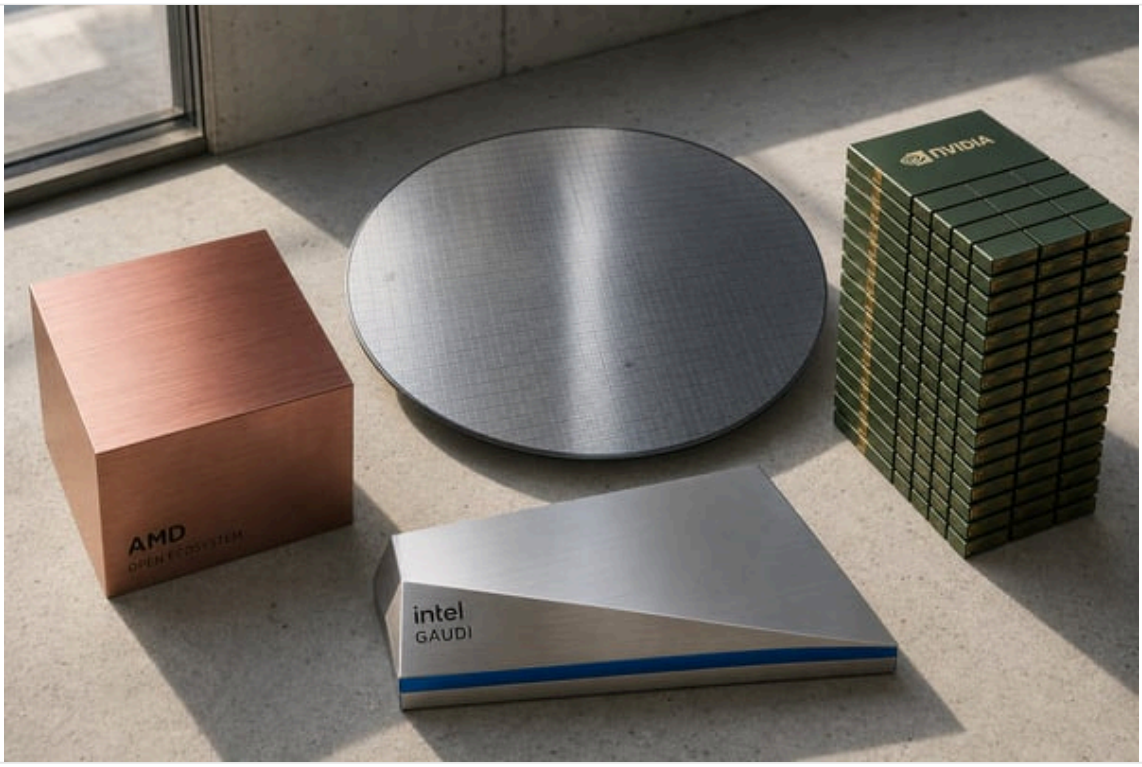
NVIDIA vs AMD vs Intel vs Cerebras AI Accelerator Comparison

Published July 10, 2026 36 min read



```
# Executive Summary

The AI accelerator market in 2026 remains dominated by **NVIDIA**, which reported **\$62.3 billion** in quarterly Data Ce

**AMD** is the most direct GPU-based challenger, reporting Q1 2026 Data Center segment revenue of **\$5.8 billion**, up 5

**Intel's Gaudi 3** accelerator, priced by industry analysts at roughly **\$16,000** per unit in volume against [\$30,000

**Cerebras Systems** takes the most architecturally distinct approach: its third-generation **Wafer-Scale Engine (WSE-3)*

On raw power efficiency, Cerebras's own analysis claims a 2.2 times improvement in performance per watt over a comparable

# Introduction and Background

Artificial intelligence (AI) accelerators, purpose-built processors designed to speed up the matrix multiplication and te

The stakes are enormous.
```NVIDIA alone generated **\$215.9 billion** in total revenue for fiscal 2026, up 65% from the prior year, with Data Cen

This comparison focuses on four contenders that between them account for the overwhelming majority of merchant (non-capti

Each vendor is assessed on architecture and capabilities, real-world adoption evidenced by named customer deployments, an

# NVIDIA: The Data Center Standard

## Capabilities

NVIDIA's current data center lineup spans the Hopper-generation **H100** and **H200** GPUs and the newer Blackwell-genera

## Adoption

NVIDIA's ecosystem advantage rests on the maturity of its CUDA software stack, which remains the default target for most

## Strengths and Limitations

NVIDIA's principal strength is ecosystem depth: CUDA, cuDNN, TensorRT-LLM, and the broader NVIDIA AI Enterprise software

# 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 HBM

## Adoption

AMD's Instinct fleet has moved from niche deployments to hyperscaler-scale commitments in the past two years. At its 2024

## 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 t
```
```

```
# 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 th

## Adoption

Intel's most concrete enterprise deployment is with **IBM Cloud**, the first cloud service provider to offer Gaudi 3 acce

## Strengths and Limitations

Intel's clearest differentiator is acquisition price: industry analyst estimates cited by trade press put bulk per-unit G

# Cerebras: Wafer-Scale Computing

## Capabilities

Cerebras takes a fundamentally different design philosophy. Rather than networking many discrete chips, the **Wafer-Scale

## Adoption

Cerebras's commercial momentum accelerated sharply in the first half of 2026. In January, OpenAI signed an agreement to d

## Strengths and Limitations

Cerebras's headline strength is raw single-model inference speed. In its own benchmarking against a 405-billion-parameter

# Feature Comparison

_Table 1_ below summarizes the core specifications of each vendor's current flagship data center accelerator, drawing on

| **Specification** | **NVIDIA GB200 NVL72 (per GPU: B200)** | **AMD Instinct MI355X** | **Intel Gaudi 3** | **Cerebras W
| --- | --- | --- | --- | --- |
| **Architecture** | Blackwell, TSMC 4nm process | CDNA 4, TSMC process | Gaudi 3, matrix and tensor cores | Wafer-scale,
| **Peak AI compute** | Rack: 1,440 PFLOPS NVFP4 (sparse) | Up to 10.1 PFLOPS FP8 (sparsity) per GPU | Up to 1.8 PFLOPS F
| **Memory capacity** | 13.4TB HBM3E across rack | 288GB HBM3E per GPU | 128GB HBM2e per accelerator | 44GB on-chip SRAM
| **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)
| **Approximate system power** | 14.3kW (DGX B200, 8 GPUs) (Source: [arxiv.org](https://arxiv.org/html/2503.11698v1#:~:te
| **Approximate list price** | \~\$500,000 (DGX B200, 8 GPUs, estimate) | Not publicly listed (OEM/hyperscale contracts)
| **Software stack** | CUDA, TensorRT-LLM, NVIDIA AI Enterprise | ROCm (open source) | SynapseAI, PyTorch integration | C

_Table 1_ makes the divergence in design philosophy explicit; the underlying specification figures are drawn from the NVI

# Performance and Benchmarks

MLCommons, the industry consortium behind the MLPerf benchmark suite, published **MLPerf Training v6.0** results in June

On the inference side, an April 2026 analysis of MLPerf Inference v6.0 results derived per-GPU throughput by dividing sys

| **Accelerator** | **GPT-OSS 120B (tok/s per GPU)** | **LLaMA 2 70B (tok/s per GPU)** | **SDXL (samples/s 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_ (Source: [

On the GPT-OSS-120B benchmark, NVIDIA's B200 delivered roughly 2.5 times the per-GPU throughput of the H200, a gap the an

Independent, vendor-adjacent benchmarking from Artificial Analysis offers a complementary view focused on realistic produ

Cerebras does not participate in the standard MLPerf Inference harness, instead publishing benchmarks independently verif

# Data Analysis and Evidence

The quantitative case for NVIDIA's continued market leadership rests on both revenue and independent market-share estimat

_Table 3_ below assembles the most recent quarterly financial data for the three publicly traded vendors, alongside Cereb

| **Company** | **Most Recent Quarterly Data Center / AI Revenue** | **Year-over-Year Growth** | **Source Period** |
| --- | --- | --- | --- |
| NVIDIA | \$62.3 billion (Data Center, Q4 FY2026) (Source: [nvidianews.nvidia.com](http://nvidianews.nvidia.com/news/nvi
| AMD | \$5.8 billion (Data Center segment, Q1 2026), \$10.25 billion total company revenue (Source: [datacenterdynamics.
| Intel | \$5.1 billion (Data Center and AI segment, Q1 2026); \$13.6 billion total revenue (Source: [intc.com](https://w
| Cerebras | Not yet reported as a public company; \$48.8 billion IPO valuation (Source: [cnbc.com](https://www.cnbc.com/

_Table 3_ shows NVIDIA's single quarter of Data Center revenue exceeds the combined trailing quarterly Data Center or AI

On the training-versus-inference split that determines which architecture matters most for a given buyer, Mordor Intellig

# 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

## 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

## 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

## 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 202

# Implications and Future Directions

The near-term trajectory of the AI accelerator market points toward continued architectural specialization rather than co
```

```
Power availability is emerging as an equally consequential constraint as raw silicon supply. Both the AMD-Meta and OpenAI

Geopolitical and supply-chain risk will also continue to shape vendor strategy. NVIDIA's \$4.5 billion H20 export-control

# Frequently Asked Questions (FAQs)

**Which AI accelerator has the best power efficiency in 2026?** On a raw performance-per-watt basis for training, Cerebra

**How does NVIDIA H100 compare to AMD MI300X and Cerebras WSE-3 in practice?** The H100 offers 80GB of memory and up to 3

**Is Intel Gaudi competitive with NVIDIA and AMD for AI training and inference?** Gaudi 3 is positioned primarily for inf

**What is the Cerebras Wafer-Scale Engine and how does it differ from a GPU?** The WSE-3 is a single 46,225 square millim

**How big is the overall AI accelerator market, and who is growing fastest?** Mordor Intelligence estimates the market at

**Which accelerator is best for training versus inference?** Training workloads at frontier scale remain best served by N

# Conclusion

As of July 2026, the "nvidia vs amd vs intel vs cerebras" comparison resolves less into a single winner than into four di

For technical buyers, the practical takeaway is that accelerator selection should be workload-driven rather than brand-dr
```

Tags: 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