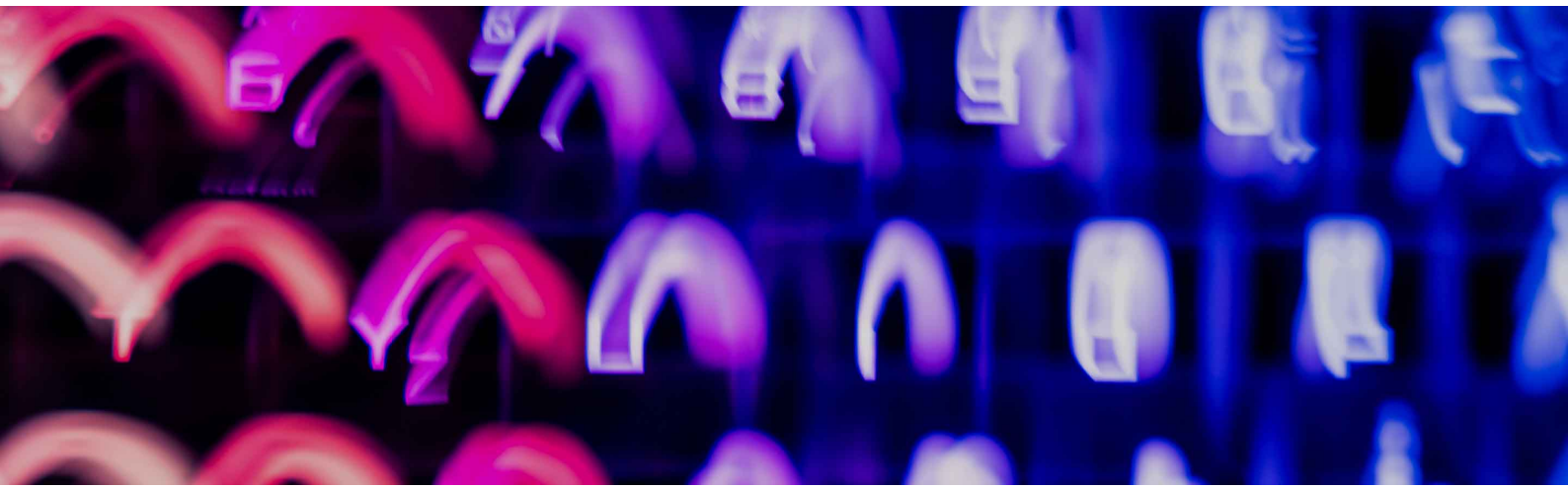


Breaking the GPU stronghold: emerging competition in AI infrastructure

Technology / Article
October 02, 2025

The accelerator market is actively seeking alternatives to Nvidia, fostering a diversified ecosystem where credible options—from established players to emerging start-ups—are rapidly gaining traction.

Nvidia has long defined the AI compute stack. Its strength comes not only from GPU performance but also from a tightly integrated ecosystem spanning chips, systems (DGX and HGX), software (CUDA and cuDNN), and cloud distribution (DGX Cloud and CoreWeave). But as AI models scale in both size and complexity, the costs and risks of relying on a single vendor are becoming untenable.

High margins, limited supply, and strategic lock-in are pushing customers and competitors to act. A new wave of challengers, including rival chipmakers, hyperscaler-designed silicon, and architecture-first start-ups, is gaining momentum. In fact, we project Nvidia's market share will fall from about 90 percent today to 70 percent by 2030. In this article, we shed light on the forces that are reshaping the AI infrastructure landscape, profile the emerging competition, and highlight where disruption is already taking hold.

1. The market forces fueling a shift away from Nvidia GPUs

Nvidia has become the default provider of AI compute, controlling nearly all of the AI accelerator market, but its dominance is starting to dwindle. Four market forces are accelerating this shift:

High margins are attracting competition and pushing buyers to seek alternatives. Nvidia's data-center GPUs deliver gross margins of about 75 percent, well above industry norms. While this has sustained exceptional profitability for Nvidia, it has also increased the cost of AI infrastructure, a burden that even the largest hyperscalers are finding difficult to justify at scale. High margins are also attracting competition. With the market size projected to exceed $400 billion, established chipmakers, hyperscalers, and start-ups are introducing cost-competitive, "good enough" alternatives. These solutions that may not match Nvidia's full-stack performance but offer sufficient capability at lower price points for a broadening range of workloads.

Server OEMs are being squeezed out of the value chain. Server OEMs such as Supermicro, Dell, and HPE have seen their margins collapse from 10 to 12 percent to only 3 to 4 percent on Nvidia-based systems. Nvidia's reference designs, including HGX and DGX, tightly constrain what OEMs can build, leaving little room for differentiation. Most of the system value is captured upstream by Nvidia, while OEMs are relegated to low-margin assembly. This shift is incentivizing OEMs to explore partnerships with alternative chip vendors where they have more room to innovate, integrate, and rebuild profitability.

Hyperscalers are using custom silicon to reduce their dependence. As AI infrastructure spending accelerates, hyperscalers are under pressure from investors to improve cost efficiency and demonstrate a strong ROI on capex. Nvidia's premium pricing makes it more difficult to profitably scale AI cloud infrastructure. Beyond cost, dependence on a single vendor for foundational compute has become a strategic concern. Risks of vendor lock-in, constrained supply, and limited design flexibility have prompted a decisive response: hyperscalers are investing in custom accelerators tailored to their own training and inference workloads. These include Google's TPU, Amazon's Trainium, Microsoft's Maia, and Meta's MTIA, each optimized for internal scale, workload fit, and long-term cost control.

Nvidia is expanding its reach downstream and upstream. Nvidia's reach now extends beyond chips and systems into cloud services and distribution. Through CoreWeave, a GPU cloud provider with only about 2 percent of global revenue but more than 5 percent of Nvidia's supply, Nvidia has created an alternative channel outside the major hyperscalers. It has also launched DGX Cloud, a managed service operated on hyperscaler infrastructure but controlled by Nvidia, and Lepton, a marketplace aggregating GPU capacity from smaller providers such as Lambda and Crusoe. By moving into services and distribution, Nvidia risks competing with its largest customers, incentivizing hyperscalers and emerging cloud players like Cloudflare to accelerate alternatives and reduce reliance on Nvidia.

These moves highlight a broader strategy: Nvidia is no longer just a component supplier. It is shaping the architecture, economics, and control of AI compute as demand accelerates and cloud dynamics evolve.

2. A new wave of competition

A growing set of challengers is targeting specific weaknesses in Nvidia's model: rising costs, supply constraints, and workload inefficiencies. Rather than replicate Nvidia's full stack, these players focus on targeted use cases where they can compete effectively. They fall broadly into three categories:

Competing GPU providers

AMD has emerged as the most serious contender. Its MI350X and soon-to-be-launched MI450X products are competitive on key hardware specs such as high-bandwidth HBM3E memory, FP4/FP8 support, and efficient chiplet packaging. AMD's biggest commercial win to date is a large-scale partnership with Oracle Cloud Infrastructure (OCI), which plans to deploy tens of thousands of MI300X GPUs to support enterprise AI workloads. While AMD still trails Nvidia in developer ecosystem and software stack maturity, its ROCm platform is improving steadily and gaining adoption.

Intel, in contrast, has faced headwinds with Gaudi 3. The accelerator was delayed by about a year, with 2024 shipment projections slashed by about 30 percent amid persistent software and production challenges. Still, Gaudi 3 has now reached limited deployment, and IBM Cloud announced itself as the first public provider to offer Gaudi 3 instances, aiming to deliver cost-effective generative AI acceleration to customers. Even so, Gaudi 3 is a niche alternative, with Intel refocusing its strategy on future architectures and targeted enterprise offerings.

Neither player matches Nvidia's full-stack integration, but AMD is beginning to close the hardware gap, while Intel remains in rebuilding mode.

Hyperscaler custom silicon

All four major hyperscalers have developed their own AI accelerators, designed specifically for internal training and inference workloads:

Google's Tensor Processing Unit (TPU v7) is a custom accelerator optimized for TensorFlow and JAX, supporting both training of large models such as Gemini and inference across core products including Search, Gmail, and YouTube. Initially reserved for internal use, TPUs are now available through Google Cloud's Vertex AI and GKE, enabling adoption by AI labs and enterprises. In 2025, OpenAI began renting TPUs to scale ChatGPT inference more cost-effectively, joining early adopters such as Anthropic and Apple. Enterprises including Ford, Banco BV, and GSK are also leveraging TPU-backed infrastructure for workloads ranging from medical imaging to fraud detection and network optimization.

AWS has developed two custom accelerators, Trainium2 for training and Inferentia2 for inference, designed to lower costs and improve performance versus GPU-based instances. They power internal services such as Alexa and CodeWhisperer and are available through EC2 and SageMaker. Customers including Anthropic, Datadog, Scale.ai, and Money Forward report up to 80 percent lower inference costs and 50 percent training savings. Beyond AI labs, enterprises in healthcare, finance, and compliance use AWS silicon for workloads such as medical summarization, fraud detection, and document classification.

Microsoft has developed Maia, a custom accelerator for training and inference of large models within Azure. The first chip, Maia 100, launched in late 2023, supports workloads such as GPT-4, Bing AI, and Microsoft Copilot. Maia remains early-stage, tightly integrated with Azure, and not available to external

customers. Adoption is limited today, but its rollout underscores Microsoft's long-term strategy to reduce reliance on GPU.

Meta has developed Meta Training and Inference Accelerator (MTIA), an in-house chip designed for inference workloads across Facebook, Instagram, and WhatsApp. The first generation, launched in 2023, supported simple AI tasks; MTIA v2, introduced in 2024, expanded compute and memory bandwidth for more complex recommendation and ranking models. Unlike general-purpose GPUs used to train models such as Llama, MTIA is focused on efficient inference at scale within Meta's data centers. Adoption remains internal and early-stage, reflecting Meta's strategy of vertical optimization rather than broad commercial deployment.

These initiatives are long-term bets to reduce dependence on Nvidia, improve economics at scale, and strengthen control over future AI infrastructure. While GPUs remain essential for many workloads, hyperscalers are steadily migrating critical use cases to silicon they design and optimize in-house.

New-age architecture start-ups

A third wave of competition is emerging from start-ups that are rethinking AI compute from the ground up. Rather than competing head-on with Nvidia's GPUs, these players build around its blind spots, addressing specific bottlenecks such as latency, memory bandwidth, or power efficiency.

Groq has developed LPU v2, a deterministic, compiler-driven architecture optimized for ultra-low-latency inference. By removing caches and branch predictors, it eliminates runtime variability, making it well-suited for real-time applications such as conversational AI and algorithmic trading. Groq has demonstrated industry-leading token-level latency and is positioning itself as the reference platform for high-throughput, low-latency workloads. The company recently launched an inference data center in Helsinki with Equinix, secured a sovereign AI infrastructure deal in Canada through AIFabric, and received a $1.5 billion commitment from Saudi Arabia to expand GroqCloud in the Middle East.

Cerebras with WSE-3 takes a radically different approach with its wafer-scale engine, a single chip the size of a dinner plate with 900,000 cores. This architecture enables trillion-parameter training on a single system, reducing the need for complex cluster orchestration. Cerebras systems are deployed in national labs and pharma R&D, where scale and parallelization efficiency are critical. UAE-based G42 accounts for more than $100 million in contracts to deploy Condor Galaxy AI supercomputers, while the US Department of Defense awarded a $45 million contract for real-time battlefield simulations. Cerebras has also partnered with Meta to accelerate Llama inference (up to 18x GPU speed) and with Perplexity AI to power one of the fastest AI search models available.

SambaNova with SNL40 delivers rack-scale systems based on a dataflow architecture, optimized for fine-tuned LLMs and enterprise deployment. Its integrated software platform, SambaFlow, enables rapid deployment without large ML ops teams. SambaNova has built traction with government agencies and Fortune 100 enterprises. Key deployments include Accenture, Argonne National Laboratory, and a strategic partnership with SoftBank Corp. to power inference workloads across APAC using models such as Llama and Qwen.

Beyond these leaders, start-ups including **Tenstorrent** (training and inference across data center and edge), **d-Matrix** (transformer-optimized inference acceleration), and **Untether AI** (ultra-efficient inference) are introducing differentiated architectures that rethink data movement, memory access, and scalability.

While still niche compared with mainstream GPU deployments, these companies provide credible alternatives for buyers seeking performance advantages in specific workloads or greater control over their infrastructure.

3. How do the challengers stack up

Nvidia is still the default in AI compute, but challengers are advancing across multiple layers of the stack. Each brings a focused strength, some in hardware, others in developer ecosystems, system integration, or economics. Together, they are fragmenting what was once a single-vendor stronghold into a more diverse and workload-optimized marketplace. Figure 1 summarizes the key considerations we use to compare the alternatives.

Figure 1
AI accelerators can be compared on a variety of factors

Considerations for evaluating AI accelerators	Description
Hardware architecture	Examines core chip design choices such as memory, compute formats, and bandwidth that shape raw capability
Developer stack and ecosystem	Explores the maturity of software platforms, libraries, and integrations that enable model development and deployment
System integration and infrastructure design	Looks at how accelerators are assembled into systems and racks, and the flexibility of deployment model
Performance and efficiency	Compares throughput, latency, and energy use across different workload types
Cost, deployment, and total cost of ownership	Evaluates pricing, operating economics, and lifetime cost of scaling infrastructure
Strategic and business considerations	Considers market dynamics, ecosystem strategies, and diversification efforts that influence long-term adoption

Source: Kearney analysis

Hardware architecture: AMD leads among GPU alternatives

Among direct GPU competitors, AMD has emerged as the most credible challenger. The Instinct MI300X offers 192 GB of HBM3E memory, industry-leading bandwidth, and FP8/FP4 support, making it a strong option for large-model training and high-throughput inference. These capabilities enable dense rack-level

deployments valued by cloud providers and enterprises. Microsoft validated this positioning by launching MI300X-backed Azure ND v5 instances in early 2025, with Hugging Face as an early adopter. Meta has also deployed MI300X at scale for Llama training and inference, signaling its intent to reduce Nvidia reliance. Apple continues to emphasize proprietary silicon for on-device AI, while Google relies on TPUs for cloud training. Nvidia still leads in absolute compute and integration, but AMD now provides a viable alternative for organizations seeking cost-effective capacity and vendor diversification.

Developer stack and ecosystem: ROCm progressing but structural gaps persist

Nvidia's software moat remains formidable, but AMD's ROCm platform has begun to narrow the gap. With ROCm 6.0, released in 2025, the platform now supports PyTorch, Hugging Face, and Triton, and runs open-source models such as Llama 2, Mistral, and Llama 3.1 natively. This progress makes ROCm increasingly viable for developers prioritizing cost savings or vendor neutrality.

However, there are still structural challenges for wider adoption of ROCm:

- **Ecosystem fragmentation.** Many pretrained models and enterprise tools still default to CUDA, forcing developers to undertake manual and error-prone conversion processes.
- **Technical gaps.** ROCm continues to struggle with dynamic shape execution, multi-GPU orchestration, and fine-grained kernel optimization, leading to weaker evaluation scores and slower continuous integration.
- **Tooling immaturity.** Developer profiling and debugging tools remain less advanced than Nvidia's Nsight suite, extending development cycles.
- **Enterprise integration.** AMD lacks the deep reference integrations that Nvidia maintains with enterprise vendors like ServiceNow and SAP, complicating adoption for risk-averse IT buyers.

These gaps do not negate ROCm's progress—hyperscalers such as Microsoft and Oracle are already deploying ROCm at scale—but they represent switching costs that enterprise buyers must still weigh carefully.

System integration and infrastructure design: AMD extends reach, Cerebras reinvents the model

System-level integration has become a new arena of competition. Nvidia's DGX and HGX platforms lock OEMs into tightly prescribed reference designs, capturing much of the system value chain. AMD, in contrast, offers more flexibility by partnering with OEMs such as Dell and Supermicro, enabling enterprises to tailor rack-scale deployments to their specific workload and economic needs. AMD's position was strengthened by its $4.9 billion acquisition of ZT Systems in 2025, a top hyperscaler ODM. AMD retained ZT's design and customer teams while selling the manufacturing arm to Sanmina for $3 billion, creating a preferred partnership model. This structure enables AMD to offer vertically integrated yet modular systems that combine Instinct GPUs, EPYC CPUs, and Pensando DPUs. In effect, AMD can now compete head-to-head with Nvidia's vertically integrated systems while still appealing to enterprises that value openness and configurability.

Cerebras has taken a different approach by reimagining the fundamental unit of system design. Its wafer-scale engine (WSE-3) consolidates compute into a single massive chip, eliminating the need for multi-GPU clustering. For specialized workloads such as training trillion-parameter models in national labs or pharma R&D, this approach dramatically simplifies orchestration and offers unique performance advantages. While limited in scope, Cerebras illustrates how challengers can compete not by replicating Nvidia's playbook but by breaking from it entirely.

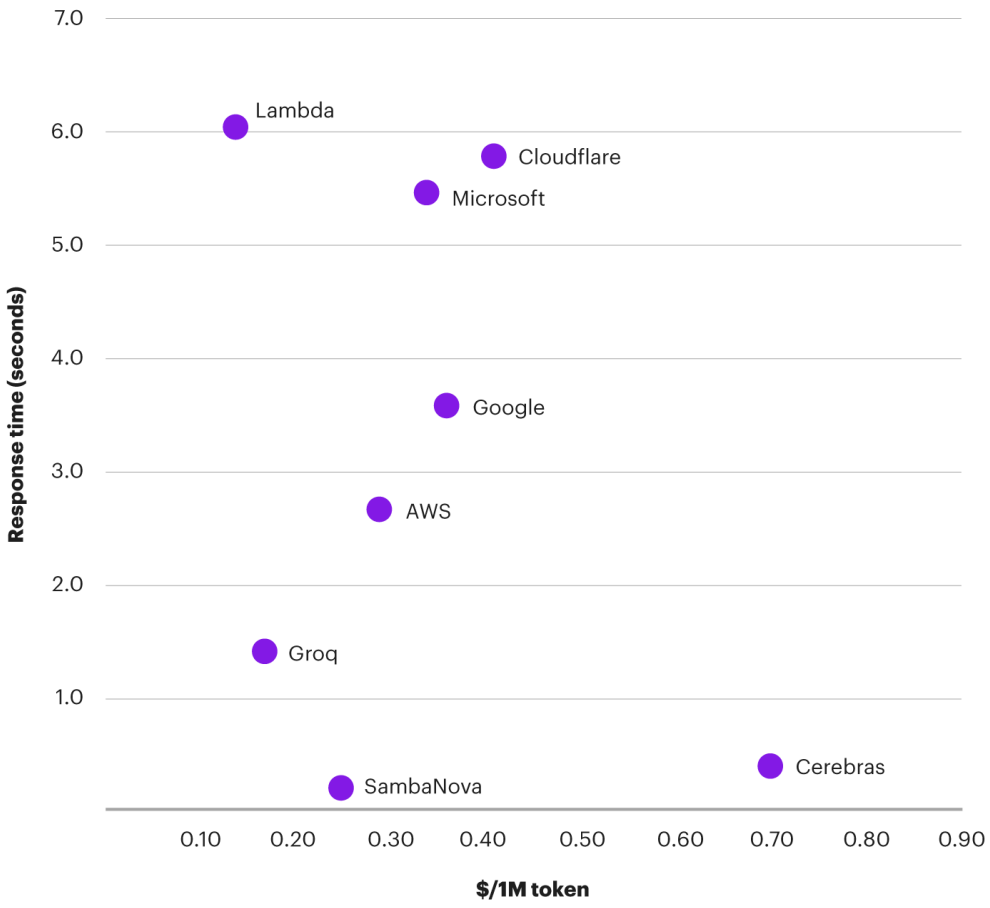
Performance and efficiency: workload-specific leaders emerge

Nvidia continues to lead on broad, general-purpose benchmarks, but challengers are establishing clear advantages in targeted workloads. AMD's MI300X delivers competitive throughput and energy efficiency at lower cost, strengthening its case for large-scale deployments. Groq's compiler-driven architecture provides unmatched real-time inference latency, making it well-suited for conversational AI and financial applications. Cerebras' wafer-scale design enables trillion-parameter training on a single system, eliminating the complexity of multi-GPU orchestration. Meanwhile, hyperscaler silicon such as Google's TPU v5p and AWS Trainium demonstrates strong cost-normalized efficiency within their respective ecosystems.

Groq, SambaNova, and Cerebras systems have lower response times than Nvidia systems deployed by hyperscalers and challengers such as Coreweave and Lambda (see figure 2).

Figure 2
Groq and SambaNova offer cheaper options for specific models with slower response times

Cost vs. response time of emerging AI accelerators



Sources: Artificial Analysis LLM API Providers Leaderboard; Kearney analysis

Cost, deployment, and TCO: AWS and AMD challenge Nvidia's premium

Cost has become a decisive factor as AI infrastructure scales. Nvidia continues to command premium pricing, but buyers are increasingly evaluating alternatives on total cost of ownership. AWS Trainium has delivered up to 50 percent savings in training and 80 percent in inference versus GPUs. AMD's MI300X is priced 20 to 30 percent lower than Nvidia equivalents, and its broad OEM ecosystem enables rack-level designs that further reduce TCO.

Recent MLPerf and WWT benchmarks confirm AMD's advantage on cost-normalized inference throughput, especially in dense deployments. These results are prompting enterprises and hyperscalers to question whether Nvidia's premium remains justified as model sizes and inference demand surge.

Strategic and business considerations: hyperscaler silicon reshapes market power

Beyond hardware economics, hyperscalers are pursuing custom silicon as a hedge against Nvidia's market power. Google has scaled TPU adoption across both internal services and external customers, with OpenAI, Anthropic, and Apple all joining as major TPU users in 2025. AWS has expanded the use of Trainium and Inferentia internally (Alexa and CodeWhisperer) and externally with clients such as Scale.ai, achieving significant cost savings. Microsoft's Maia and Meta's MTIA remain primarily internal, powering Copilot and recommendation engines, but represent strategic bets on architectural independence.

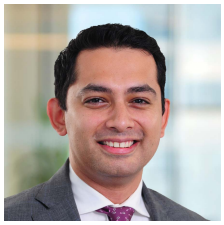
In addition, Meta and Oracle have publicly embraced AMD for training capacity, while innovators such as Groq and SambaNova are being tapped for specialized workloads like conversational AI or turnkey deployments. The message is clear: organizations are no longer comfortable concentrating their AI futures in the hands of a single vendor. Custom silicon and alternative platforms provide leverage in pricing, supply, and long-term architectural direction.

4. A more varied mix on the horizon

The AI accelerator market is evolving from a single-vendor stronghold into a pluralistic, workload-optimized ecosystem, reshaping economics, competition, and enterprise choice. Nvidia will still be the dominant player, but by 2030, its market share is expected to fall below 70 percent as challengers carve out a larger share. AMD is likely to secure 5 to 10 percent of the market by scaling its MI350/450 series across hyperscalers and enterprise customers. Hyperscaler-designed silicon including Google's TPU, AWS' s Trainium, and Microsoft's Maia could capture 15 to 20 percent as internal deployments expand and external offerings mature. Niche innovators such as Groq and Cerebras will chip away at the edges, gaining up to 5 percent in specialized inference and large-scale training workloads. Despite these shifts, Nvidia will continue to dominate, sustained by its broad ecosystem, software leadership, unmatched end-to-end integration, and its ability to outspend competitors in R&D on chip development.

The authors wish to thank Aanuj Taneja and Shiven Mahajan for their valuable contributions to this article.

Authors



Sridhar Narasimhan
Partner

Archit Johar
Principal

Bhavtosh Jadeja
Consultant

Abhilash Jain
Consultant

About Kearney

For 100 years, Kearney has been a leading management consulting firm and trusted partner to three-quarters of the Fortune Global 500 and governments around the world. With a presence across more than 40 countries, our people make us who we are. We work impact first, tackling your toughest challenges with original thinking and a commitment to making change happen together. By your side, we deliver—value, results, impact.