How does AMD's displacement of Nvidia in the AI chip market reconfigure the power structures of global semiconductor supply chains to affect the governance, cost efficiency, and technological sovereignty of large language model deployment?

May 29, 2026 | SnugLab Research | readme.snuglab.com

Executive Summary

AMD's entry into the AI chip market introduces competitive pressure that lowers deployment costs and reduces vendor lock-in, but it does not yet fundamentally reconfigure global semiconductor supply chain power structures to definitively shift the balance of cost efficiency and governance stability away from Nvidia's integrated ecosystem. While AMD offers compelling price-performance for LLM inference and an open-source software stack, Nvidia retains a dominant market share and its mature CUDA ecosystem continues to offer superior efficiency for training workloads and short-term cloud rentals, suggesting a dual-lead dynamic rather than a displacement.

Key Findings

AMD Creates a Dual-Lead Dynamic, Reinforcing Geographic Concentration

AMD's market share growth has not displaced Nvidia's structural dominance, as Nvidia retains approximately 80-92% of the AI accelerator market [9, 10]. Instead, AMD's expansion creates a dual-lead dynamic where both companies compete for the same constrained manufacturing resources, reinforcing geographic concentration in East Asia [4, 6, 10]. Both companies rely on TSMC's advanced nodes, with AI chip demand consuming approximately 60% of TSMC's N3 wafer output in 2026 and projected to reach 86% in 2027 [6]. HBM production further strains this capacity, consuming three times the wafer space per bit of commodity DRAM and nearly four times with HBM4 [6]. This competition reinforces geographic concentration, as 100% of advanced (below 10 nanometers) manufacturing capacity is located in Taiwan (92%) and South Korea (8%) [4]. The specific manufacturing allocation mechanism involves prioritizing limited N3 and 2nm wafer slots based on demand, with AMD's upcoming MI400 series expected to be

the first GPU produced on TSMC's 2nm process [10].

Nvidia and AMD have secured dual-sourced HBM contracts with SK Hynix and Samsung, backed by TSMC capacity reservations [1, 2, 3, 13, 15, 18]. Nvidia holds approximately 60% of total CoWoS allocation and over 70% of the CoWoS-L capacity required for dual-die Blackwell/Rubin chips [3, 12]. Nvidia intentionally diversified its HBM4 sources, inviting Samsung to negotiate 2026 HBM4 pricing just one week after securing supply with SK Hynix, explicitly seeking diversified sources due to surging HBM4 demand [19]. AMD's commitment of its MI400 to Samsung HBM4, following previous HBM3E supplies, also demonstrates a strategic dual-sourcing approach [13, 15].

Cost Efficiency for LLM Deployment Varies by Workload and Scale

AMD's hardware price advantage is substantial but does not automatically guarantee lower total cost of ownership (TCO) for LLM deployment, as software optimization overheads and operational expenses can compound to negate initial savings [5, 10, 16]. AMD's MI300X accelerators are priced at roughly half the cost of Nvidia's H100 and offer 2.4x the high-bandwidth memory (HBM) capacity, making them highly attractive for memory-bound inference workloads [9, 10, 15]. Cloud pricing for MI300X generally runs 40-60% below H100 rates at comparable providers [10]. For inference, AMD's MI300X, MI325X, and MI350X models offer competitive or superior cost per token at most batch sizes, particularly for memory-bound LLM serving [10, 12, 15]. The MI300X shows a 40% latency advantage over the H100 on Llama2 70B workloads [4, 5, 8, 12]. As model parameter counts exceed 100B, AMD's memory capacity and bandwidth advantages become more pronounced; for example, the H100 cannot fit DeepSeekV3 670B into a single node, while the MI300X can run it on 4-5 GPUs [12, 20].

However, for training workloads, Nvidia generally delivers better performance-per-TCO due to its mature software ecosystem and higher effective utilization, unless organizations invest in significant AMD-specific kernel optimization [10]. For short-term GPU rentals (under six months), Nvidia also generally offers better performance per dollar due to a more competitive rental market [12]. On-premise deployment involves significant operational expenses for electricity, cooling, maintenance, and staffing, with staffing costs alone ranging from $225,000 to $300,000 over three years [5, 16]. Cloud solutions remain more cost-effective when GPU utilization rates fall below 70% [16].

Technological Sovereignty Enhanced by Open Ecosystems, Limited by

Shared Supply Chains

Technological sovereignty entails controlling the entire AI value chain, from hardware to software and data, to ensure strategic autonomy and prevent dependency on foreign or single-vendor ecosystems [7, 11, 14]. AMD's open-source ROCm software ecosystem reduces the vendor lock-in associated with Nvidia's proprietary CUDA, allowing organizations to build "Sovereign AI Solutions" that enhance strategic autonomy and governance over their AI infrastructure [11, 17]. This openness aligns with sovereignty goals by allowing nations to scale domestic AI capabilities without being tethered to a single corporate ecosystem [11, 17]. AMD has collaborated on initiatives like the U.S. Department of Energy's El Capitan supercomputer and Finland's "Compute to Impact" program to foster domestic AI capabilities [11].

However, diversifying procurement toward AMD does not fully de-risk geographic supply chain vulnerabilities. Both AMD and Nvidia rely on the same constrained TSMC wafers and HBM, meaning the global manufacturing concentration in East Asia remains a shared vulnerability [4, 6, 10]. While AMD provides a credible second source, maintaining dual hardware and software ecosystems fragments procurement standards and increases operational complexity, demanding additional engineering labor to optimize kernels and prevent performance gaps [10, 17].

Governance Power Shifts Towards Software Frameworks and Hyperscalers

The concurrent competition between Nvidia, AMD, and custom silicon forces LLM developers to adopt modular, hardware-agnostic architectures to avoid vendor lock-in and optimize deployment costs [9, 10, 17]. This architectural shift transfers governance power away from chipmakers and toward software frameworks, hyperscalers, and state initiatives [7, 17]. Standardizing on open frameworks like ROCm allows developers to maintain control and flexibility, shifting the governance of LLM development from hardware manufacturers to the maintainers of these software stacks [17].

Hyperscalers are also centralizing governance through their custom silicon, with Google running over 75% of Gemini on TPUs and AWS Trainium processing over 50% of Bedrock tokens [10]. States are establishing regulatory standards, such as the EU's Cloud and AI Development Act (CAIDA), to define "sovereign cloud" and control the AI value chain [6]. The transition from CUDA to ROCm generally decentralizes LLM governance by distributing standard-setting authority across a broader coalition of

hardware vendors and independent contributors, despite the significant influence of hyperscalers.

Implications

AMD's growing presence in the AI chip market implies a more competitive landscape for LLM deployment, offering organizations greater choice and potentially lower costs, particularly for inference workloads. This competition encourages the adoption of open-source software ecosystems, which can enhance technological sovereignty by reducing reliance on single-vendor proprietary solutions. However, the continued geographic concentration of advanced semiconductor manufacturing means that supply chain risks remain largely undiversified at the foundational hardware level. Organizations must carefully weigh the initial hardware cost savings from AMD against potential software optimization overheads and the established maturity of Nvidia's ecosystem, especially for training-intensive applications or short-term cloud rentals. The shift towards modular, hardware-agnostic architectures will empower software developers and hyperscalers, but true strategic independence will require sustained investment in domestic AI infrastructure and talent to manage diverse hardware and software stacks.

Limitations and Caveats

This report draws from a source pool where a significant portion consists of blog posts, social media analyses, and press releases, with fewer peer-reviewed or government reports. Consequently, conclusions, particularly those involving specific market share projections or detailed cost-per-token benchmarks, should be treated as provisional and subject to ongoing market dynamics. Direct quantitative figures for AMD's exact B2B pricing tiers compared to Nvidia's H100 and H200 in Q2 2026 are limited, with available data providing relative pricing rather than specific dollar amounts. The long-term impact of AMD's market share growth on global semiconductor supply chain power structures is still unfolding, and the full extent of its reconfiguring effects on governance, cost efficiency, and technological sovereignty will require further observation.

Sources

[1] [gov] Crs Product - congress.gov - https://www.congress.gov/crs-product/R48642
[2] [edu] Semiconductor Supply Chains Ai And Economic Statecraft - cetas.turing.ac.uk -

https://cetas.turing.ac.uk/publications/semiconductor-supply-chains-ai-and-economic-statecraft

[3] [edu] U S Ai Statecraft - cset.georgetown.edu - https://cset.georgetown.edu/publication/u-s-ai-statecraft/

[4] Strengthening The Global Semiconductor Supply Chain In An Un - semiconductors.org - https://www.semiconductors.org/strengthening-the-global-semiconductor-supply-chain-in-an-uncertain-era/

[5] [preprint] Html - arxiv.org - AUTHORS UNAVAILABLE - https://arxiv.org/html/2509.18101v3

[6] Build Ai Dont Block Access The European Unions Digital Sover - laweconcenter.org - https://laweconcenter.org/resources/build-ai-dont-block-access-the-european-unions-digital-sovereignty-trap/

[7] [peer-reviewed] Digital Disintegration: Techno-Blocs and Strategic Sovereignty in the AI Era - Authors: Stephen Weymouth - Journal: International Organization - https://www.cambridge.org/core/journals/international-organization/article/digital-disintegration-technoblocs-and-strategic-sovereignty-in-the-ai-era/DD86C6FD3FDD7FBBADEF100C6935D577

[8] Special Focus Semiconductor Value Chains Dc772986 - oecd.org - https://www.oecd.org/en/publications/2025/09/economic-security-in-a-changing-world_78f3b129/full-report/special-focus-semiconductor-value-chains_dc772986.html

[9] [social] Amd Vs Nvidia Comprehensive Ai Chip Market Analysis Report A - linkedin.com - https://www.linkedin.com/pulse/amd-vs-nvidia-comprehensive-ai-chip-market-analysis-report-aujla-behpc

[10] Amd Vs Nvidia Ai Gpu Market Share 2026 - siliconanalysts.com - https://siliconanalysts.com/analysis/amd-vs-nvidia-ai-gpu-market-share-2026

[11] Sovereign Ai - amd.com - https://www.amd.com/en/solutions/ai/sovereign-ai.html

[12] Amd Vs Nvidia Inference Benchmark Who Wins Performance Cost - newsletter.semianalysis.com - https://newsletter.semianalysis.com/p/amd-vs-nvidia-inference-benchmark-who-wins-performance-cost-per-million-tokens

[13] [blog] Llm Inference Benchmarking How Much Does Your Llm Inference - developer.nvidia.com - https://developer.nvidia.com/blog/llm-inference-benchmarking-how-much-does-your-llm-inference-cost/

[14] [preprint] Delivery.Cfm - papers.ssrn.com - AUTHORS UNAVAILABLE - https://papers.ssrn.com/sol3/Delivery.cfm/5312977.pdf?abstractid=5312977&mirid=1

[15] Amd Vs Nvidia Ai Workloads Performance 2025 - sanj.dev - https://sanj.dev/post/amd-vs-nvidia-ai-workloads-performance-2025

[16] [blog] Llm Inference On Premise Vs Cloud - spheron.network - https://www.spheron.network/blog/llm-inference-on-premise-vs-cloud/

[17] [blog] Amd Vs Nvidia Gpu - fluence.network - https://www.fluence.network/blog/amd-vs-nvidia-gpu/

[18] 261577980 Skhynix Hbm4 Samsung Nvda Micron Mu Tradingkey - tradingkey.com - https://www.tradingkey.com/analysis/stocks/us-stocks/261577980-skhynix-hbm4-samsung-nvda-micron-mu-tradingkey

[19] Samsung Hbm4 Nvidia 2026 Shipments - digitimes.com - https://www.digitimes.com/news/a20251127PD231/samsung-hbm4-nvidia-2026-shipments.html

[20] [blog] Nvidia H100 Vs Amd Mi300x Vs Intel Gaudi3 Best Gpu For Ai Tr - hostrunway.com - https://www.hostrunway.com/blog/nvidia-h100-vs-amd-mi300x-vs-intel-gaudi3-best-gpu-for-ai-training-llm-inference/