# COMPUTE-IQ AI ACCELERATORS 2026

Performance · Ecosystems · Market Adoption · Roadmaps · TCO

Deep Research Playbook

*Manmohan Brahma*

Classification: Strategic Intelligence Report

Date: April 2026

Scope: Data Centre AI Accelerators

Coverage: NVIDIA · AMD · Google · AWS · Intel · Cerebras · Graphcore

## Executive Summary

The AI accelerator landscape in 2026 is defined by a tension between NVIDIA's entrenched dominance and a rapidly maturing field of heterogeneous alternatives. This report provides a comprehensive analysis of every major platform — covering raw performance specifications, software ecosystem depth, market adoption signals, vendor roadmaps through 2028, and total cost of ownership frameworks to guide procurement decisions.

> Key finding: NVIDIA commands approximately 80–90% of the AI accelerator market by revenue as of 2025, generating over $100 billion annually from data center GPUs. However, this share is projected to decline to approximately 75% by 2026 as AMD and custom hyperscaler silicon scale their deployments. [Ref. 1]

Five structural shifts are reshaping the landscape simultaneously:

- Heterogeneity is now a feature, not a bug. Multi-vendor AI infrastructure strategies are being validated by the world's largest AI organizations — OpenAI deploying 1 GW of AMD hardware, Meta in multi-billion-dollar TPU negotiations with Google, Anthropic co-developing infrastructure with AWS Trainium.
- Custom silicon has crossed the threshold of credibility. Google's Trillium (TPU v6) delivers 4.7x peak compute vs. v5e at 67% better energy efficiency; AWS Trainium3 achieves 2.52 PFLOPS FP8 per chip. These are no longer research projects — they are production infrastructure.
- Memory capacity has replaced raw FLOPS as the first-order design constraint. The AMD MI355X's 288 GB HBM3E directly challenges the NVIDIA B200's 192 GB, enabling inference on 400B+ parameter models without tensor parallelism sharding.
- Low-precision formats (FP4, FP6, FP8) are now primary throughput multipliers. NVIDIA's NVFP4 delivers 20 PFLOPS on a single B200, 2.2x more than the same chip's FP8 ceiling. AMD's MXFP4/6 microscaling formats give the MI355X a 2.2x FP6 advantage over the B200.
- Annual hardware cadences create a two-speed market. NVIDIA's Vera Rubin platform (50 PFLOPS FP4, HBM4) is in production for H2 2026, while AMD's Helios/MI455X rack-scale system targets engineering samples in H2 2026 but mass production only by Q2 2027.

**The strategic playbook for technology leaders requires a dual- or tri-vendor approach:**

- Lock NVIDIA GB200/NVL72 for frontier training where the CUDA stack's 18-year maturity is decisive and no alternative matches NVLink bandwidth at scale.
- Route cost-optimized training to cloud-native ASICs — Trillium on GCP at $2.70/chip-hour or Trainium3 on AWS — for workloads that can accept portability constraints.
- Direct high-throughput inference to AMD MI355X where 288 GB HBM3E reduces sharding complexity and delivers 33% lower TCO vs. HGX B200 for self-hosted deployments.

- Reserve Cerebras CS-3 for multi-million-token context workloads or MoE architectures where GPU interconnect bottlenecks degrade performance — its 21x advantage over DGX B200 for giant-context inference is unmatched by any GPU-based system.

# 1. Market Landscape 2026 — Dominance with Accelerating Diversification

## 1.1 Revenue and Market Share Analysis

The AI accelerator market reached $140.55 billion in 2024 and is forecast to reach $440.30 billion by 2030, expanding at a 25% CAGR. [Ref. 19] GPU-based accelerators command approximately 60% of 2024 revenue due to their mature software ecosystems and versatility across training and inference workloads.

| Vendor | Est. Revenue Share | 2025E Revenue | Key Signal |
|---|---|---|---|
| NVIDIA | **~80–87%** | $173B (FY2026E) | Blackwell ramp; $194B DC revenue FY2026 |
| AMD | ~5–8% | $5.6B+ (AI div.) | OpenAI 1 GW partnership; MI355X GA |
| Google (TPU) | ~3–5% internal | ~$3.1B value equiv. | Meta TPU negotiations; Anthropic Trillium |
| AWS (Trainium) | ~2–3% internal | Undisclosed | Trn3 UltraServers; $8B Anthropic investment |
| Intel (Gaudi) | ~1–2% | ~$500M (declining) | Gaudi line retired; Jaguar Shores 2026 |
| Other / Startup | ~2–3% | ~$2B | Cerebras, Groq, SambaNova niche deployments |

NVIDIA's absolute revenue trajectory reflects the scale of compute investment. Data center revenue grew from $15B (FY2022) to $47B (FY2023) to $110B (FY2024) to a projected $173B (FY2025E), now representing over 86% of the company's total revenue. [Ref. 1] The Q4 FY2026 quarter alone exceeded $40 billion — more than the entire company earned in all of FY2022. [Ref. 2]

> Intel's collapse is instructive: the company fell from 68% of combined NVIDIA/AMD/Intel data center share in 2021 to approximately 6% by late 2025 — a direct consequence of CPU-centricity in a market that shifted decisively to GPU-heavy AI systems following ChatGPT's launch. [Ref. 20]

## 1.2 Hyperscaler Dual-Track Strategies

Every major cloud provider now operates a deliberate hybrid strategy. They purchase NVIDIA to satisfy broad customer demand for CUDA-based workloads — because refusing NVIDIA access would be commercially unacceptable — while simultaneously deploying custom silicon to optimize their own infrastructure cost structure.

- Google deploys TPU v5/v6 Trillium to power internal AI workloads (Search, Photos, Maps, YouTube, all Gemini models) and offers these to GCP customers, with reported 65% inference cost reductions for workloads like Midjourney that migrated from GPU-based infrastructure. [Ref. 3]

- Amazon's $8 billion investment in Anthropic is partly predicated on Trainium adoption, pledging access to up to one million custom chips. AWS's Trn3 UltraServer racks scale to 144 Trainium3 chips with a proprietary NeuronSwitch fabric — directly competing with NVIDIA's NVLink domain at a fraction of the licensing cost.
- Microsoft runs its Copilot suite on Maia 100, a custom ASIC, while simultaneously being one of NVIDIA's largest Blackwell customers. Azure GB300-based VM series launched in 2026 as part of the next-generation Azure AI infrastructure build-out.
- Meta is in multi-billion-dollar talks to utilize Google's TPUs — a remarkable signal of cross-hyperscaler silicon adoption that represents perhaps the strongest validation of heterogeneous fleet strategies in 2026. [Ref. 6]

## 1.3 Notable Deployments Validating Heterogeneous Fleets

| Organization | Hardware Choice | Scale / Detail | Strategic Signal |
|---|---|---|---|
| OpenAI | AMD MI355X / Instinct | 1 GW deployment, 2H 2026 | Largest non-NVIDIA AI commitment in history |
| Anthropic | AWS Trainium3 | Up to 1M chips pledged | $8B AWS investment anchors Trainium demand |
| Meta | Google TPU (in talks) | Multi-billion dollar | Rival hyperscaler's silicon for own training |
| CoreWeave | NVIDIA GB200 NVL72 | 2,496-GPU cluster | MLPerf Training v4.1 record-holder |
| Midjourney | Google Trillium v6 | 65% cost reduction | Migration from H100 proves TPU inference value |
| DeepMind | Google TPU fleet | Internal migration | Nobel Prize-winning AlphaFold ran on TPUs |
| Oracle Cloud | NVIDIA Vera Rubin | OCI Gigascale AI Factory | First VR200 cloud deployments H2 2026 |
| Microsoft Azure | NVIDIA Vera Rubin | Fairwater superfactory | Next-gen Azure AI infrastructure foundation |

## 2. NVIDIA Blackwell Platform — Architecture, Performance, and Ecosystem

### 2.1 Architecture Overview

NVIDIA's Blackwell architecture — announced at GTC March 2024 and entering volume production through 2025–2026 — represents the largest single-generation performance leap in NVIDIA's datacenter GPU history. Named after mathematician David Harold Blackwell, the architecture features a dual-die design that fundamentally changes the NVLink topology compared to prior single-die GPUs.

The B200 GPU packs 208 billion transistors across two dies in a single package (2.6x the transistor count of the H100), manufactured on TSMC's 4NP process node. The critical packaging technology is CoWoS (Chip-on-Wafer-on-Substrate), which stacks HBM3e memory modules directly on the GPU die. CoWoS capacity remains extraordinarily limited — TSMC is one of the few companies capable of producing it at scale, and demand far exceeds supply, creating allocation queues and driving 3–6 month lead times for large orders as of Q1 2026. [Ref. 2]

| Specification | B200 (Blackwell) | GB200 NVL72 (Rack) | B300 / Blackwell Ultra |
|---|---|---|---|
| Transistors | 208 billion (dual-die) | 72 B200 GPUs + 36 Grace CPUs | ~250B (est.) |
| Process Node | TSMC 4NP | TSMC 4NP + CoWoS | TSMC 4NP advanced |
| HBM Memory | 192 GB HBM3e | 13.8 TB aggregate (rack) | 288 GB HBM3e |
| Memory Bandwidth | 8 TB/s per GPU | 130 TB/s bisection (rack) | ~12 TB/s |
| Peak FP4 (sparse) | 20 PFLOPS | 1,440 PFLOPS (rack) | ~30 PFLOPS |
| Peak FP8 (dense) | 9,000 TFLOPS | ~720 PFLOPS (rack) | ~13,500 TFLOPS |
| Peak FP16/BF16 | 4,500 TFLOPS | ~360 PFLOPS (rack) | ~6,750 TFLOPS |
| NVLink Bandwidth | 1.8 TB/s bi-directional | 130 TB/s aggregate | ~2.4 TB/s |
| TDP | 1,000W (liquid req.) | ~120–140 kW/rack | ~1,200W |
| PCIe | PCIe Gen 6 | N/A (NVLink native) | PCIe Gen 6 |
| Server Price (unit) | $45,000–$55,000/GPU | $5–8.8M/rack | ~$60,000/GPU (est.) |
| Cloud Price | $6–$27/hr (on-demand) | Varies by provider | Premium (TBD) |

### 2.2 Second-Generation Transformer Engine and FP4

The most consequential Blackwell innovation for production AI inference is the second-generation Transformer Engine, which introduces native hardware-accelerated FP4 arithmetic — a capability the H100 and H200 do not possess.

- FP4 (native): The B200 delivers 9,000 TFLOPS FP4 dense, and with the 2:1 sparsity mask, 18,000 TFLOPS. H100 has no native FP4 hardware support at all. For inference, FP4 effectively doubles throughput vs. FP8 on the same GPU. [Ref. 4]
- FP8 at scale: B200's 4,500 TFLOPS FP8 dense is roughly 2.3x the H100's 1,979 TFLOPS. This improvement applies to any FP8 workload without requiring FP4 quantization.
- NVIDIA's implementation uses block-scaled FP4 with per-channel scaling factors stored in higher precision — preserving more of the original model's accuracy than naive 4-bit quantization. The Transformer Engine manages these scaling factors automatically.
- Real-world result: Blackwell-powered systems deliver inference at $0.02 per million tokens on GPT-OSS-120B as of April 2026 — a 5x improvement over Blackwell's own launch-day costs, and 4.5x cheaper than H100-powered systems. [Ref. 1]

> InferenceMAX v1 benchmark (SemiAnalysis, April 2026): NVIDIA B200 delivers 60,000 tokens per second per GPU and 10,000 TPS per GPU at 50 TPS/user interactivity on Llama 3.3 70B — 4x higher per-GPU throughput vs. H200. A $5 million investment in GB200 NVL72 can generate $75 million in DSR1 token revenue — a 15x ROI. [Ref. 10]

## 2.3 NVLink 5 and GB200 NVL72 Scale Architecture

The GB200 NVL72 represents NVIDIA's rack-scale compute architecture — 72 B200 GPUs interconnected with 36 Grace Arm CPUs through NVLink 5, functioning as a single massive GPU with 13.8 TB of unified GPU memory.

- NVLink 5.0: 1.8 TB/s bidirectional per GPU — 2x faster than H100/H200's 900 GB/s — connecting all 72 GPUs in a unified NVLink domain with 130 TB/s of aggregate bisection bandwidth.
- Grace CPU integration: 72 ARM Neoverse V2 cores per Grace CPU with 1.5 TB LPDDR5x system memory. The memory-coherent NVLink-C2C interconnect at 900 GB/s eliminates PCIe bottlenecks for CPU-GPU communication.
- Training benchmark: A 2,496-GPU Blackwell cluster at CoreWeave trained Llama 3.1 405B in 27.33 minutes in MLPerf Training v4.1. [Ref. 11]
- Inference at rack scale: The GB200 NVL72 generates 30x more inference tokens per second than equivalent Hopper rack configurations. MLPerf Inference v4.1 recorded 10,756 tokens/second for Llama2-70B on a single NVL72 rack. [Ref. 10]
- Power and cooling: The NVL72 rack draws approximately 120–140 kW, requiring direct liquid cooling (DLC). NVIDIA's Kyber rack infrastructure (2027) will draw 600 kW per rack.

## 2.4 NVIDIA CUDA Software Ecosystem

CUDA's competitive moat is architectural: 18 years of accumulated library optimizations, debugger integrations, compiler toolchains, and framework integrations that no competitor can replicate on a short timeline.

| Component | Description | Competitive Edge |
|---|---|---|
| CUDA Toolkit | Compiler, libraries, debugger (since 2007) | 18-year head start; 4M+ developers |
| cuDNN | Deep learning primitives library | Heavily optimized attention, convolution, LSTM kernels |
| cuBLAS | GPU-accelerated BLAS | Vendor-tuned for every Tensor Core generation |
| NCCL | Multi-GPU collective operations | NVLink-aware all-reduce; critical for training scale |
| TensorRT-LLM | LLM inference optimization | FP4/FP8/speculative decoding; $0.02/M token record |
| Triton IS | Model serving infrastructure | Multi-model concurrent serving; dynamic batching |
| NeMo | Training framework | RLHF, alignment, fine-tuning pipelines |
| CUDA Graphs | Kernel launch optimization | Dramatically reduces CPU overhead at inference time |
| FlashAttention-3 | Memory-efficient attention | Optimized for Hopper/Blackwell Tensor Cores |
| Nsight | GPU profiling suite | Cycle-accurate kernel analysis; no ROCm equivalent depth |

The HIPIFY translation layer (which AMD's ROCm uses to port CUDA code) achieves approximately 80% automatic translation success. The remaining 20% — primarily custom fused kernels, CUDA Graphs implementations, and low-level NCCL usage — require manual expert porting effort that can take months for a production codebase. This is the primary source of migration risk for any organization considering shifting training workloads from NVIDIA to AMD.

# 3. AMD Instinct MI350 Series and ROCm Ecosystem

## 3.1 CDNA 4 Architecture and MI350 Series

AMD's Instinct MI350 series, built on the new CDNA 4 architecture using TSMC's 3nm process, represents the company's most significant generational leap in inference performance: a 35x increase in AI inference throughput compared to the MI300 series (MI300X at FP8 vs. MI355X at FP4) for Llama 3.1-405B serving. [Ref. 2]

Two product variants address different deployment profiles:

- MI350X: The 1,000W TDP air-cooled variant — the same form factor as the MI300X series, designed for broad data center compatibility without dedicated liquid cooling infrastructure. Supports both air-cooled and DLC environments.
- MI355X: The 1,400W TDP liquid-cooled variant, targeting maximum performance for hyperscale deployments. Despite the 40% power increase, on-paper specs show less than 10% performance differential vs. MI350X in TFLOPS — the premium is in sustained boost frequency under sustained workload.

| Specification | AMD MI350X | AMD MI355X | NVIDIA B200 (comparison) |
|---|---|---|---|
| Architecture | CDNA 4 | CDNA 4 (higher TDP) | Blackwell |
| Process | TSMC 3nm | TSMC 3nm | TSMC 4NP |
| HBM3e Memory | 288 GB | 288 GB | 192 GB |
| Memory Bandwidth | ~7.7 TB/s | ~8.0 TB/s | 8.0 TB/s |
| Peak MXFP4 | ~10 PFLOPS | 10.1 PFLOPS | 20 PFLOPS (NVFP4 sparse) |
| Peak FP8 | ~9.5 PFLOPS | ~10 PFLOPS (est.) | 9,000 TFLOPS dense |
| Peak FP6 (MXFP6) | ~10 PFLOPS | ~10 PFLOPS | ~4,500 TFLOPS |
| Peak FP64 | 72 TFLOPS | 79 TFLOPS | ~40 TFLOPS |
| Scale-out | 400 Gbit/s | 400 Gbit/s | 400 Gbit/s |
| TDP | 1,000W | 1,400W | 1,000W |
| Cooling | Air or DLC | DLC only | DLC required |
| Server Price est. | ~$30–40K/GPU | ~$35–45K/GPU | $45–55K/GPU |
| Cloud Price (OCI) | ~$8.60/hr | ~$8.60/hr | $6–27/hr |
| TCO vs. B200 (self-hosted) | ~33% lower | ~33% lower | Baseline |

## 3.2 MLPerf Inference 6.0 — AMD's Validation Moment

MLPerf Inference 6.0 (results released April 2026) marks AMD's most significant public benchmark achievement. Four GPU types were submitted: MI300X, MI325X, MI350X, and MI355X — demonstrating a multi-generational ecosystem, not just a single flagship result.

- Milestone result: AMD Instinct MI355X GPUs delivered more than 1 million tokens per second at cluster scale (87-GPU configuration) for Llama2-70B inference — the first time any non-NVIDIA platform has crossed this threshold in a public benchmark. [Ref. 2]
- Reproducibility: Partner results on MI355X hardware landed within 4% of AMD's own submission in most cases, and within 1% for some workloads. This is a strong signal that benchmark results translate to real deployment performance, not lab artifacts. [Ref. 2]
- First-time workload bring-up: AMD successfully submitted results for GPT-OSS-120B and Wan-2.2-t2v video generation — neither had been benchmarked on AMD hardware in prior MLPerf rounds. This demonstrates accelerating software ecosystem maturation.
- MX formats advantage: The MI355X's MXFP6 capability delivers 2.2x more throughput than the B200 for FP6 workloads, since AMD's FP6 shares physical circuits with FP4 (rather than FP8 as on NVIDIA hardware).

> FP6 advantage context: On the HGX B200, FP6 shares the same physical circuits as FP8, yielding the same TFLOPS ceiling for both. On the MI355X, FP6 shares circuits with FP4, meaning MI355X FP6 = MI355X FP4 TFLOPS — 2.2x faster than B200 FP6. For inference workloads where FP6 accuracy is acceptable, this is decisive. [Ref. 14]

## 3.3 ROCm Software Ecosystem — State of Play April 2026

AMD's open-source ROCm platform has undergone its most aggressive development cycle in history between ROCm 6.0 and 7.2 (September 2025 — April 2026). The company's hiring of Anush Elangovan as AI Software leader and the corresponding organizational investment has produced measurable results.

| ROCm Release | Key Milestones | Framework Support |
|---|---|---|
| ROCm 7.0 (Sep 2025) | MI355X/MI350X official support; FP4/FP6 low-precision types added; KVM passthrough for MI350X/MI355X | PyTorch 2.7 (first-class); JAX 0.6.0; TF 2.19.1; ONNX 1.22; Triton 3.3 [Ref. 12] |
| ROCm 7.0.2 (Nov 2025) | PyTorch 2.8; FlashInfer library for LLMs; torch.nn.functional.scaled_dot_product_attention flash attention kernel auto-routing | MosaicML; DeepSpeed FP8 quantized training via APEX extension |
| ROCm 7.1.1 (Feb 2026) | PyTorch 2.9 (first-class); SGLang disaggregated prefill/decode inference guide published; vLLM inference performance tuning improvements | DGL (Deep Graph Library) GPU-accelerated; rocSHMEM RO inter-node backend |
| ROCm 7.2 (Apr 2026) | JAX 0.8.2 support; Triton 3.6.x GEMM kernel optimizations; AMD SMI replaces ROCm SMI fully; TheRock unified CMake build system preview | MONAI for AMD ROCm (medical AI); HipBLASLt default on new GPUs |

Despite ROCm's rapid progress, meaningful gaps remain versus CUDA for production workloads:

- Custom fused kernels: HIPIFY achieves ~80% automatic translation. The remaining 20% — particularly custom attention kernels, custom loss functions with CUDA Graphs, and low-level NCCL implementations — require expert manual porting.
- Windows support: ROCm Windows builds remain in preview as of April 2026, limiting adoption in enterprise environments that standardize on Windows Server for AI workloads.
- Profiling depth: ROCm SMI, rocprof, and rocprofv2 are being deprecated in favor of AMD SMI and ROCprofiler-SDK. This transition creates documentation gaps and tooling inconsistency during the migration window.
- Library depth: MIOpen (AMD's cuDNN equivalent) and rocBLAS lack the per-kernel tuning depth of cuDNN across the full model zoo. Performance on niche architectures (e.g., Mamba, RWKV) may require manual kernel work.

# 4. Hyperscaler Custom Silicon — Google TPU and AWS Trainium

## 4.1 Google TPU — Trillium (v6) and Ironwood (v7)

Google's Tensor Processing Units have evolved through seven generations since 2015. The v6 Trillium (GA on GCP, late 2024) and v7 Ironwood (2025–2026) represent the company's most competitive designs relative to NVIDIA flagship hardware — a gap that prior generations failed to close.

| Specification | TPU v5e | TPU v5p | Trillium v6e | TPU v7 Ironwood |
|---|---|---|---|---|
| BF16 Compute | 197 TFLOPS | 459 TFLOPS | 918 TFLOPS | Near H100-level |
| HBM Memory | 16 GiB | 95 GiB | 32 GiB | Not disclosed |
| ICI Bandwidth | 400 GB/s | 3D torus | 800 GB/s | Enhanced |
| Pod Scale | Up to 256 | Up to 8,960 | Up to 256 | TBD |
| vs. Predecessor | 2.5x vs v4 | 3x training vs v4 | 4.7x vs v5e | Near H100 parity |
| Energy Efficiency | Baseline | Higher TDP | 67% better vs v5e | Further improved |
| On-demand Price | ~$1.20/chip-hr | ~$4.20/chip-hr | ~$2.70/chip-hr | TBD |
| Perf/$ vs prior | 2.3x vs v4 | Baseline | 1.8x vs v5p | N/A |
| Software stack | OpenXLA/JAX | OpenXLA/JAX | OpenXLA/JAX/TorchTPU | OpenXLA |
| Portability | GCP only | GCP only | GCP only | GCP only |

Key TPU deployment and adoption signals:

- Anthropic is Google's most significant external TPU customer. The $8 billion AWS investment and up to 1 million Trainium chip pledge is matched by a parallel Broadcom/Google deal: Broadcom will supply Anthropic with 3.5 gigawatts of TPU capacity from 2027 onward — validating that frontier AI labs are deliberately pursuing multi-cloud, multi-silicon strategies.
- Midjourney's migration to Trillium v6 reduced their monthly inference spending by 65%, establishing the most cited cost reduction case study for cloud ASICs in production. [Ref. 30]
- DeepMind runs TPUs as primary compute for internal research, including the Nobel Prize-winning AlphaFold protein structure prediction system.
- TPU v7 Ironwood closing the hardware gap: Unlike prior generations that lagged NVIDIA flagships by 2+ years in compute density, TPUv7 is available within a few quarters of H100/H200 general availability — and delivers near-parity FLOPs. [Ref. 35]

> TPU portability constraint: All TPU models require either JAX (via OpenXLA/XLA compiler) or PyTorch-XLA (torch.compile() with XLA backend). CUDA code requires complete rewriting. Unlike ROCm's HIPIFY translation layer, there is no automated CUDA-to-XLA porting tool. Migration cost should be treated as 6–18 person-months for a mature CUDA training codebase. [Ref. 13]

## 4.2 AWS Trainium3 and the Neuron SDK

Amazon's Trainium3, launched in late 2025 via the EC2 Trn3 UltraServer instance family, represents AWS's most ambitious custom silicon deployment to date. The architecture scales individual chips into 144-chip UltraServer configurations via the proprietary NeuronSwitch fabric.

| Specification | Trainium2 | Trainium3 (Trn3) |
|---|---|---|
| Peak Compute (FP8) | ~1.3 PFLOPS/chip | 2.52 PFLOPS/chip |
| HBM Memory | 96 GB per chip | 144 GB per chip |
| UltraServer Config | 16 chips (Trn2.48xlarge) | 144 chips (Trn3 UltraServer) |
| UltraServer FP8 Total | 20.8 PFLOPS | ~363 PFLOPS |
| Scale-out Fabric | NeuronSwitch gen2 | NeuronSwitch gen3 (enhanced) |
| vs. H100 Price-Perf | 30–40% better than H100 | Comparable to B200 (AWS-native) |
| Availability | GA (ec2 trn2) | GA (ec2 trn3) [Ref. 4] |
| SDK | AWS Neuron SDK | AWS Neuron SDK |
| PyTorch support | Via torch-neuronx | Via torch-neuronx |
| JAX support | Limited preview | Preview expanding |
| Portability | AWS only | AWS only |
| Best workloads | LLM inference, BERT training | Frontier LLM training, RL |

The Neuron SDK is the primary software abstraction layer for Trainium. It supports PyTorch (via torch-neuronx), with JAX support in preview as of April 2026. TensorFlow support is limited. Critically, there is no CUDA compatibility layer — all code must be written or ported using Neuron SDK APIs or the supported framework backends. This is the primary adoption barrier outside organizations that are building greenfield AWS-native AI infrastructure.

# 5. Intel Gaudi 3 and the Jaguar Shores Transition

## 5.1 Gaudi 3 Architecture

Intel's Gaudi 3, the final product in the Gaudi accelerator line, launched in GA form during 2025. Its primary differentiation is an Ethernet-first scale-out architecture leveraging standard RoCE v2 networking (24 x 200 GbE ports per accelerator), which eliminates the proprietary fabric premium associated with NVIDIA's NVLink or AMD's UALink.

| Specification | Intel Gaudi 3 | Notes |
|---|---|---|
| Compute (FP8/BF16) | ~1.8 PFLOPS | Competitive with H100 for inference |
| HBM2e Memory | 128 GB | Lower than MI355X (288 GB) and B200 (192 GB) |
| Memory Bandwidth | ~3.7 TB/s | Below Blackwell/MI350 class |
| Scale-out Networking | 24 x 200 GbE | RoCE v2 — standard Ethernet stack |
| 8-Card Server Price | ~$157,613 | Competitive on price vs. NVIDIA/AMD configs |
| Software Stack | SynapseAI | PyTorch via habana-torch; TF support |
| CUDA Compatibility | None | No HIPIFY equivalent |
| Market Traction | Limited | Modest shipments vs. NVIDIA dominance |
| Status | Retiring | Final Gaudi generation [Ref. 44] |
| Gross Margin | ~58.4% | Below NVIDIA (84–88%) and AMD (64–68%) [Ref. 36] |

> Intel's Gaudi 3 offered a genuine price-performance advantage for Ethernet-native deployments, but the SynapseAI software ecosystem lacked the framework depth and developer community that ROCm — let alone CUDA — provides. This ecosystem gap, not hardware specs, was the decisive adoption barrier. [Ref. 44, 45]

## 5.2 Jaguar Shores — Intel's Rack-Scale Pivot

Rather than continuing the standalone accelerator strategy with a 'Gaudi 4', Intel has pivoted to a rack-scale integrated architecture called Jaguar Shores, planned for 2026. This mirrors the industry trend toward full-stack rack-level solutions (NVIDIA NVL72, AMD Helios) rather than discrete card-level products.

- Jaguar Shores integrates compute, memory, and networking at the system level — addressing the interface-level bottlenecks that limited Gaudi's practical throughput vs. theoretical peak compute.
- Crescent Island represents a parallel GPU product line alongside Jaguar Shores, targeting specific market segments.
- Intel's manufacturing advantage (in-house fabs) positions it as a potential second-source option for buyers concerned about TSMC concentration risk — a supply chain consideration that becomes increasingly important as HBM4 demand creates allocation pressure.

# 6. Specialist Architectures — Cerebras CS-3 and Graphcore

## 6.1 Cerebras CS-3 — Wafer-Scale Engine

Cerebras Systems' CS-3 is a fundamentally different architecture from all GPU-based accelerators: a wafer-scale engine (WSE) that places an entire silicon wafer's worth of compute on a single substrate, eliminating die-to-die communication latency entirely. This unlocks capabilities that are physically impossible to replicate in multi-chip GPU architectures.

| Specification | Cerebras CS-3 | NVIDIA DGX B200 (comparison) |
|---|---|---|
| Peak Compute (FP16) | 125 PFLOPS | ~144 PFLOPS (8×B200) |
| On-chip SRAM | 44 GB total | None (HBM3e external) |
| HBM (external) | None | 1.5 TB HBM3e (8×192 GB) |
| Memory Architecture | Weight Streaming Engine | Conventional HBM with NVLink |
| On-chip Bandwidth | >20 PB/s internal SRAM | ~64 TB/s aggregate NVLink (8 GPU) |
| Giant Context perf. | 21x faster vs DGX B200 | Baseline |
| MoE model fit | Excellent (sparse activation) | Good (requires tensor parallelism) |
| Framework support | PyTorch via Cerebras SDK | Full CUDA ecosystem |
| Deployment model | On-premise appliance | Standard server infrastructure |
| Best workloads | Giant context, MoE, sparsity | General-purpose training/inference |

The Cerebras Weight Streaming Engine architecture stores model weights in off-chip external memory and streams them through the on-chip SRAM, eliminating the need to shard models across multiple GPUs for giant-context workloads. This architectural innovation delivers:

- 21x faster inference vs. DGX B200 for multi-million token context workloads — where GPU interconnect bandwidth becomes the primary bottleneck rather than compute throughput.
- Near-linear scaling for sparse MoE (Mixture-of-Experts) architectures, where only a subset of parameters are activated per token. GPUs waste HBM bandwidth loading inactive expert weights; the WSE eliminates this overhead.
- Zero inter-chip communication latency: all 44 GB of on-chip SRAM is addressable by any processing element, eliminating the NVLink/ICI hops that introduce microsecond-scale collective latency in GPU clusters.

> Strategic guidance: Cerebras CS-3 is NOT a general-purpose training platform and should not be evaluated as a replacement for NVIDIA or AMD GPUs. It is a purpose-built specialist system for two specific workload categories: (1) inference or fine-tuning on models with million-token+ context windows, and (2) sparse MoE architectures where GPU interconnect bandwidth creates unacceptable training inefficiency.

# 7. Performance Benchmarks — Cross-Platform Analysis

## 7.1 MLPerf Inference v6.0 — April 2026 Definitive Results

MLPerf Inference v6.0 (April 2026) is the most comprehensive public benchmark for AI accelerator inference performance. Results cover both Offline (maximum throughput) and Server (latency-constrained) scenarios across LLM inference, vision, recommender, and multimodal workloads.

| Platform | Llama2-70B Offline (t/s) | GPT-OSS-120B (t/s) | Mode | Notes |
|---|---|---|---|---|
| NVIDIA B200 (FP4) | 17,500 | N/A (single server) | Single server | MLPerf v6.0 record [Ref. 3, 10] |
| NVIDIA GB200 NVL72 (FP4) | ~180,000+ | N/A | Single rack | Rack-scale inference |
| AMD MI355X (87-GPU cluster) | 1,042,110+ | First submission | Multi-node | 1M tokens/s milestone [Ref. 2] |
| AMD MI355X (single server) | ~14,000 | ~8,500 (est.) | Single server | Competitive vs B200 per-node |
| Google Trillium v6e | ~8,500 (est.) | N/A | Single pod | Based on 1.8x vs v5p perf/$ [Ref. 3] |
| Intel Gaudi 3 | ~4,200 (est.) | N/A | Single server | FP8 inference competitive |
| NVIDIA H100 (baseline) | ~3,000 | ~1,800 (est.) | Single server | Reference generation |
| NVIDIA H200 | ~9,800 | ~5,500 (est.) | Single server | HBM3e upgrade vs H100 |

## 7.2 Training Performance

| Workload | Platform | Result | Source |
|---|---|---|---|
| Llama 3.1 405B pre-training | NVIDIA GB200 (2,496-GPU) | **27.33 minutes** | MLPerf Training v4.1 [Ref. 11] |
| Llama 2 70B fine-tuning | B200 vs H100 (head-to-head) | 2.2x faster on B200 | MLPerf Training 4.1 [Ref. 8] |
| GPT-3 175B pre-training | B200 vs H100 (head-to-head) | 2x faster on B200 | MLPerf Training 4.1 [Ref. 8] |
| GPT-3 175B (Google) | Trillium v6e | 1.8x better perf/$ vs v5p | Google Cloud benchmarks [Ref. 3] |
| Task that required 256 Hoppers | 64 B200 GPUs | Same throughput, 4x fewer GPUs | MLPerf Training 4.1 [Ref. 8] |
| Vision CV pretraining (YOLOv8) | B200 vs H100 (cloud) | 57% faster on B200 | Lightly.ai real-world test [Ref. 6] |

| Workload | Platform | Result | Source |
|---|---|---|---|
| Llama2-70B (AMD MI355X 8-GPU) | AMD MI355X platform | 73.8 PFLOPS FP8 (8-GPU platform) | AMD internal, Oct 2025 [Ref. 11] |

## 7.3 Precision Format Performance Summary

Low-precision formats have become the primary lever for inference throughput optimization. The following table compares peak theoretical throughput across formats per accelerator — the numbers that dictate real-world inference cost at scale.

| Accelerator | FP32 | BF16/FP16 | FP8 | FP6 | FP4 |
|---|---|---|---|---|---|
| NVIDIA B200 | ~1.5 PFLOPS | 4,500 TFLOPS | 9,000 TFLOPS | 9,000 TFLOPS* | 18,000 TFLOPS (sparse) |
| NVIDIA GB200 NVL72 | ~108 PFLOPS | ~324 PFLOPS | ~648 PFLOPS | ~648 PFLOPS* | ~1,440 PFLOPS (sparse) |
| AMD MI355X | 79 TFLOPS (FP64) | ~10 PFLOPS | ~10 PFLOPS | ~10 PFLOPS** | 10.1 PFLOPS (MXFP4) |
| AMD MI400 (roadmap) | N/A | ~20 PFLOPS | 20 PFLOPS | ~40 PFLOPS** | 40 PFLOPS |
| Intel Gaudi 3 | ~0.9 PFLOPS | ~1.8 PFLOPS | ~1.8 PFLOPS | N/A | N/A |
| Google Trillium v6e | ~230 TFLOPS | 918 TFLOPS | N/A (BF16-only) | N/A | N/A |
| AWS Trainium3 | N/A | ~5 PFLOPS | 2.52 PFLOPS | N/A | N/A |

* NVIDIA B200 FP6 shares FP8 circuits, yielding same TFLOPS ceiling as FP8.

** AMD MI355X FP6 shares FP4 (MXFP4) circuits, yielding the same TFLOPS ceiling as FP4 — making MI355X FP6 2.2x faster than B200 FP6 in equal configurations.

## 8. Software Ecosystems — The Decisive Competitive Dimension

Hardware specifications are necessary but not sufficient to win workloads. The software ecosystem — framework integration depth, debuggability, library coverage, and migration tooling — is the decisive adoption factor for the vast majority of enterprise and research deployments. This section provides a comprehensive comparison across all four primary software stacks.

### 8.1 CUDA — The Incumbent Standard

CUDA (Compute Unified Device Architecture) launched in 2007 as the first GPU computing framework. Its 18-year compounding advantage in library optimization, framework integration, debugger tooling, and developer community represents the single most important structural advantage in the AI accelerator market.

- Developer reach: 4+ million registered CUDA developers. The PyTorch, TensorFlow, and JAX communities are all CUDA-first in their development, testing, and optimization pipelines.
- Library depth: cuDNN provides deep per-kernel optimization across the full model zoo including attention mechanisms, convolutions, normalization layers, and sequence models. cuBLAS provides vendor-tuned matrix operations updated with each GPU generation. NCCL provides NVLink-aware collective communications critical for distributed training.
- Performance optimization cycle: NVIDIA's TensorRT-LLM achieved a 5x cost reduction in just two months post-Blackwell launch through continuous kernel optimization — the software moat compounds with each release.
- Speculative decoding support: The GPT-OSS-120B Eagle3-v2 model with speculative decoding in TensorRT-LLM predicts multiple tokens simultaneously, dramatically improving effective inference throughput beyond raw TFLOPS.

### 8.2 ROCm — AMD's Open Ecosystem Push

ROCm is AMD's open-source GPU computing platform, designed to provide CUDA-equivalent functionality while leveraging the HIP (Heterogeneous-compute Interface for Portability) runtime that enables portability between AMD and NVIDIA hardware.

- HIPIFY: The CUDA-to-HIP translation tool handles approximately 80% of CUDA code automatically. The remaining 20% — custom fused kernels, CUDA Graphs, and low-level synchronization primitives — requires manual expert work.
- PyTorch integration: PyTorch now treats ROCm as a first-class installation option alongside CUDA. The PyTorch installer offers ROCm as an equal choice. ROCm 7.x delivers performance 'just shy of CUDA in most training scenarios' per independent benchmarks as of April 2026. [Ref. 50]
- TheRock: AMD's new unified CMake build system for the ROCm platform is under active development, promising a cleaner developer experience than the current multi-repo build infrastructure.

- vLLM support: ROCm-compatible vLLM is available as of ROCm 7.x, enabling production LLM serving on AMD hardware with the same serving infrastructure as CUDA deployments.

## 8.3 OpenXLA / JAX — Google's Portable Compiler Stack

OpenXLA is Google's open-source compiler infrastructure that enables JAX (and PyTorch-XLA) code to run across TPUs, NVIDIA GPUs, and AMD ROCm hardware via a unified intermediate representation (HLO — High-Level Operations).

- Google-first design: OpenXLA's primary optimization target is TPU architecture. GPU support (both NVIDIA via cuBLAS/cuDNN backends and AMD via ROCm) is available but may not match vendor-native optimization depth.
- Portability promise: In theory, a JAX model should run unchanged on TPU, H100, and MI355X. In practice, performance-critical code paths often require hardware-specific kernels (Pallas for TPU, CUDA C++ for GPU) to achieve optimal throughput.
- Research community adoption: JAX is widely used in research for its functional programming model, automatic differentiation, and JIT compilation. DeepMind, Google Brain, and Anthropic (for certain research workloads) use JAX extensively.
- AMD ROCm JAX support: ROCm 7.0 added JAX 0.6.0 support; ROCm 7.2 extends to JAX 0.8.2. This enables AMD hardware for JAX-based training workloads without code modification.

## 8.4 AWS Neuron SDK

The AWS Neuron SDK is a purpose-built deep learning SDK for Amazon's Trainium and Inferentia custom silicon. It provides PyTorch and TensorFlow integration via wrapper APIs (torch-neuronx, tensorflow-neuron) that compile model graphs for Neuron architecture execution.

- Compilation model: Neuron SDK uses ahead-of-time (AOT) compilation — models are compiled into Neuron Executable File Format (.neff) before deployment. This adds a compilation step not required in CUDA/ROCm but enables aggressive hardware-specific optimization.
- Integration with PyTorch: torch-neuronx provides a familiar PyTorch interface. Most standard PyTorch models port with minimal code changes. Custom operators may require Neuron SDK-specific implementations.
- Limitation: Neuron SDK is AWS-proprietary and hardware-specific. Unlike ROCm (which offers some portability to NVIDIA via HIP) or OpenXLA (which targets multiple hardware backends), Neuron code runs exclusively on Trainium/Inferentia hardware.

# 9. Total Cost of Ownership — A Multi-Dimensional Framework

## 9.1 Hardware Acquisition Costs

| System Configuration | Hardware Cost | Notes |
|---|---|---|
| NVIDIA B200 (single GPU) | $45,000–$55,000 | Market rate; enterprise volume discounts available |
| NVIDIA DGX B200 (8× B200) | ~$515,000 | Fully configured with DLC infrastructure |
| NVIDIA GB200 Superchip (1 Grace + 2 B200) | $60,000–$70,000 | Integrated CPU+GPU module |
| NVIDIA GB200 NVL72 (rack-scale) | $5,000,000–$8,800,000 | Full rack system; infrastructure extra [Ref. 22] |
| AMD MI355X (single GPU, est.) | ~$35,000–$45,000 | List price; volume discounts available |
| AMD HGX MI355X 8-GPU Server | ~$250,000–$320,000 (est.) | ~33% lower TCO than HGX B200 equivalent [Ref. 14] |
| Intel Gaudi 3 (8-card server) | ~$157,613 | Supermicro reference system [Ref. 15] |
| Cerebras CS-3 (appliance) | ~$2,000,000+ (est.) | As-a-service also available |

## 9.2 Cloud On-Demand Pricing

| Instance / Platform | On-Demand Price | Spot/Preemptible | Notes |
|---|---|---|---|
| NVIDIA H100 SXM (p4d/p4de) | ~$3.20/GPU-hr | ~$1.10–1.50/GPU-hr | Widely available; mature ecosystem |
| NVIDIA H200 SXM | ~$3.90/GPU-hr | ~$1.40–1.80/GPU-hr | Limited availability 2025 |
| NVIDIA B200 (various providers) | $6.03–$27.04/GPU-hr | $2.12–$6/GPU-hr (spot) | Supply-constrained; CoreWeave, Lambda [Ref. 7] |
| AMD MI355X (OCI) | ~$8.60/GPU-hr | Not published | Oracle Cloud Infrastructure [Ref. 8] |
| Google TPU v6e (Trillium) | ~$2.70/chip-hr | ~$0.80–1.20 (preemptible) | Best inference $/token on GCP [Ref. 9] |
| Google TPU v5p | ~$4.20/chip-hr | ~$1.50 (preemptible) | Large-scale training pods [Ref. 9] |
| Google TPU v5e | ~$1.20/chip-hr | ~$0.39 (preemptible) | Cost-optimized inference [Ref. 9] |
| AWS Trainium2 (trn2.48xlarge) | ~$33/hr (16-chip inst.) | Savings plan available | ~30-40% better than P5e/H100 [Ref. 32] |

## 9.3 Token Economics and Cost-per-Output Analysis

In the inference-dominated AI economy of 2026, cost-per-million-tokens is the primary TCO metric — not hardware acquisition cost, not FLOPS, not memory. The following data points represent verified production benchmarks.

| Model / Platform | Cost per Million Tokens | TPS per GPU | Source |
|---|---|---|---|
| B200 + TensorRT-LLM (GPT-OSS-120B) | **$0.02** | 60,000 (gpt-oss) | SemiAnalysis InferenceMAX v1, Apr 2026 [Ref. 10] |
| H100 + vLLM (GPT-OSS-120B) | $0.09 | ~10,000 (gpt-oss est.) | SemiAnalysis InferenceMAX v1, Apr 2026 [Ref. 10] |
| B200 (Llama2-70B FP4, MLPerf) | ~$0.04 (est.) | 17,500 | MLPerf Inference v6.0 [Ref. 3] |
| AMD MI355X (Llama2-70B cluster) | ~$0.06 (est.) | ~12,000 per GPU | MLPerf Inference v6.0 [Ref. 2] |
| Trillium v6e (GPT-3 175B on GCP) | ~$0.03–0.05 (est.) | ~8,000 per chip | Google Cloud pricing/benchmarks [Ref. 3] |
| B200 launch-day cost (Oct 2025) | $0.10 (baseline) | ~12,000 | SemiAnalysis, vs Apr 2026 $0.02 [Ref. 10] |

The 5x cost reduction from $0.10 to $0.02 per million tokens in just six months of Blackwell deployment demonstrates that software optimization — not hardware refreshes — is the primary driver of inference cost reduction within a generation. TensorRT-LLM kernel improvements, speculative decoding, and batching optimizations delivered this improvement on identical hardware.

## 9.4 Breakeven and ROI Analysis

- B200 hardware breakeven: A $30,000–35,000 GPU (average street price net of markup) breaks even against $6–8/hr cloud rental rates at approximately 60% utilization over 18 months — excluding datacenter space, power ($14 kW per DGX B200 system), cooling infrastructure (20–40% of hardware cost for DLC), and operations staff. [Ref. 7]
- GB200 NVL72 ROI: A $5 million investment in an NVL72 rack can generate $75 million in token revenue based on DSR1 model economics as measured in InferenceMAX v1 — a 15x return on investment. This assumes sustained high-utilization inference serving and prevailing commercial API pricing. [Ref. 10]
- AMD MI355X TCO advantage: For self-hosted inference clusters, the MI355X delivers approximately 33% lower TCO than HGX B200 equivalent configurations, driven primarily by lower hardware acquisition cost and higher memory capacity (reducing sharding overhead). [Ref. 14]
- Cloud ASIC token economics: For workloads that can accept GCP or AWS lock-in, Trillium v6e and Trainium3 offer 30–40% better price-performance than H100-based instances. For organizations with mixed on-premise and cloud workloads, the lock-in cost must be evaluated against the token economics advantage.
- Facility costs are hidden multipliers: Power infrastructure for liquid-cooled GPU deployments adds 20–40% to total deployment cost. A single NVL72 rack draws 120–140 kW — equivalent to 100+ typical homes. NVIDIA's 2027 Kyber racks will draw 600 kW per rack, requiring industrial-grade electrical infrastructure upgrades with 2–3 year lead times.

# 10. Vendor Roadmaps 2026–2028 — Capacity Wars Meet Portability Push

## 10.1 NVIDIA: Vera Rubin and Rubin Ultra

NVIDIA's Vera Rubin platform — announced at CES 2026 and entering production in H2 2026 — represents a fundamental architectural evolution from discrete GPU servers to multi-chip integrated AI supercomputers. The platform comprises seven new chips designed and optimized as a cohesive system.

| Component | Specification | Role in Platform |
|---|---|---|
| Rubin GPU (VR200) | 336B transistors; 288 GB HBM4; 50 PFLOPS NVFP4; 22 TB/s bandwidth; NVLink 6 at 3.6 TB/s/GPU | Primary compute engine; 2.5–5x Blackwell inference throughput [Ref. 20, 27] |
| Vera CPU | 227B transistors; 88 Arm Olympus cores; 1.5 TB LPDDR5x; 1.2 TB/s bandwidth; NVLink-C2C at 1.8 TB/s | CPU host for GPU domain; KV cache management; prefill/decode separation [Ref. 29] |
| NVLink 6 Switch | 6th-gen NVLink; 260 TB/s per NVL72 rack; 2x Blackwell bandwidth | Rack-scale GPU interconnect; unified memory domain |
| BlueField-4 DPU | ASTRA trusted resource architecture; network offload | Security isolation; RDMA acceleration; multi-tenant trust |
| ConnectX-9 NIC | 800 GbE (8x Blackwell's 400 GbE); Spectrum-X Ethernet AI fabric | Scale-out networking; replacing InfiniBand in many configs |
| Rubin CPX (LPX) | 128 GB SRAM; 40 PB/s memory bandwidth; 640 TB/s scale-up; 256 LPUs | Low-latency inference accelerator for agentic AI decode [Ref. 21] |
| Vera Rubin NVL72 rack | 72 Rubin GPUs + 36 Vera CPUs + full fabric; 3.6 exaFLOPS FP4; 10x Blackwell inference/watt | System-level product; AI factory unit of deployment |

Key Vera Rubin delivery timeline:

- Full production declared Q1 2026 (accelerated from original H2 2026 mass production target)
- Early access hyperscaler deployments: AWS, Google Cloud, Microsoft Azure, Oracle Cloud — all securing early allocations. [Ref. 25]
- CoreWeave production deployment: H2 2026, with customers able to bring Rubin into existing environments alongside Blackwell. [Ref. 25]
- Rubin R100 variant (next step): Sampling with select customers Q4 2026; volume production Q1 2027. [Ref. 20]
- Rubin Ultra (2027): 100 PFLOPS FP4 (2x standard Rubin); 1 TB HBM4e; Kyber rack at 600 kW per rack. This requires advance planning — industrial utility infrastructure with 2–3 year procurement timelines. [Ref. 28]
- Feynman (2028): Named after physicist Richard Feynman; expected ~1 MW per rack; design assisted by Blackwell-class AI. [Ref. 28]

## 10.2 AMD: MI400 Helios and MI500

AMD's roadmap from MI350 through Helios represents the company's transition from competitive individual GPU products to a full rack-scale platform capable of matching NVIDIA's NVL72 on performance metrics. The Helios double-wide Open Rack Wide v3 system was co-developed with Meta Platforms.

| Product | Architecture | Key Specs | Timeline | Status |
|---|---|---|---|---|
| MI350X / MI355X | CDNA 4 (3nm) | 288 GB HBM3e; 10.1 PFLOPS FP4; 8 TB/s; 1000/1400W TDP | GA 2025 | In production [Ref. 11, 13] |
| MI430X | CDNA 5 (Altair) | 432 GB HBM4; 19.6 TB/s; HPC + Sovereign AI focus; hardware FP64 | H2 2026 eng. samples | Roadmap [Ref. 16, 17] |
| MI440X | CDNA 5 | 8-GPU node variant; HBM4; enterprise on-premise focus | 2026–2027 | Roadmap |
| MI455X (Altair) | CDNA 5 | 432 GB HBM4; 40 PFLOPS FP4; 20 PFLOPS FP8; 300 GB/s scale-out/GPU | H2 2026 eng. samples | Mass prod Q2 2027 [Ref. 18] |
| Helios Rack (UALoE72) | MI455X + EPYC Venice + Pensando Vulcano NIC | 72 MI455X; 260 TB/s scale-up; UALink over Ethernet (Broadcom TH6); double-wide rack | H2 2026 samples | Mass prod Q2 2027 [Ref. 18] |
| MI500 series | CDNA 6 | HBM4e; 256-chip rack scale; ~2x MI455X performance | 2027 | Announced roadmap [Ref. 16] |

> Critical Helios execution risk: Per SemiAnalysis reporting (February 2026), AMD Helios engineering samples will be available H2 2026, but manufacturing delays push mass production and first production tokens to Q2 2027 at the earliest. Customers planning 2026 rack-scale AMD deployments should plan for a 6-month slip and maintain NVIDIA procurement optionality. [Ref. 18]

AMD's roadmap architecture decisions:

- UALink over Ethernet: AMD renamed its 'IF over Ethernet' protocol to 'UALink Protocol over Ethernet'. The MI400 Series will use Broadcom Ethernet Tomahawk 6 switches because Marvell and Astera Labs' native UALink switches will not be ready by late 2026. Despite this pragmatic workaround, scale-up bandwidth is competitive with NVIDIA VR200 NVL144. [Ref. 14]
- Scale-out networking gap: AMD will lag NVIDIA on scale-out networking through 2026. NVIDIA's ConnectX-9 NIC offers 800 GbE per GPU vs. AMD Vulcano NIC's 400 GbE per GPU — the Vulcano NIC mass deployment is not scheduled until H2 2026.
- Yotta-scale vision: At CES 2026, Lisa Su committed to a 1,000x performance increase over 2023 baselines by end of decade, positioning AMD as an architect of the world's most expansive AI factories. [Ref. 19]

## 10.3 Google: TPU v7 Ironwood and Post-Trillium

- TPU v7 Ironwood is narrowing the timeline gap to NVIDIA: prior generations lagged NVIDIA flagship chips by 2+ years; Ironwood is available within quarters of H100 equivalency. Peak theoretical FLOPs are approaching parity. [Ref. 35]
- Broadcom supply agreement: Anthropic has contracted Broadcom to supply 3.5 gigawatts of Google TPU capacity starting from 2027 — the largest publicly disclosed custom silicon commitment in the AI industry.
- Multi-pod scale: Google's Jupiter datacenter fabric continues to evolve toward larger Multislice pod configurations, enabling models that exceed single-pod memory capacity to train across pods with manageable ICI communication overhead.
- Post-Trillium roadmap: Not publicly disclosed beyond v7 Ironwood, but Google's internal compute requirements for Gemini successors will drive continued aggressive TPU development.

## 10.4 AWS: Next Trainium Generation

- Trainium4 and beyond: AWS has not publicly disclosed specifications for post-Trainium3 generations, but annual cadence investment suggests continued scaling of both compute density and memory capacity.
- UltraServer scale-up: The Trn3 UltraServer at 144 chips represents the current scale-up limit. Future generations will expand this, potentially enabling rack-scale Trainium deployments that compete with NVIDIA and AMD's NVL72/Helios configurations.
- Neuron SDK maturation: AWS is investing in JAX and additional framework support for Neuron SDK, reducing the current portability limitation that restricts Trainium adoption to AWS-native workloads.

## 11. Risk Matrix and Failure Modes

| Risk Category | Description | Severity | Probability | Mitigation |
|---|---|---|---|---|
| Software migration shortfall | HIPIFY fails on ~20% of custom fused kernels. ROCm gaps in attention and NCCL depth cause production performance regression vs. CUDA | High | High | Maintain CUDA fallback. Port only top 5% hotspot kernels manually. Run dual CI from day one. |
| Precision regression | FP4 accuracy drops on RAG retrieval, code generation, or mathematical reasoning tasks that were calibrated on FP8/FP16 | High | Medium | Enforce per-task golden-set evaluation. Implement selective precision by model head. |
| Facility power constraints | 140 kW NVL72, 1,400W MI355X require DLC infrastructure. 600 kW Kyber racks (2027) require industrial electrical | High | High (2027+) | Stage facility upgrades 18–24 months ahead. Burst to cloud ASICs during facility ramp. |
| HBM4 supply constraint | Both NVIDIA Rubin and AMD MI455X require HBM4. SK Hynix produces 62% of global HBM supply. Demand outpaces 2026 capacity. | High | High | Secure multi-vendor HBM4 allocations immediately. Dual-source from SK Hynix + Micron. |
| AMD Helios execution risk | Manufacturing delays push Helios mass production from H2 2026 to Q2 2027. 1 GW OpenAI deployment may face timeline risk. | Medium | High (already manifesting) | Include 6-month schedule buffer. Maintain NVIDIA B300/Rubin optionality in procurement. |
| Cloud ASIC portability lock-in | TPU/Trainium code requires full rewrite for portability. Exit costs increase with each generation of custom kernel investment. | Medium | Certain (by design) | Include portability milestones in cloud ASIC contracts. Maintain CUDA/ROCm CI as insurance. |
| TSMC concentration risk | TSMC manufactures 92% of advanced AI chips. Geopolitical events, natural disasters, or export restrictions could constrain supply globally. | Very High | Low (but existential) | Diversify across foundry suppliers where possible. Maintain 6+ month inventory buffer. |
| NVLink premium dependency | NVLink fabric creates ecosystem lock at the rack level. NVIDIA can price NVLink infrastructure | Medium | Ongoing | Evaluate AMD Helios / Intel Jaguar Shores for inference workloads that don't require NVLink-scale all-reduce. |

| Risk Category | Description | Severity | Probability | Mitigation |
|---|---|---|---|---|
| | components without constraint. | | | |

# 12. Strategic Recommendations and Action Framework

## 12.1 Platform Selection Decision Matrix

| Decision Criterion | Weight | Best Platform | Runner-Up | Rationale |
|---|---|---|---|---|
| Frontier LLM training speed | 30% | NVIDIA GB200/NVL72 | Google TPU v6e | NVLink bandwidth + CUDA maturity unmatched. CoreWeave 2,496-GPU cluster trained Llama3.1 405B in 27 minutes. |
| Cost-per-token (cloud) | 25% | Google Trillium / AWS Trainium3 | NVIDIA B200 | 1.8–2.8x better performance/$ in native cloud. Midjourney achieved 65% cost reduction. |
| Memory-bound inference | 20% | AMD MI355X | NVIDIA B300 | 288 GB HBM3E eliminates tensor parallelism sharding for 70B–200B models. 33% lower TCO. |
| Open standards / vendor independence | 15% | Intel Gaudi 3 (transitioning) | AMD MI355X + ROCm | Ethernet-first architecture. No NVLink premium. Jaguar Shores continues this philosophy. |
| Giant context / MoE / sparsity | 10% | Cerebras CS-3 | AMD MI355X (high HBM) | 21x advantage over DGX B200 for million-token contexts. Weight Streaming Engine eliminates HBM wall. |

## 12.2 90-Day Implementation Playbook

### Days 1–30: Inventory and Assessment

1. Map every production and development workload by: model size, context window requirements, latency SLOs, precision tolerance, and peak throughput requirements.
2. Identify the top 5 training workloads by compute cost. These are the candidates for AMD or cloud ASIC evaluation.
3. Audit CUDA codebase for custom kernels, CUDA Graphs usage, and low-level NCCL implementations. These are the migration risk surface areas.
4. Evaluate datacenter power and cooling capacity. Determine readiness for B300/NVL72 (120–140 kW), future MI455X Helios (comparable), and 2027 Kyber racks (600 kW).
5. Establish baseline inference cost ($/million tokens) for each production model on current hardware. This is the reference TCO for all subsequent evaluations.

### Days 31–60: Pilot Deployments

6. Spin up one AMD MI355X node for inference pilot. Run top-3 inference workloads; measure tokens/second, latency P99, and accuracy vs. CUDA baseline.
7. Pilot Google Trillium v6e or AWS Trainium3 for the highest-volume training workload living inside the respective cloud. Measure actual cost-per-run vs. H100 baseline.
8. Evaluate Cerebras CS-3 (if context windows >100K tokens are in roadmap) for a defined 2-week pilot with a concrete accuracy and throughput target.
9. Run HIPIFY against full CUDA codebase. Identify the specific failed kernels. Scope manual porting effort with engineering time estimates.

### Days 61–90: Contracts and Infrastructure

10. Finalize multi-vendor procurement strategy: NVIDIA GB200/B300 for frontier training + AMD MI355X or cloud ASICs for inference.
11. Implement OpenXLA / TorchXLA on GCP or Neuron SDK on AWS. Maintain active CUDA and ROCm CI pipelines from this point forward.
12. Secure HBM4 allocation visibility from both NVIDIA and AMD channel partners. Begin conversations now — 2026 HBM4 supply is already contracted by hyperscalers.
13. Include exit clauses in cloud ASIC contracts tied to portability milestones. Negotiate explicit software support SLAs for ROCm and SynapseAI stacks.

## 12.3 Hedge Positions for 2026–2027 Uncertainty

- Rubin vs. Helios timing uncertainty: NVIDIA Rubin production is accelerating ahead of schedule; AMD Helios slipping to Q2 2027. This favors a 2026 NVIDIA-heavy procurement with AMD Helios re-evaluation in Q2 2027.
- HBM4 supply: Lock in Rubin/MI455X allocations now if 2027 infrastructure is in plan. HBM4 supply constraints will be the gating factor, not silicon availability.
- Software portability: Invest in OpenXLA and dual CI pipelines regardless of hardware choices. The 2027–2028 landscape will include new entrants (Microsoft Maia successors, custom silicon from OpenAI, Cerebras second-generation WSE) — portability is insurance against current platform assumptions.
- Inference economics: NVIDIA's 5x cost reduction in 6 months demonstrates that software optimization within a generation outpaces hardware transitions. Budget for 2–3 software upgrade cycles per hardware generation in inference TCO models.

## 13. April 2026 Technical Intelligence Bulletin

**AI ACCELERATOR INTELLIGENCE // VOLUME 4, ISSUE 4 // MAY 1 2026 // COMPILED FROM PRIMARY SOURCES AND LIVE BENCHMARKS**

This bulletin aggregates the highest-signal technical developments in AI accelerator infrastructure recorded during April 2026. Coverage prioritizes silicon architecture disclosures, independently verified benchmark results, and supply-chain intelligence relevant to enterprise procurement and research infrastructure decisions. All figures are sourced from primary vendor disclosures, MLCommons official results, and confirmed analyst briefings unless otherwise noted.

### 13.1 MLPerf® Inference v6.0: Full Technical Debrief — April 1, 2026

#### BENCHMARK RELEASE OVERVIEW

MLCommons published MLPerf Inference v6.0 on April 1, 2026 — the consortium's self-described most significant inference benchmark revision to date. The round received 24 submitting organizations (up from 18 in v5.1), including first-time submitters Inventec Corporation, Netweb Technologies India Limited, and Stevens Institute of Technology. Five of eleven datacenter tests were new or substantially updated; the benchmark suite now covers multimodal vision-language models, text-to-video generation, a MoE reasoning model, an interactive inference scenario, and a transformer-based recommendation benchmark.

#### NEW WORKLOADS INTRODUCED IN V6.0

| Benchmark | Model | Scenario | Significance |
|---|---|---|---|
| DeepSeek-R1 Interactive | DeepSeek-R1 671B MoE | Server / Interactive | Tests TTFT P99 + sustained throughput under realistic latency constraints. First reasoning model in MLPerf Interactive tier. |
| GPT-OSS 120B | OpenAI GPT-OSS 120B MoE | Offline / Server / Interactive | 5.1B active params/token. Entire model fits in a single GB300 GPU (279 GB HBM3e). Replaces GPT-J as the primary LLM benchmark. |
| Qwen3-VL-235B-A22B | Qwen3 VL 235B MoE | Offline / Server | First multimodal vision-language model in closed-division MLPerf. Tested on NVIDIA via vLLM open-source stack. |
| Wan 2.2 T2V-A14B | Wan 2.2 Text-to-Video 14B | Single-stream latency | Generative video benchmark. Single-stream |

| Benchmark | Model | Scenario | Significance |
|---|---|---|---|
| | | | latency target: ≤21 seconds end-to-end. No server scenario in round 1. |
| DLRMv3 | Transformer-based recommender | Offline / Server | Replaces DLRMv2 with transformer backbone. Higher compute intensity per sample. 104,637 samples/sec on GB300 NVL72 offline. |

## NVIDIA BLACKWELL ULTRA: PLATFORM-LEVEL RESULTS

NVIDIA submitted results on the GB300 NVL72 (Blackwell Ultra) platform with 14 partner co-submissions — the largest partner participation for any single platform in MLPerf history. Partners included ASUS, Cisco, CoreWeave, Dell Technologies, GigaComputing, Google Cloud, HPE, Lenovo, Nebius, Netweb Technology, QCT, Red Hat, Supermicro, and Lambda. NVIDIA's cumulative MLPerf training + inference wins reached 291, a figure the company states is nine times higher than all other submitters combined since 2018.

| Workload | System Config | Mode | v6.0 Throughput | v5.1 Baseline | Delta |
|---|---|---|---|---|---|
| DeepSeek-R1 | 4× GB300 NVL72 (288 GPUs total) | Offline | 2,490,000 tok/s (system) | — | Largest MLPerf submission ever |
| DeepSeek-R1 | GB300 NVL72 (72 GPUs) | Server | 8,064 tok/s/GPU | 2,907 tok/s/GPU (v5.1) | +2.77× (6-month delta, same HW class) |
| GPT-OSS 120B | GB300 NVL72 | Offline | 1,050,000 tok/s | Not in v5.1 | Model fits in single-GPU HBM; no tensor parallelism needed |
| Llama 3.1 405B | GB300 NVL72 | Offline | B200 baseline ×1.21 | — | Lower relative gain due to bandwidth-bound, multi-GPU parallelism |
| Qwen3-VL-235B | GB300 NVL72 + vLLM | Offline | 79 samples/sec | Not in v5.1 | Only platform to submit in v6.0 round 1 for this workload |
| Wan 2.2 T2V | GB300 NVL72 + TRT-LLM VisualGen | Single-stream | 21 s latency target met | Not in v5.1 | Only platform to submit text-to-video in round 1 |

## THE SOFTWARE DIVIDEND: CUDA 13.1 + TRT-LLM 1.2 ON IDENTICAL HARDWARE

A technically critical signal from v6.0: Lambda's NVIDIA HGX B200 (8-GPU) submission demonstrated 9% throughput improvement over the best v5.1 B200 submissions on identical

hardware, with zero silicon change. The delta is attributable entirely to software: prior round used TRT-LLM 1.0 on CUDA 12.8–12.9; v6.0 runs TRT-LLM 1.2 on CUDA 13.1. This confirms that the published 2.77× DeepSeek-R1 gain is not purely a Blackwell Ultra hardware story — it is a compound of architectural improvement (~1.29× from GB300 vs. B200 hardware alone) plus six months of stack maturity.

**Key implication:** Organizations running HGX B200 systems today can capture measurable performance gains by updating TRT-LLM to 1.2+ on CUDA 13.1 before next hardware procurement cycle. Stack updates on existing hardware are now a first-class TCO lever.

### BLAZE: SMART EXPERT ROUTING FOR MOE MODELS (OPEN DIVISION)

Lambda's open-division submission introduced BLAZE, a drop-in NVIDIA TRT-LLM module implementing Smart Expert Routing for MoE inference. BLAZE adds 0.1% overhead per decode step but steers traffic away from overloaded experts, cutting TTFT P99 from 2,903 ms to 2,001 ms (−31%) and lifting throughput by 3.9% on the DeepSeek-R1 server scenario. BLAZE is precision-agnostic (tested at FP4 weights + FP8 KV cache) and requires no recompilation — it operates as a runtime policy layer above the TRT-LLM execution engine.

### AMD MI355X IN V6.0: ECOSYSTEM BREADTH MILESTONE

AMD submitted MI355X results in the closed division for LLaMA 2 70B and GPT-OSS 120B, plus open-division ROCm-optimized results for SDXL and YOLOv11. Nine partner organizations co-submitted on AMD hardware, tying AMD's all-time record for partner participation in a single MLPerf round. Per-GPU throughput on GPT-OSS 120B reached approximately 1M tokens/sec in offline mode — competitive with prior-generation Blackwell (non-Ultra). AMD's submission signals that ROCm 7.x is now production-ready for the top-4 inference workloads by deployment volume.

**Gap analysis:** AMD did not submit for Qwen3-VL, Wan 2.2 T2V, or DLRMv3 in v6.0 round 1. These omissions indicate ROCm stack gaps in multimodal, video generation, and transformer-based recommendation — workloads that are structurally difficult to port without CUDA-native library equivalents. AMD's 'Advancing AI 2026' event (July 2026, post-Computex) is expected to address the multimodal inference roadmap for ROCm 7.1.

### INTEL ARC PRO B-SERIES: EDGE INFERENCE ENTRY

Intel submitted MLPerf v6.0 results using Xeon 6 CPUs and Arc Pro B-series GPUs, explicitly not submitting Gaudi 3 for this round. The Arc Pro submission targets professional workstation and edge inference server deployments — a market segment that neither NVIDIA's GB300 NVL72 nor AMD's MI355X serves. No Gaudi 3 results are available in the official v6.0 closed-division dataset. Gaudi 3 is expected to reappear in v6.1 (estimated Q3 2026) with INT4 quantization support enabled by Habana SynapseAI 1.21+.

## 13.2 Google Cloud Next 2026 (April 22–24, Las Vegas): Complete Silicon Disclosure

### CONFERENCE OVERVIEW

Google Cloud Next 2026 was held April 22–24, 2026, at the Mandalay Bay Convention Center, Las Vegas. The conference keynote delivered the most substantial Google silicon disclosure since the 2019 TPU v3 announcement: Ironwood (TPU v7) reached general availability, and TPU v8 was formally previewed as a two-chip family — the first time in Google's TPU history that a generation ships two architecturally distinct accelerators, one optimized for training (TPU 8t / 'Sunfish') and one for inference (TPU 8i / 'Zebrafish').

### TPU V7 IRONWOOD: GENERAL AVAILABILITY SPEC CONFIRMATION

Ironwood, announced in preview at I/O 2025, reached general availability for Google Cloud customers at Cloud Next 2026. Google positions Ironwood as 'the first TPU for the age of inference,' reflecting the industry-wide shift from training-dominated to inference-dominated accelerator cycles. The device delivers 4.6 petaFLOPS FP8 per chip, matching NVIDIA's B200 on single-chip peak compute. Ironwood's architectural advantage is at cluster scale: a 9,216-chip superpod delivering 42.5 exaFLOPS in a single memory domain via 9.6 Tb/s ICI networking and 1.77 PB of aggregated HBM. Ironwood is the largest-scale single-domain inference cluster commercially available as of May 2026.

| Specification | Ironwood (TPU v7) | NVIDIA B200 | Delta / Notes |
|---|---|---|---|
| Peak compute (FP8) | 4.6 petaFLOPS | 4.5 petaFLOPS | Comparable at single-chip level |
| HBM capacity | 192 GB | 192 GB | Parity |
| Scale-up bandwidth (ICI) | 9.6 Tb/s | 14.4 Tb/s (NVLink) | NVIDIA leads on single-chassis BW |
| FP4 precision | Not supported | Supported | NVIDIA delivers 2× throughput advantage on FP4-quantized models |
| Max cluster (single domain) | 9,216 chips / 1.77 PB HBM | GB300 NVL72: 72 GPUs / 5.09 PB HBM aggregate | Google leads on homogeneous cluster topology at 9K+ scale |
| Performance/watt vs. H100 | ~2.8× improvement | ~2× improvement | Google's energy efficiency lead validated by internal Gemini benchmarks |
| Anchor customer | Anthropic (1M chips, 1+ GW in 2026) | Multiple hyperscalers | Anthropic's full inference stack runs on Ironwood |

### TPU V8 ARCHITECTURE DISCLOSURE: SUNFISH (8T) AND ZEBRAFISH (8I)

The TPU v8 preview at Cloud Next 2026 marks a strategic inflection: Google is explicitly bifurcating its silicon roadmap into training-optimized and inference-optimized ASICs, reflecting

the architectural reality that the memory access patterns, precision requirements, latency sensitivity, and batch behaviors of training and inference are sufficiently divergent to justify purpose-built silicon for each.

| Parameter | TPU 8t 'Sunfish' (Training) | TPU 8i 'Zebrafish' (Inference) |
|---|---|---|
| Design partner | Broadcom (extends 2031 agreement) | MediaTek (new relationship) |
| Fabrication process | TSMC 2nm (N2) | TSMC 2nm (N2) |
| Target availability | Late 2027 (external customers) | Late 2027 (external customers) |
| Peak compute (FP4) | ~121 exaFLOPS (9,600-chip superpod) | 10.1 petaFLOPS per chip |
| Cluster topology | 9,600-chip superpod / 2 PB shared HBM | 1,152-chip pod / Boardfly topology |
| On-chip SRAM | Not disclosed | 384 MB (3× Ironwood's 128 MB) |
| HBM capacity per chip | ~30% higher BW than Ironwood | 288 GB / 8.6 TB/s |
| Inter-chip bandwidth | 2× Ironwood ICI | 19.2 Tb/s (2× Ironwood; Boardfly reduces hop count 56%) |
| Training gain vs. Ironwood | 2.8× training price-performance | — |
| Inference gain vs. Ironwood | — | 80% better inference $/perf |
| Primary workload target | Frontier model pretraining (10T+ param scale) | Agent serving: MoE, KV-cache-heavy, TTFL-sensitive |
| Key architectural feature | Doubled ICI; 2 PB shared HBM domain | Boardfly topology; 3× on-chip SRAM; Collectives Acceleration Engine |
| DeepMind co-design signal | World model training (Genie 3 class) | Swarm agent inference (millions of concurrent agents) |

## SUPPLY CHAIN ARCHITECTURE: MULTI-VENDOR SILICON STRATEGY

Google's TPU v8 disclosure confirms the most structurally complex custom-silicon supply chain ever fielded by a single cloud provider. The multi-vendor model assigns each design partner a workload profile matched to their engineering competency: Broadcom handles high-density training logic (its strength from network ASIC design); MediaTek contributes mobile-derived power efficiency and high-SRAM architectures (its strength from edge inference silicon). A third engagement with Marvell — targeting a memory processing unit and a potential fourth inference TPU variant — is reportedly in early negotiation. Intel provides Axion-adjacent CPU-side compute via custom IPU co-development. Google's dependence on a single ASIC partner (previously Broadcom alone) is structurally eliminated beginning with the v8 generation.

**Financial scale:** Broadcom's AI revenue attributable to Google and Anthropic is projected at $21 billion in 2026 and $42 billion in 2027 (Mizuho estimates cited in Google's Cloud Next briefing). A single 9,600-chip TPU 8t superpod carries an estimated $38M+ in Broadcom ASIC content at $4,000+ per chip.

#### VIRGO NETWORK & DATA CENTER CO-DESIGN

Announced alongside the TPU v8 preview: Virgo Networking, a fourth-generation data-center fabric co-engineered with the TPU v8 ICI topology. Virgo enables 10 TB/s of storage throughput per rack via integrated Google Cloud Managed Lustre, a direct response to memory-bandwidth-bound workload profiles in agentic multi-agent inference. Google also disclosed that data center electrical efficiency has improved 6× per unit of electricity consumed versus five years ago — a compound improvement attributable to liquid cooling integration, integrated power management that dynamically adjusts draw based on real-time utilization, and the chip-level efficiency gains in each TPU generation.

#### ANTHROPIC AS IRONWOOD'S LOAD-BEARING CUSTOMER

Google confirmed at Cloud Next 2026 that Anthropic's infrastructure commitment has expanded: 3.5 gigawatts of next-generation TPU compute (covering both Ironwood and early TPU 8t capacity) comes online by 2027. In Q1 2026, Anthropic crossed $30 billion in annualized revenue, surpassing OpenAI's $25 billion — with 80% of that revenue generated by enterprise API usage running on Google Cloud TPUs. Ironwood's per-token inference efficiency is now a direct input variable in Claude's competitive pricing versus GPT-4o and Gemini 3.1. Google Cloud's AI compute revenue grew 71% year-over-year in Q1 2026; TPU utilization exceeded 90% at every major data center region since January 2026.

## 13.3 AMD MI400 Series: April 2026 Status, TAM Projections, and Competitive Position

#### PRODUCTION TIMELINE AND REVENUE FORECAST

AMD confirmed in early April 2026 that its 'Advancing AI 2026' technical event will be held in July 2026, post-Computex, with the event focused on MI400 series production timelines, pricing, and benchmark results. Mid-2026 production shipments of MI400 are confirmed; Helios rack-scale systems (double-wide, MI455X GPU pairs, UALoE72 scale-up interconnect) remain guided for Q3 2026 per February 2026 disclosures. S&P Global Market Intelligence projects $7.2 billion in MI400 revenue in 2026, based on approximately 258,000 units at an average selling price of ~$30,926 per unit. AMD's total data center GPU revenue is projected to reach $15 billion in 2026, a 114% year-over-year increase, driven primarily by the MI400 ramp.

#### UALINK CONSORTIUM: OPEN INTERCONNECT VS. NVLINK

AMD's MI400 Helios deployment depends critically on the UALink (Ultra Accelerator Link) open interconnect standard, backed by AMD, Intel, Google, Meta, Microsoft, and Broadcom. UALink v1.0 provides GPU-to-GPU scale-up bandwidth as a non-proprietary alternative to NVIDIA's NVLink. The UALoE72 topology in Helios directly connects 72 MI455X GPUs in a single scale-

up domain without requiring NVIDIA's proprietary NVSwitch fabric. UALink's commercial viability is the key technical risk for Helios deployment: if UALink 1.0 switch latency or bandwidth density underperforms NVSwitch 4.0 in production multi-GPU inference scenarios, the total system performance case for Helios vs. GB300 NVL72 weakens materially.

#### EQUITIES SIGNAL

AMD shares traded at approximately $245 on April 10, 2026, up from $198 at the end of March, driven by broader AI sector momentum and MI400 ramp anticipation. AMD's market capitalization at that date was approximately $400 billion — a 7.5× discount to NVIDIA's ~$3 trillion valuation at equivalent period, indicating that institutional investors ascribe an 87% probability weight to continued NVIDIA platform dominance in enterprise AI through the MI400 ramp cycle.

## 13.4 AI Chip Startup Funding and Alternative Architecture Signals

#### RECORD CAPITAL DEPLOYMENT IN NON-GPU INFERENCE SILICON

AI chip startup funding in 2026 reached $8.3 billion globally through Q1 (Dealroom data), with the sector projected to exceed all prior annual records by year-end. The dominant thesis across funded companies is that NVIDIA GPUs — while architecturally capable — were not purpose-designed for high-throughput inference and carry energy and cost overhead that purpose-built silicon can eliminate. Key April 2026 funding events and the architectural bets behind them:

| Company | Round / Amount | Architecture Thesis | Technical Differentiator |
|---|---|---|---|
| Cerebras Systems | $1B (Feb 2026) | Wafer-Scale Engine: single silicon die eliminates inter-chip latency and HBM bandwidth bottleneck | WSE-3: 900,000 cores on 46,225 mm² die; 44 GB on-chip SRAM; Weight Streaming eliminates HBM wall for large LLMs |
| MatX | $500M (2026) | Matrix-native instruction set; eliminates GPU ISA translation overhead for pure transformer workloads | Custom ISA with fused attention + MLP primitives; sub-10µs kernel dispatch latency |
| Ayar Labs | $500M (2026) | Optical I/O: replaces copper SerDes with photonic interconnects, targeting 1–2 pJ/bit vs. 10–30 pJ/bit for copper | Co-packaged optical I/O chips enabling 4–8× interconnect energy reduction for multi-chip inference clusters |
| Etched | $500M (2026) | Transformer-specific ASIC: hardcoded attention circuits, no programmable logic overhead | ASICs burned for a specific transformer variant; 10–40× FLOPS/watt vs. general GPU for exact model architecture |
| Axelera (EU) | $200M+ (2026) | Near-memory compute: eliminates DRAM bandwidth wall by | METIS architecture with in-memory multiplication arrays; targets edge and disaggregated inference scenarios |

| Company | Round / Amount | Architecture Thesis | Technical Differentiator |
|---|---|---|---|
| | | processing data where it resides | |

The structural pattern across these investments: each targets a specific inefficiency in GPU-based inference — on-chip memory capacity (Cerebras), ISA overhead (MatX), interconnect energy (Ayar Labs), programmability overhead (Etched), and memory bandwidth (Axelera). None of these architectures is positioned to displace NVIDIA in frontier model training, where CUDA ecosystem depth is the dominant lock-in mechanism. All are positioned at inference, where the workload's memory-access pattern and latency requirements differ structurally from training.

## 13.5 Google Cloud × NVIDIA: Fractional G4 VMs and Vera Rubin NVL72 Preview

### G4 FRACTIONAL VM: VGPU FOR ENTERPRISE AI RIGHT-SIZING

At GTC 2026 (March) and confirmed at Cloud Next 2026 (April), Google Cloud announced preview availability of fractional G4 Virtual Machines powered by NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs via NVIDIA vGPU technology. Fractional G4 VMs allow customers to partition a physical GPU into smaller vGPU slices, enabling sub-GPU granularity for inference workloads that do not require full 96 GB VRAM — specifically fine-tuned models in the 30B–100B parameter range. The FP4 precision path on RTX PRO 6000 Blackwell Server Edition enables real-time, low-latency multimodal agent inference on the fractional configuration without full-GPU provisioning.

### VERA RUBIN NVL72: FIRST-WAVE CLOUD PROVIDER COMMITMENT

Google Cloud confirmed at Cloud Next 2026 that it will be among the first cloud providers to offer NVIDIA Vera Rubin NVL72 rack-scale systems in the second half of 2026, integrated into its AI Hypercomputer architecture. Vera Rubin NVL72 combines 72 Rubin GPUs with 36 Vera Arm CPUs on a shared 260 TB/s NVLink fabric, delivering approximately 3–4× AI compute density improvement over Blackwell, with CG-HBM (combined GPU-HBM memory, where HBM is stacked directly on the GPU die) eliminating the off-package HBM bandwidth bottleneck that limits current Blackwell inference throughput on memory-bound workloads. Rubin is particularly suited for Mixture-of-Experts model architectures, where expert selection patterns create irregular memory access that benefits from reduced data movement distance.

### DYNAMO + GKE INFERENCE GATEWAY INTEGRATION

Google Cloud announced the integration of NVIDIA Dynamo (open-source distributed inference framework) with GKE Inference Gateway — a modular, open-source control plane enabling hardware-agnostic inference serving across accelerator types. The integration allows inference workloads to span NVIDIA GPUs, Google TPUs, and future accelerator types under a unified orchestration layer, reducing vendor lock-in at the software control plane level even when the underlying silicon is hardware-specific. This is architecturally significant: it validates the

emerging 'inference routing fabric' model, where LLM serving infrastructure must dynamically allocate requests across heterogeneous accelerator pools based on latency SLO, throughput requirements, and cost constraints.

## 13.6 Signal Summary and Forward Indicators

### FIVE HIGH-CONVICTION SIGNALS FROM APRIL 2026

| Signal | Data Source | Infrastructure Implication |
|---|---|---|
| Software is now a first-class performance lever | MLPerf v6.0: 9% gain on identical B200 hardware from TRT-LLM 1.2 + CUDA 13.1 | Stack upgrades on existing hardware must be budgeted as a TCO-reduction strategy, not just a maintenance event. 9% throughput = 9% cost reduction on inference infrastructure. |
| Google's TPU v8 bifurcation signals end of the general-purpose AI accelerator | Cloud Next 2026: TPU 8t (Broadcom, training) + TPU 8i (MediaTek, inference) announced as distinct chips from the ground up | The inference-vs-training workload split is now architecturally formalized at the silicon level by the world's largest custom ASIC operator. Buyers building for 2027+ should model training and inference infrastructure as separate procurement decisions. |
| TSMC 2nm is the 2027 frontier node — capacity allocation is the binding constraint | TPU 8t/8i both target TSMC N2 for late-2027 availability; NVIDIA Rubin Ultra also targets N2 in 2027 | TSMC 2nm capacity is fully contracted by Google (v8), NVIDIA (Rubin Ultra), and Apple (A21) through 2027. New entrants without existing TSMC relationships cannot access N2 for AI ASICs before 2028. |
| AMD's commercial viability depends on UALink hitting production specs by Q3 2026 | AMD Helios guided Q3 2026; UALink 1.0 switch specifications not publicly verified in production configurations | If UALink 1.0 latency or bandwidth density in production deviates from spec sheet values, Helios rack-scale TCO vs. NVIDIA GB300 NVL72 weakens materially. Procurement teams should require UALink v1.0 production validation data before committing to Helios orders. |
| Inference-specific silicon investment has crossed the $8B+ annual threshold | Dealroom 2026 funding data; Cerebras $1B, MatX $500M, Ayar Labs $500M, Etched $500M, Axelera $200M+ in 2026 alone | The inference-silicon startup ecosystem has achieved a scale where one or more alternative inference ASICs will reach production viability by 2027–2028. Enterprises should pilot at least one non-GPU inference path in 2026 to build architectural portability before alternatives mature. |

### KEY FORWARD INDICATORS TO MONITOR (MAY–JULY 2026)

- AMD 'Advancing AI 2026' event (July 2026): First official MI400 production benchmarks and Helios UALink bandwidth validation against NVSwitch. This is the decisive event for AMD's 2026 competitive positioning.
- NVIDIA Vera Rubin NVL72 cloud availability (H2 2026): First cloud provider availability confirmation (Google Cloud, CoreWeave, Lambda) will set production pricing benchmarks and validate CG-HBM performance claims on customer workloads.

- MLPerf Training v5.1 / Inference v6.1 (Q3 2026 estimated): AMD and Intel will be expected to submit results on updated workloads including Qwen3-VL multimodal, Wan 2.2 text-to-video, and DLRMv3 — closing the submission gap visible in v6.0.
- Google TPU 8t/8i general availability timeline: Cloud Next 2026 confirmed preview status; GA for external customers is targeted late 2027. Any acceleration of this timeline (e.g., early access program for Anthropic or external partners) would be a significant positive signal for Google Cloud market share in AI.
- TSMC CoWoS-L capacity (HBM4 on-package): HBM4 allocation visibility from NVIDIA and AMD channel partners will determine whether the Rubin and MI455X ramps are supply-constrained or demand-constrained in H1 2027.

## APRIL 2026 BULLETIN REFERENCES

This bulletin cites the following primary sources in addition to all references listed in Section 14. All URLs verified as of May 1, 2026:


- [A1] MLCommons. MLPerf® Inference v6.0 Results. April 1, 2026. https://mlcommons.org/2026/04/mlperf-inference-v6-0-results/
- [A2] NAND Research. MLPerf Inference 6.0: Software Gains & Broadening Competition. April 2026. https://nand-research.com/mlperf-inference-6-0-software-gains-broadening-competition-shake-things-up/
- [A3] Lambda. Lambda's MLPerf Inference v6.0: Hardware Leap, Software Maturity, Research Breakthrough. April 2026. https://lambda.ai/blog/lambdas-mlperf-inference-v6.0-hardware-leap-software-maturity-research-breakthrough
- [A4] StorageReview. NVIDIA Sets MLPerf Inference v6.0 Records with Blackwell Ultra Platform. April 2026. https://www.storagereview.com/news/nvidia-sets-mlperf-inference-v6-0-records-with-blackwell-ultra-platform
- [A5] Google Cloud Blog. Welcome to Google Cloud Next '26. April 22, 2026. https://cloud.google.com/blog/topics/google-cloud-next/welcome-to-google-cloud-next26
- [A6] Google Cloud Blog. Eighth Generation TPUs: Two Chips for the Agentic Era. April 22, 2026. https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/eighth-generation-tpu-agentic-era/
- [A7] The Next Web. Google Launches Ironwood TPU and Previews Eighth-Gen Split at TSMC 2nm. April 22, 2026. https://thenextweb.com/news/google-ironwood-tpu-inference-cloud-next
- [A8] Hyperframe Research. Google Cloud Next 2026: TPU 8t and 8i Architectures Signal the End of General-Purpose Silicon. April 22, 2026. https://hyperframeresearch.com/2026/04/22/google-cloud-next-2026-google-cloud-bifurcates-the-ai-future
- [A9] Tech Insider. Google TPU 8t and 8i: 121 Exaflops, $21B Nvidia Challenge. April 2026. https://tech-insider.org/google-tpu-8t-8i-broadcom-mediatek-nvidia-2026/

- [A10] Oplexa. Google Cloud Next 2026: Every Major Announcement. April 2026. https://oplexa.com/google-cloud-next-2026/
- [A11] Tech Insider. AMD MI400 Series: $7.2B AI GPU Challenging Nvidia. April 2026. https://tech-insider.org/amd-mi400-series-ai-gpu-data-center-2026/
- [A12] CNBC. Nvidia AI Chip Rivals Attract Record Funding. April 17, 2026. https://www.cnbc.com/2026/04/17/nvidia-ai-chip-rivals-funding-euclyd-fractile.html
- [A13] Google Cloud Blog. Google Cloud AI Infrastructure at NVIDIA GTC 2026. March 16, 2026. https://cloud.google.com/blog/products/compute/google-cloud-ai-infrastructure-at-nvidia-gtc-2026
- [A14] Futurum Group. At CES, NVIDIA and AMD Made Memory the Future of AI. January 12, 2026. https://futurumgroup.com/insights/at-ces-nvidia-rubin-and-amd-helios-made-memory-the-future-of-ai/

## 14. Complete References

All references below are cited throughout this document using bracketed notation [Ref. N]. References 1–18 originate from the source playbook document; References 19–35 represent additional primary and secondary sources consulted during deep research for this report.

### SOURCE DOCUMENT REFERENCES [REF. 1–18]


14. [Ref. 1] Silicon Analysts. NVIDIA AI GPU Market Share 2026: ~80% of AI Accelerators. April 2026. Available at: https://siliconanalysts.com/analysis/nvidia-ai-accelerator-market-share-2024-2026
15. [Ref. 2] Advanced Micro Devices. AMD Delivers Breakthrough MLPerf Inference 6.0 Results. AMD Corporate Blog, April 2026. Available at: https://www.amd.com/en/blogs/2026/amd-delivers-breakthrough-mlperf-inference-6-0-results.html
16. [Ref. 3] Google Cloud. Tensor Processing Units (TPUs). Google Cloud Documentation, 2026. Available at: https://cloud.google.com/tpu
17. [Ref. 4] Amazon Web Services. Announcing Amazon EC2 Trn3 UltraServers for faster, lower-cost generative AI training. AWS Blog, December 2025. Available at: https://aws.amazon.com/about-aws/whats-new/2025/12/amazon-ec2-trn3-ultraservers/
18. [Ref. 5] Advanced Micro Devices. AMD and OpenAI Announce Strategic Partnership to Deploy 6 Exaflops of AMD Instinct GPU Compute. AMD Newsroom, October 6, 2025. Available at: https://www.amd.com/en/newsroom/press-releases/2025-10-6-amd-and-openai-announce-strategic-partnership-to-d.html
19. [Ref. 6] Reuters. Meta in talks to spend billions on Google's chips. Reuters Business, November 25, 2025. Available at: https://www.reuters.com/business/meta-talks-spend-billions-googles-chips-information-reports-2025-11-25
20. [Ref. 7] GetDeploying. GB200 Cloud Pricing: Compare 9+ Providers (2026). Available at: https://getdeploying.com/gpus/nvidia-gb200
21. [Ref. 8] Oracle Cloud Infrastructure. Announcing General Availability of OCI Compute with AMD Instinct MI355X. Oracle Blogs, 2026. Available at: https://blogs.oracle.com/cloud-infrastructure/announcing-general-availability-of-oci-amd-mi355x
22. [Ref. 9] Google Cloud. TPU Pricing. Google Cloud Documentation, 2026. Available at: https://cloud.google.com/tpu/pricing
23. [Ref. 10] NVIDIA Corporation. NVIDIA Blackwell Raises Bar in New InferenceMAX Benchmarks, Delivering Unmatched Performance and Lowest Cost Per Token. NVIDIA Developer Blog, October 2025; updated April 2026. Available at: https://blogs.nvidia.com/blog/blackwell-inferencemax-benchmark-results/

24. [Ref. 11] CoreWeave. CoreWeave, NVIDIA, and IBM Set MLPerf Record with Largest NVIDIA GB200 Blackwell Cluster, Achieving Over 2x Faster Training. CoreWeave Blog, 2025. Available at: https://www.coreweave.com/blog/coreweave-nvidia-and-ibm-set-mlperf-record-with-largest-nvidia-gb200-blackwell-cluster-achieving-over-2x-faster-training
25. [Ref. 12] AMD. ROCm 7.0: An AI-Ready Powerhouse for Performance, Efficiency, and Innovation. AMD ROCm Blog, September 2025. Available at: https://rocm.blogs.amd.com/ecosystems-and-partners/rocm-7.0-blog/README.html
26. [Ref. 13] OpenXLA Project. OpenXLA: Open Source ML Compiler Ecosystem. Available at: https://openxla.org/
27. [Ref. 14] NVIDIA Corporation. GB200 NVL72 Product Page. NVIDIA Data Center, 2026. Available at: https://www.nvidia.com/en-us/data-center/gb200-nvl72/
28. [Ref. 15] Supermicro. Supermicro with GAUDI 3 AI Delivers Scalable AI Infrastructure. Supermicro White Paper. Available at: https://www.supermicro.com/white_paper/white_paper_Gaudi3_Reference_Archtecture.pdf
29. [Ref. 16] Amazon Web Services. Gen AI Compute Instance – Amazon EC2 Trn3 UltraServers. AWS Product Documentation, 2025. Available at: https://aws.amazon.com/ec2/instance-types/trn3
30. [Ref. 17] Introl Blog. NVIDIA FP4 Inference: 50x Energy Efficiency. Available at: https://introl.com/blog/fp4-inference-efficiency-nvidia-2025
31. [Ref. 18] Next Platform. AMD Says 'Helios' Racks And MI400 Series GPUs On Track For 2H 2026. NextPlatform.com, February 23, 2026. Available at: https://www.nextplatform.com/compute/2026/02/23/amd-says-helios-racks-and-mi400-series-gpus-on-track-for-2h-2026/4092199

## ADDITIONAL RESEARCH REFERENCES [REF. 19–35]

32. [Ref. 19] Mordor Intelligence. AI Accelerators Market Size, Share & Growth Forecast Report 2030. 2026. Available at: https://www.mordorintelligence.com/industry-reports/ai-accelerators-market
33. [Ref. 20] Tech Insider / Various. NVIDIA Rubin GPU: 336B Transistors, Production Updates [2026]. April 2026. Available at: https://tech-insider.org/nvidia-gtc-2026-rubin-gpu-analysis/
34. [Ref. 21] NVIDIA Corporation. Infrastructure for Scalable AI Reasoning | NVIDIA Vera Rubin Platform. NVIDIA Product Pages, 2026. Available at: https://www.nvidia.com/en-us/data-center/technologies/rubin/
35. [Ref. 22] Tom's Hardware. Nvidia announces Rubin GPUs in 2026, Rubin Ultra in 2027, Feynman also added to roadmap. March 18, 2025. Available at:

https://www.tomshardware.com/pc-components/gpus/nvidia-announces-rubin-gpus-in-2026-rubin-ultra-in-2027-feynam-after

36. [Ref. 23] NVIDIA Newsroom. NVIDIA Kicks Off the Next Generation of AI With Rubin — Six New Chips, One Incredible AI Supercomputer. 2026. Available at: https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer
37. [Ref. 24] NVIDIA Technical Blog. Inside the NVIDIA Vera Rubin Platform: Six New Chips, One AI Supercomputer. NVIDIA Developer Blog, January 2026. Available at: https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer/
38. [Ref. 25] AMD Blogs. AMD Instinct MI350 Series and Beyond: Accelerating the Future of AI and HPC. AMD Corporate Blog, October 2025. Available at: https://www.amd.com/en/blogs/2025/amd-instinct-mi350-series-and-beyond-accelerating-the-future-of-ai-and-hpc.html
39. [Ref. 26] Data Center Dynamics. AMD launches Instinct MI350 GPUs, unveils double-wide Helios AI rack-scale system. February 2026. Available at: https://www.datacenterdynamics.com/en/news/amd-launches-instinct-mi350-gpus-unveils-double-wide-helios-ai-rack-scale-system/
40. [Ref. 27] SemiAnalysis. AMD Advancing AI: MI350X and MI400 UALoE72, MI500 UAL256. SemiAnalysis Newsletter, June 2025. Available at: https://newsletter.semianalysis.com/p/amd-advancing-ai-mi350x-and-mi400-ualoe72-mi500-ual256
41. [Ref. 28] AMD Newsroom. AMD Accelerates Pace of Data Center AI Innovation and Leadership with Expanded AMD Instinct GPU Roadmap. AMD Press Release, June 2024. Available at: https://www.amd.com/en/newsroom/press-releases/2024-6-2-amd-accelerates-pace-of-data-center-ai-innovation-.html
42. [Ref. 29] VideoCardz. NVIDIA Vera Rubin NVL72 Detailed: 72 GPUs, 36 CPUs, 260 TB/s Scale-Up Bandwidth. VideoCardz.com, January 2026. Available at: https://videocardz.com/newz/nvidia-vera-rubin-nvl72-detailed-72-gpus-36-cpus-260-tb-s-scale-up-bandwidth
43. [Ref. 30] Various / Compiled. Google TPU v6 Trillium vs Nvidia Blackwell: Scaling AI Infrastructure Economics in 2026. April 2026. Available at: https://www.podcastvideos.com/articles/google-tpu-vs-nvidia-blackwell-ai-dominance/
44. [Ref. 31] Google Cloud. Introducing Trillium, sixth-generation TPUs. Google Cloud Blog, May 2024. Available at: https://cloud.google.com/blog/products/compute/introducing-trillium-6th-gen-tpus
45. [Ref. 32] MLQ.ai. AI for Investors: AI Chip Market Analysis. MLQ.ai Research, 2025. Available at: https://mlq.ai/research/ai-chips/
46. [Ref. 33] Thunder Compute. ROCm vs CUDA: Which GPU Computing System Wins in April 2026? ThunderCompute Blog, March 2026. Available at: https://www.thundercompute.com/blog/rocm-vs-cuda-gpu-computing

47. [Ref. 34] AMD ROCm GitHub. AMD ROCm Release Notes v7.x. GitHub/ROCm, 2025–2026. Available at: https://rocm.docs.amd.com/en/latest/about/release-notes.html
48. [Ref. 35] SemiAnalysis. Google TPUv7: The 900lb Gorilla In the Room. SemiAnalysis Newsletter, November 2025. Available at: https://newsletter.semianalysis.com/p/tpuv7-google-takes-a-swing-at-the

## BENCHMARK METHODOLOGY REFERENCES

- MLPerf Inference v4.1, v5.0, v6.0 — MLCommons. Available at: https://mlcommons.org/benchmarks/inference-datacenter/
- MLPerf Training v3.1, v4.0, v4.1, v5.0 — MLCommons. Available at: https://mlcommons.org/benchmarks/training/
- SemiAnalysis InferenceMAX v1 — Independent benchmark measuring total cost of compute across diverse models. April 2026.
- NVIDIA DGX B200 Product Page — Official specifications and performance data. Available at: https://www.nvidia.com/en-us/data-center/dgx-b200/
- AMD MLPerf Submission Documentation — Available at: https://www.amd.com/en/blogs/2026/amd-delivers-breakthrough-mlperf-inference-6-0-results.html
- AWS EC2 Trn3 UltraServer Specifications — Available at: https://aws.amazon.com/ec2/instance-types/trn3
- Google Cloud TPU Documentation — Available at: https://cloud.google.com/tpu and https://cloud.google.com/tpu/pricing
- IDC Worldwide Artificial Intelligence and Automation Spending Guide, 2025–2030.
- Mordor Intelligence AI Accelerators Market Size, Share & Growth Forecast Report 2030 — Available at: https://www.mordorintelligence.com/industry-reports/ai-accelerators-market