accelerators as a service to enterprise customers, with support for IBM's watsonx AI platform following in Q2 2025 ([intel.com](https://www.intel.com)). IBM testing showed a single Gaudi 3 card generating over 5,000 tokens per second on IBM's granite-8b model while supporting more than 100 concurrent users with sub-20-millisecond inter-token latency, and the platform is designed to scale from a single 8-accelerator node at 9.6TB/s of throughput up to a 1,024-node cluster of 8,192 accelerators delivering 9.830 petabytes per second (PB/s) of aggregate throughput ([intel.com](https://www.intel.com)). To accelerate that adoption, Intel and IBM have run promotional pricing offering customers a discount on Gaudi 3 accelerators on IBM Cloud so they can "run twice as long for the same price" ([intel.com](https://www.intel.com)). Intel's Data Center and AI segment revenue reached **\$5.1 billion** in Q1 2026, part of an overall quarterly beat that CEO Lip-Bu Tan attributed to a "sixth consecutive quarter of revenue above our expectations" ([intc.com](https://www.intc.com)).

Strengths and Limitations

Intel's clearest differentiator is acquisition price: industry analyst estimates cited by trade press put bulk per-unit Gaudi 3 pricing at approximately **\$16,000**, versus **\$30,000-plus** for an NVIDIA H100, a roughly 2 times discount for comparable memory-bound inference workloads ([tweaktown.com](https://www.tweaktown.com)) ([tweaktown.com](https://www.tweaktown.com)). Gaudi's reliance on standard Ethernet rather than proprietary interconnects also lowers switching and networking costs for buyers who do not want NVLink lock-in, per Intel's own positioning. The central limitation is competitive visibility: Intel's official MLPerf Inference v6.0 submissions in April 2026 covered Xeon 6 CPUs and Arc Pro GPUs rather than Gaudi 3, meaning no directly comparable, independently audited per-accelerator throughput figures for Gaudi 3 exist in the latest MLCommons dataset ([spheron.network](https://www.spheron.network)). This absence of recent third-party benchmark data, combined with revenue that remains a fraction of NVIDIA's, means buyers evaluating Gaudi 3 must weigh cost savings against a comparatively thin body of recent, standardized performance evidence.

Cerebras: Wafer-Scale Computing

Capabilities

Cerebras takes a fundamentally different design philosophy. Rather than networking many discrete chips, the **Wafer-Scale Engine 3 (WSE-3)** is fabricated as a single piece of silicon measuring 46,225 square millimeters, the largest AI chip ever built, containing 4 trillion transistors and 900,000 AI-optimized cores ([cerebras.ai](https://www.cerebras.ai)). An independent, peer-reviewed comparison published on arXiv in March 2025 confirms the WSE-3 integrates "900,000 AI-optimized cores, and 44 GB of on-chip SRAM," achieving "a peak computing performance of 125 petaflops" and a memory bandwidth of 21 petabytes per second (arxiv.org) (arxiv.org). This is the third generation of the architecture: the original 2019 WSE-1 held over 1.2 trillion transistors, 400,000 AI-optimized cores, and 18GB of high-speed SRAM with 9 petabytes per second of memory bandwidth, while the WSE-2 grew to 2.6 trillion transistors, 850,000 cores, and 40GB of on-chip SRAM with 20 petabytes per second of memory bandwidth and 220 petabits per second of fabric bandwidth. The wafer is packaged into the **CS-3**

system, which decouples compute from memory using an external "MemoryX" appliance that scales from 1.2 terabytes up to 1.2 petabytes, allowing a single CS-3 to hold models with up to 24 trillion parameters without the multi-GPU partitioning schemes GPU clusters require ([arxiv.org](#)).

Adoption

Cerebras's commercial momentum accelerated sharply in the first half of 2026. In January, OpenAI signed an agreement to deploy 750 megawatts of Cerebras AI chips in tranches through 2028, a deal worth more than **\$10 billion**, with Cerebras CEO Andrew Feldman calling it a way of "bringing the world's leading AI models to the world's fastest AI processor" ([cnbc.com](#)) ([cnbc.com](#)). CNBC reported Cerebras's wafer-scale chips can "deliver responses up to 15 times faster than GPU-based systems" according to the company's own release accompanying the OpenAI deal ([cnbc.com](#)). Separately, Amazon Web Services announced a deal to bring Cerebras chips into AWS data centers alongside Amazon's own Trainium3 silicon ([cnbc.com](#)). These commitments underpinned Cerebras's IPO, which Nasdaq scheduled for May 14, 2026, pricing at a valuation of up to **\$48.8 billion**, more than double the **\$23 billion** private valuation the company had announced only three months earlier ([cnbc.com](#)) ([cnbc.com](#)).

Strengths and Limitations

Cerebras's headline strength is raw single-model inference speed. In its own benchmarking against a 405-billion-parameter Llama 3.1 model, Cerebras reported output of 969 tokens per second, describing the result as "12x faster than best GPU result," at a published price of \$6 per million input tokens and \$12 per million output tokens ([cerebras.ai](#)) ([cerebras.ai](#)). On power efficiency, Cerebras's own head-to-head comparison against an NVIDIA DGX B200 found "the CS-3 consumes 23kW peak while the DGX B200 consumes 14.3 kW," but delivers 125 petaflops versus 36 petaflops for the DGX B200, which Cerebras calculates as "a 2.2x improvement in performance per watt" ([cerebras.ai](#)) ([cerebras.ai](#)) ([cerebras.ai](#)). The chief limitation is cost and manufacturing complexity: the independent arXiv analysis estimates a single CS-3 system costs approximately \$2 million to \$3 million, versus roughly \$350,000 for an NVIDIA DGX H100 and \$500,000 for a DGX B200, a 5 to 7 times price premium per system that only pencils out for buyers whose workloads specifically benefit from wafer-scale memory bandwidth and single-device simplicity ([arxiv.org](#)). The same paper notes each CS-3 server "is said to consume 23KW" of power, a figure that complicates deployment in facilities not already provisioned for high-density liquid cooling ([arxiv.org](#)). Cerebras's prior fundraising history also illustrates customer-concentration risk: CNBC reported the company withdrew a 2025 IPO attempt after its prospectus revealed "heavy reliance" on a single customer, Microsoft-backed G42, which was simultaneously a major buyer and a Cerebras investor, before the OpenAI and AWS deals diversified that base ahead of the successful 2026 listing ([cnbc.com](#)).

Feature Comparison

Table 1 below summarizes the core specifications of each vendor's current flagship data center accelerator, drawing on official vendor datasheets, product pages, and the independent arXiv architectural comparison.

Specification	NVIDIA GB200 NVL72 (per GPU: B200)	AMD Instinct MI355X	Intel Gaudi 3	Cerebras WSE-3 (CS-3 system)
Architecture	Blackwell, TSMC 4nm process	CDNA 4, TSMC process	Gaudi 3, matrix and tensor cores	Wafer-scale, TSMC 5nm process
Peak AI compute	Rack: 1,440 PFLOPS NVFP4 (sparse)	Up to 10.1 PFLOPS FP8 (sparsity) per GPU	Up to 1.8 PFLOPS FP8/BF16	125 petaflops per WSE-3 chip (arxiv.org)
Memory capacity	13.4TB HBM3E across rack	288GB HBM3E per GPU	128GB HBM2e per accelerator	44GB on-chip SRAM plus 1.2TB to 1.2PB external MemoryX
Memory bandwidth	576TB/s across rack	8TB/s per GPU	3.7TB/s per accelerator	21 petabytes/s on-wafer
Interconnect	130TB/s NVLink Switch across 72 GPUs	Infinity Fabric, 128GB/s per link	Standard Ethernet (RoCE)	On-wafer 2D mesh (no external interconnect needed)
Approximate system power	14.3kW (DGX B200, 8 GPUs) (arxiv.org)	750W per MI300X GPU (amd.com)	Up to 900W (water-cooled OAM)	23kW per CS-3 system (arxiv.org)
Approximate list price	~\$500,000 (DGX B200, 8 GPUs, estimate)	Not publicly listed (OEM/hyperscale contracts)	~\$16,000 per accelerator in volume (tweaktown.com)	~\$2M to \$3M per CS-3 system (estimate) (arxiv.org)
Software stack	CUDA, TensorRT-LLM, NVIDIA AI Enterprise	ROCm (open source)	SynapseAI, PyTorch integration	Cerebras SDK, layer-by-layer execution model

Table 1 makes the divergence in design philosophy explicit; the underlying specification figures are drawn from the NVIDIA GB200 NVL72 datasheet, AMD's Instinct MI350 series and MI300X product pages, Intel's Gaudi product page, and the independent arXiv wafer-scale comparison cited throughout the Capabilities sections above. NVIDIA and AMD compete primarily on a per-GPU, rack-aggregated basis, where discrete accelerators are networked together and the interconnect fabric becomes a first-order performance variable, whether that is NVIDIA's proprietary 130TB/s NVLink Switch or AMD's Infinity Fabric. Intel's Gaudi 3 undercuts both on acquisition price while trading away proprietary interconnect performance for standard Ethernet compatibility. Cerebras inverts the entire model: instead of scaling out across many chips, it scales a single chip to the size of an entire silicon wafer, eliminating inter-chip communication for workloads that fit within one WSE-3, at the cost of a substantially higher unit price and power draw per system. Budget-tier options exist on every vendor's roadmap, from AMD's MI350P PCIe card (144GB HBM3E, 4TB/s bandwidth) to Intel's lower-wattage 450W Gaudi 3 configurations, giving buyers a way to trade peak throughput for lower acquisition and operating cost within a single vendor family.

Performance and Benchmarks

MLCommons, the industry consortium behind the MLPerf benchmark suite, published **MLPerf Training v6.0** results in June 2026 and **MLPerf Inference v6.0** results in April 2026, providing the most current independently audited data on training and inference throughput. NVIDIA submitted results across every benchmark and reported the fastest time-to-train at scale on every category it entered, including a DeepSeek-V3 671B mixture-of-experts model trained in 2.02 minutes on a GB300 NVL72 cluster of 8,192 GPUs ([developer.nvidia.com](#)). NVIDIA also reported that software optimizations alone lifted GB300 training throughput on DeepSeek-V3 from 1,298 to 1,648 TFLOPS per GPU in three months without any hardware changes, describing "DeepSeek-V3 training throughput increasing by 1.3x in just three months" through kernel fusion, CUDA graph, and communication-overlap improvements ([developer.nvidia.com](#)).

On the inference side, an April 2026 analysis of MLPerf Inference v6.0 results derived per-GPU throughput by dividing system-level scores by accelerator count. *Table 2* below summarizes the closed-division, per-GPU results for the GPT-OSS 120B and LLaMA 2 70B language model benchmarks, plus the Stable Diffusion XL (SDXL) image-generation and YOLOv11 object-detection tasks.

Accelerator	GPT-OSS 120B (tok/s per GPU)	LLaMA 2 70B (tok/s per GPU)	SDXL (samples/s per GPU)	YOLOv11 (QPS per GPU)
NVIDIA B200 (GB200 NVL72)	~7,800	~17,500	~14.2	~28,400
NVIDIA H200 SXM	~3,100	~7,800	~6.3	~15,200
AMD MI355X (closed division)	~2,600	~6,200	~4.8 (open division)	~11,800 (open division)

Source: MLCommons Inference v6.0, closed division unless noted, as analyzed by Spheron Network in April 2026 ([spheron.network](#)).

On the GPT-OSS-120B benchmark, NVIDIA's B200 delivered roughly 2.5 times the per-GPU throughput of the H200, a gap the analysis attributes primarily to the combination of FP4 precision support and 8.0TB/s of HBM3e bandwidth versus the H200's 4.8TB/s ([spheron.network](#)). AMD's closed-division submission, however, put the MI355X within 20% of the H200 on both LLaMA 2 70B and GPT-OSS-120B, figures the analysis notes are directly comparable because both vendors ran under identical model and precision constraints, with AMD's larger 288GB memory capacity allowing bigger batch sizes on models that fit within it ([spheron.network](#)). Notably, Intel's official v6.0 submissions covered Xeon 6 CPUs and Arc Pro GPU workloads rather than Gaudi 3, leaving a gap in the most current independently audited inference dataset for Intel's flagship AI accelerator ([spheron.network](#)).

Independent, vendor-adjacent benchmarking from Artificial Analysis offers a complementary view focused on realistic production loads rather than the standardized MLPerf harness. Using its own System Load Test methodology on DeepSeek R1 and Llama 4 Maverick across matched 8-GPU NVIDIA H100, H200, and AMD MI300X systems, Artificial Analysis found NVIDIA held an advantage in per-query output speed and latency at low concurrency, while AMD's MI300X sustained near-perfect response rates all the way to a higher peak

system throughput, translating into a lower implied cost per token at scale ([artificialanalysis.ai](#)). For accuracy, the same study found no statistically significant difference between AMD and NVIDIA systems on GPQA, MMLU-Pro, and MATH-500 evaluations at a 95% confidence interval, indicating that hardware choice among GPU vendors does not, on its own, materially affect model output quality.

Cerebras does not participate in the standard MLPerf Inference harness, instead publishing benchmarks independently verified by Artificial Analysis. On Llama 3.1 405B, Cerebras reported 969 output tokens per second and a 240-millisecond time to first token, roughly 12 times faster than the best comparable GPU-based result at the time of publication, and 8 times faster than SambaNova and 75 times faster than AWS on the same evaluation ([cerebras.ai](#)). Because Cerebras, NVIDIA, and AMD use different benchmarking harnesses, Cerebras favoring single-request latency and token throughput and MLCommons favoring standardized offline and server scenarios, buyers evaluating "best ai accelerator for training and inference" claims should treat vendor-reported multiples cautiously and, where possible, request workload-specific proof-of-concept testing before committing to a platform.

Data Analysis and Evidence

The quantitative case for NVIDIA's continued market leadership rests on both revenue and independent market-share estimation. NVIDIA's fiscal 2026 Data Center segment generated **\$193.7 billion**, up 68% year over year, with the fourth quarter alone contributing a record **\$62.3 billion**, up 75% from the prior year ([nvidianews.nvidia.com](#)) ([nvidianews.nvidia.com](#)). Mordor Intelligence's independent industry analysis separately estimates NVIDIA held approximately 80% of global AI training revenue in 2024, a concentration the research firm attributes to "CUDA lock-in, integrated software libraries, and a mature partner ecosystem" ([mordorintelligence.com](#)). Even accounting for accelerating diversification toward AMD, Intel, and custom silicon, GPUs as a processor category retained an estimated 59.20% revenue share of the overall AI accelerator market in 2025, with the balance split between application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), and CPU and neural processing unit (NPU) hybrids ([mordorintelligence.com](#)).

Table 3 below assembles the most recent quarterly financial data for the three publicly traded vendors, alongside Cerebras's newly disclosed public-market valuation, to illustrate the scale gap that still separates NVIDIA from its challengers even as those challengers post strong percentage growth.

Company	Most Recent Quarterly Data Center / AI Revenue	Year-over-Year Growth	Source Period
NVIDIA	\$62.3 billion (Data Center, Q4 FY2026) (nvidianews.nvidia.com)	+75%	Quarter ended January 25, 2026
AMD	\$5.8 billion (Data Center segment, Q1 2026), \$10.25 billion total company revenue (datacenterdynamics.com)	+57% (Data Center); +38% (total)	Quarter ended late March 2026
Intel	\$5.1 billion (Data Center and AI segment, Q1 2026); \$13.6 billion total revenue (intc.com)	+22% (DCAI); +7% (total)	Quarter ended March 28, 2026
Cerebras	Not yet reported as a public company; \$48.8 billion IPO valuation (cnbc.com)	Valuation up from \$23 billion three months prior	IPO priced May 14, 2026

Table 3 shows NVIDIA's single quarter of Data Center revenue exceeds the combined trailing quarterly Data Center or AI revenue of AMD and Intel by more than a factor of five, underscoring why analysts continue to describe the market as NVIDIA-led rather than genuinely multipolar, even as AMD's 57% and Intel's 22% year-over-year segment growth rates indicate real share gains are occurring at the margin, and AMD's own guidance for the following quarter called for total revenue of \$11.2 billion (datacenterdynamics.com). On power efficiency specifically, the most rigorous side-by-side data point comes from the independent arXiv comparison, which normalizes NVIDIA and Cerebras hardware to equivalent rack space and finds a CS-3-based rack (2 systems, 46kW total) delivers substantially more raw FLOPS per rack than a comparably sized H100 rack (32 GPUs, 41.6kW) or B200 rack (24 GPUs, 43.9kW), even though the CS-3 rack draws marginally more total power (arxiv.org). The paper's abstract concludes the results highlight the advantages of wafer-scale integration in performance per watt and memory scalability, while cautioning that "work is required to address cost-effectiveness and long-term viability" of the approach at commercial scale (arxiv.org).

On the training-versus-inference split that determines which architecture matters most for a given buyer, Mordor Intelligence estimates training consumed 57.30% of 2025 AI accelerator revenue, while inference spending is forecast to grow at a faster 26.10% CAGR through 2031 as more models move from research into always-on production serving (mordorintelligence.com). That shift favors architectures optimized for latency and cost-per-token, a category where Cerebras's inference-focused wafer-scale design and AMD's memory-dense MI300X and MI350 series are increasingly positioned to compete, even where they trail NVIDIA on aggregate training throughput. Beyond the four vendors profiled here, broader industry data underlines both the scale and the geographic concentration of the AI accelerator buildout: North America commanded a 43.50% share of the global AI accelerator market in 2025, while Asia-Pacific posted the fastest growth at a 27.00% CAGR, driven partly by Chinese firms raising their domestic AI chip budget share from 30% to 46% amid the shift away from NVIDIA described above (trendforce.com).

Case Studies and Real-World Examples

xAI's Colossus and the NVIDIA Hopper-to-Blackwell Buildout

xAI's Colossus supercomputer, built in partnership with Supermicro, connects 100,000 NVIDIA Hopper Tensor Core GPUs using NVIDIA's Spectrum-X Ethernet networking platform, and Supermicro describes it as designed "to take xAI's Grok AI to another era" through liquid-cooled, direct-to-chip cooling at unprecedented density (supermicro.com). The cluster illustrates both the scale NVIDIA's ecosystem can support and the infrastructure demands that scale imposes: 100,000 H100-class GPUs at up to 700W TDP each implies tens of megawatts of continuous power draw before accounting for cooling and networking overhead, a demand profile that has become a defining constraint on where hyperscale AI training clusters can physically be sited. The Colossus build also demonstrates the practical value of NVIDIA's Spectrum-X Ethernet and NVLink ecosystem at scale, since coordinating 100,000 GPUs across a single logical training job depends on the kind of mature networking and orchestration software that remains NVIDIA's most defensible advantage against GPU-based challengers.

Meta and AMD's 6-Gigawatt Instinct Partnership

In February 2026, AMD and Meta announced what AMD called a "definitive multi-year, multi-generation partnership to deploy up to 6 gigawatts of AMD Instinct GPUs," with the first gigawatt of shipments, based on a custom MI450-architecture GPU paired with AMD's sixth-generation "Venice" EPYC CPUs, scheduled to begin in the second half of 2026 (ir.amd.com). Meta CEO Mark Zuckerberg framed the agreement around compute diversification, stating the company was "excited to form a long-term partnership with AMD to deploy efficient inference compute" as part of a broader strategy to avoid single-vendor dependency. AMD structured the deal with a performance-based warrant for up to 160 million shares of AMD common stock, vesting as Meta's Instinct GPU purchases scale from the initial 1 gigawatt toward the full 6-gigawatt commitment (ir.amd.com). This deal built on Meta's existing production use of MI300X, which by late 2024 was already "serving all live traffic on Llama 405B" according to AMD's own disclosure (ir.amd.com), making Meta one of the clearest examples of a hyperscaler moving an AMD accelerator from pilot to primary production inference infrastructure at gigawatt scale.

IBM Cloud and Intel Gaudi 3 for Regulated-Industry Inference

IBM Cloud became the first cloud service provider to offer Intel Gaudi 3 accelerators as a service, targeting enterprise customers in highly regulated industries such as financial services, government, healthcare, and telecommunications who require hybrid cloud deployment options and strict AI governance (intel.com). IBM's own internal testing found a single Gaudi 3 card could generate over 5,000 tokens per second on IBM's granite-8b model while supporting over 100 concurrent users at under 20 milliseconds of inter-token latency, and the architecture is designed to scale linearly from an 8-accelerator single node up to a 1,024-node, 8,192-accelerator cluster capable of 9.830 petabytes per second of aggregate throughput (intel.com). This case demonstrates the specific niche Gaudi 3 is chasing: cost-conscious enterprise inference deployment inside a hybrid cloud, rather than frontier-scale training, where IBM's messaging emphasizes total cost of ownership and an open ecosystem over raw peak performance.

OpenAI, Cerebras, and the Wafer-Scale Inference Bet

In January 2026, OpenAI signed an agreement to deploy 750 megawatts of Cerebras wafer-scale chips in tranches through 2028, a deal CNBC valued at more than \$10 billion (cnbc.com). Cerebras's release describing the partnership claimed its chips can "deliver responses up to 15 times faster than GPU-based systems," positioning the deployment specifically around a model that "writes code," a latency-sensitive workload where Cerebras's single-request throughput advantage is most pronounced (cnbc.com). This case is notable for its context: Cerebras had previously withdrawn a planned 2025 IPO amid investor concern over "heavy reliance" on a single customer, Microsoft-backed G42, which was both a major buyer and a Cerebras investor (cnbc.com). The OpenAI deal, layered on top of the subsequent AWS partnership, materially diversified Cerebras's customer concentration in the months immediately before its successful May 2026 IPO, illustrating how a single marquee inference deployment can reshape a chip vendor's fundraising trajectory as much as any benchmark result.

Implications and Future Directions

The near-term trajectory of the AI accelerator market points toward continued architectural specialization rather than convergence on a single winning design. NVIDIA's own roadmap disclosures suggest this: the company's Rubin platform, unveiled alongside fiscal 2026 results, is designed to deliver "up to a 10x reduction in inference token cost, compared with the NVIDIA Blackwell platform," an explicit acknowledgment that inference economics, not just raw training throughput, now drive purchasing decisions at the margin (nvidianews.nvidia.com). Mordor Intelligence's forecast that inference will grow at a 26.10% CAGR through 2031, faster than training's growth rate, reinforces the expectation that future accelerator generations across all four vendors will increasingly optimize for cost-per-token and latency rather than peak FLOPS alone.

Power availability is emerging as an equally consequential constraint as raw silicon supply. Both the AMD-Meta and OpenAI-Cerebras deals are structured and reported in gigawatts rather than unit counts, a framing shift that reflects the reality that data center electricity capacity, not chip fabrication, is now frequently the binding constraint on deployment timelines. NVIDIA's own \$100 billion OpenAI commitment is likewise denominated in gigawatts, with the company estimating 10 gigawatts equates to 4 million to 5 million GPUs (cnbc.com). This dynamic advantages accelerators with superior performance-per-watt, a metric where Cerebras's arXiv-verified claims and NVIDIA's own claimed per-watt improvement for GB200 over air-cooled H100 infrastructure, cited earlier in this report, will likely become as commercially important as headline FLOPS figures.

Geopolitical and supply-chain risk will also continue to shape vendor strategy. NVIDIA's \$4.5 billion H20 export-control charge in a single quarter of fiscal 2026 demonstrates how quickly regulatory changes can disrupt even the market leader's revenue recognition (sec.gov), and TrendForce's July 2026 reporting that Bernstein projects NVIDIA's Chinese AI semiconductor market share falling to roughly 8% in 2026, with domestic champion Huawei's share predicted to surpass 50%, creates an opening for both Chinese domestic accelerators and, indirectly, for AMD, Intel, and Cerebras to capture non-China demand that might otherwise have concentrated further with NVIDIA (trendforce.com). Finally, custom silicon from hyperscalers and Broadcom's ASIC design partnerships, forecast by Mordor Intelligence to represent a \$60 billion to \$90 billion opportunity by 2027 for a company CNBC now values at over \$1.6 trillion, will continue to erode merchant GPU share at the margin, particularly for steady-state inference workloads where a purpose-built ASIC's efficiency advantage outweighs the flexibility of a general-purpose GPU (mordorintelligence.com) (cnbc.com).

Frequently Asked Questions (FAQs)

Which AI accelerator has the best power efficiency in 2026? On a raw performance-per-watt basis for training, Cerebras's own analysis claims the CS-3 delivers roughly 2.2 times the performance per watt of an NVIDIA DGX B200 (cerebras.ai), while NVIDIA claims its GB200 NVL72 delivers 25 times more performance at the same power as air-cooled H100 infrastructure specifically. Because these are vendor-reported, non-identical comparisons (one against a same-generation GPU competitor, the other against the vendor's own prior generation), buyers should treat both as directionally informative rather than directly comparable, and should request workload-specific power measurements before finalizing procurement.

How does NVIDIA H100 compare to AMD MI300X and Cerebras WSE-3 in practice? The H100 offers 80GB of memory and up to 3,958 TFLOPS of FP8 compute with sparsity at up to 700W ([nvidia.com](https://www.nvidia.com)), while the MI300X offers 192GB of memory at 750W ([amd.com](https://www.amd.com)) and independent testing shows it wins on high-concurrency throughput while the H100 wins on low-concurrency latency (artificialanalysis.ai). Cerebras's WSE-3 is not a like-for-like comparison at the chip level since a single WSE-3 replaces an entire rack of GPUs; the arXiv comparison paper found a CS-3-equipped rack outperforms both H100 and B200 racks on raw compute density while costing several times more per system (arxiv.org).

Is Intel Gaudi competitive with NVIDIA and AMD for AI training and inference? Gaudi 3 is positioned primarily for inference and price-sensitive deployments rather than frontier-scale training, with an approximately 2 times lower acquisition cost than an H100 ([twaktown.com](https://www.twaktown.com)), but the absence of Gaudi 3 results in the most recent MLPerf Inference v6.0 dataset means buyers cannot currently benchmark it against NVIDIA and AMD using the same independently audited methodology (spheron.network).

What is the Cerebras Wafer-Scale Engine and how does it differ from a GPU? The WSE-3 is a single 46,225 square millimeter chip fabricated from an entire silicon wafer, carrying 900,000 cores and 44GB of on-chip memory, versus a GPU's die size of roughly 800 to 1,600 square millimeters ([cerebras.ai](https://www.cerebras.ai)) (arxiv.org). Because all cores communicate over an on-wafer mesh rather than external interconnects, the WSE-3 eliminates a major source of latency and power overhead present in multi-GPU clusters, at the cost of a single point of failure per wafer and a substantially higher unit price.

How big is the overall AI accelerator market, and who is growing fastest? Mordor Intelligence estimates the market at \$174.69 billion in 2026, growing to \$518.12 billion by 2031 at a 24.30% CAGR (mordorintelligence.com). By percentage growth, AMD's 57% year-over-year Data Center revenue increase and Intel's 22% Data Center and AI increase in Q1 2026 both outpaced NVIDIA's underlying quarterly growth rate on a relative basis, even though NVIDIA still adds more absolute revenue in a single quarter than either competitor generates in several (datacenterdynamics.com).

Which accelerator is best for training versus inference? Training workloads at frontier scale remain best served by NVIDIA's GB200/GB300 NVL72 platforms given their independently verified MLPerf Training v6.0 leadership and mature distributed-training software stack (developer.nvidia.com). For latency-sensitive, single-model inference serving, Cerebras's CS-3 and AMD's memory-dense MI300X/MI350 series offer compelling cost and speed advantages, per the vendor and independent benchmarks cited throughout this report, while Intel's Gaudi 3 remains the lowest-cost entry point for enterprises prioritizing budget over peak throughput.

Conclusion

As of July 2026, the "nvidia vs amd vs intel vs cerebras" comparison resolves less into a single winner than into four distinct value propositions matched to different buyer priorities. NVIDIA remains the default choice for organizations that need the broadest software ecosystem, the largest verified training scale, and access to the newest architectures first, backed by fiscal 2026 Data Center revenue of \$193.7 billion and an estimated 80% share of global AI training revenue. AMD has emerged as the most credible GPU-based alternative, winning gigawatt-scale commitments from Meta and OpenAI on the strength of memory-dense Instinct accelerators and an increasingly capable open-source ROCm stack, even as CUDA's software head start persists as a real,

if narrowing, gap. Intel's Gaudi 3 occupies a genuine niche among cost-conscious enterprise buyers, particularly in regulated industries served through partners like IBM Cloud, though its absence from the latest independently audited MLPerf Inference results leaves a meaningful evidence gap for buyers who prioritize standardized benchmarking. Cerebras has translated its architecturally unique wafer-scale approach into rapid public-market validation, closing a more than \$10 billion OpenAI deal and completing 2026's largest IPO at a valuation of up to \$48.8 billion, built on demonstrated single-model inference speed rather than training-cluster scale.

For technical buyers, the practical takeaway is that accelerator selection should be workload-driven rather than brand-driven. Frontier pretraining runs at trillion-parameter scale still favor NVIDIA's NVLink-connected GPU racks; memory-bound inference at high concurrency increasingly favors AMD's larger HBM capacity; budget-constrained enterprise inference deployments inside existing Ethernet infrastructure favor Intel's Gaudi 3; and latency-critical, single-model serving at extreme token throughput favors Cerebras's wafer-scale systems. Given how quickly benchmark leadership has shifted across MLPerf Training v6.0 and Inference v6.0 rounds within a single year, buyers should treat any static comparison, including this one, as a snapshot that warrants periodic re-verification against the latest MLCommons results and vendor financial disclosures before committing to multi-year procurement contracts measured in gigawatts rather than individual chips.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <http://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2026>
2. <https://www.mordorintelligence.com/industry-reports/ai-accelerators-market>
3. <https://www.nvidia.com/en-us/data-center/gb200-nv172/>
4. <https://www.datacenterdynamics.com/en/news/amd-posts-q1-2026-data-center-revenue-of-58bn-forecasts-120bn-server-cpu-income-by-2030/>
5. <https://www.amd.com/en/products/accelerators/instinct/mi350.html>
6. <https://ir.amd.com/news-events/press-releases/detail/1279/amd-and-meta-announce-expanded-strategic-partnership-to-deploy-6-gigawatts-of-amd-gpus>
7. <https://gpusmith.com/articles/gpu-cloud-rental-prices-h100-h200>
8. <https://artificialanalysis.ai/articles/independent-analysis-of-leading-gpus-amd-nvidia>
9. <https://gpusmith.com/articles/own-vs-rent-gpus-tco-comparison>
10. <https://www.tweaktown.com/news/98977/intel-discounting-new-gaudi-3-ai-accelerator-16k-against-nvidia-h100-gpu-for-30k/index.html>
11. <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi.html>
12. <https://www.intc.com/news-events/press-releases/detail/1767/intel-reports-first-quarter-2026-financial-results>
13. <https://www.spheron.network/blog/mlperf-inference-v6-benchmark-results-2026/>
14. <https://www.cerebras.ai/chip>
15. <https://www.cnbc.com/2026/05/11/cerebras-raises-ipo-range.html>
16. <https://www.cnbc.com/2026/01/16/openai-chip-deal-with-cerebras-adds-to-roster-of-nvidia-amd-broadcom.html>
17. <https://www.cerebras.ai/blog/cerebras-cs-3-vs-nvidia-b200-2024-ai-accelerators-compared>

18. <https://arxiv.org/html/2503.11698v1>
19. <https://developer.nvidia.com/blog/nvidia-blackwell-tops-mlperf-training-6-0-with-industry-leading-scale-and-performance/>
20. <https://www.nvidia.com/en-us/data-center/h100/>
21. <https://www.supernmicro.com/en/featured/xai-colossus>
22. <https://www.sec.gov/Archives/edgar/data/1045810/000104581025000115/q1fy26pr.htm>
23. <https://www.trendforce.com/news/2026/07/07/news-chinese-firms-reportedly-raise-domestic-ai-chip-budget-share-from-30-to-46-amid-shift-from-nvidia/>
24. <https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html>
25. <https://ir.amd.com/news-events/press-releases/detail/1218/amd-unveils-leadership-ai-solutions-at-advancing-ai-2024>
26. https://www.reddit.com/r/MachineLearning/comments/1fa8vq5/d_why_is_cuda_so_much_faster_than_rocm/
27. <https://www.intel.com/content/www/us/en/customer-spotlight/stories/ibm-gaudi-3-customer-story.html>
28. <https://www.cerebras.ai/blog/llama-405b-inference>

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the [hardware library](#), [GPU references](#), [networking references](#) and [vendor research](#). [Contact GPU Smith](#) with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: <https://gpusmith.com>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.