# Why the Revolt Against Nvidia Made Broadcom the Real Toll Collector

Second Order Labs™ · June 2026

ABSTRACT

Every hyperscaler tried to escape Nvidia's pricing power by building custom chips. They didn't tear down the toll road, they split it in two and handed the bigger booth to Broadcom.

*Keywords:* Broadcom, Custom Silicon, Nvidia, AI Infrastructure, Hyperscalers, Interconnect

> *"In an inflationary world, a toll bridge would be a great thing to own because you've laid out the capital costs. You build it in old dollars, and you don't have to keep replacing it."*
>
> — Warren Buffett

For two years, every hyperscaler has been trying to get out from under Nvidia's pricing power. Amazon pushed Trainium. Google kept upgrading TPUs. Microsoft approved Athena. Meta sent part of a $35-40 billion capex plan into MTIA [1]. The pitch was straightforward: stop paying Nvidia's tax. What happened instead was stranger. The toll road stayed. It just got a second booth, and Broadcom now runs it.

A lot of tech coverage treats custom silicon like a clean fight between Nvidia and the hyperscalers, with Marvell cast as the plucky anti-Nvidia winner on the side road. The numbers do not support that story. Marvell's data center revenue rose 87% year over year to $816 million [2]. Broadcom says AI revenue should top $10 billion, more than 35% of its semiconductor business [3]. That is not a close contest. The supposed alternative to one monopoly is another company with monopoly-like economics.

> *The hyperscaler revolt didn't destroy Nvidia's toll road. It bifurcated it, and Broadcom collects the larger half.*

Call it the **Bifurcated Toll Road**. Escaping Nvidia's grip on compute does not remove the toll. It splits it. Broadcom charges on the compute off-ramp by designing the custom ASICs. Nvidia has quietly moved deeper into the network fabric, charging data centers even when its GPUs are absent. The savings from buy-

ing fewer Nvidia accelerators do not stay with the hyperscalers. They move to the two companies that control chip design and interconnect.

## I. How the Hyperscaler Revolt Built Broadcom's Toll Booth

The revolt is real. The motive is economic. Andy Jassy said AWS customers want "better price-performance for their AI workloads"[4]. That is polite corporate language for wanting out of Nvidia's margins. Performance per dollar matters more than bragging rights on peak FLOPS. Google's TPU v5e shows competitive performance per dollar against comparable GPUs in MLPerf[5]. A chip does not need to top the benchmark chart to win a budget meeting.

Still, hyperscalers are not doing this alone. They need SerDes IP, advanced packaging, and the networking blocks that turn a chip into part of a 32,000-accelerator cluster. Broadcom is the partner that keeps showing up. When Reuters reported that Broadcom had added Meta to a customer list that already included Google[6], the picture got hard to ignore. SemiAnalysis called Broadcom "the undisputed king of custom silicon"[7]. Meta's MTIA, which more than doubled the compute and memory bandwidth of the prior version[8], sits on Broadcom-supplied building blocks.

**Figure 1:** Broadcom doesn't build the hyperscalers' chips to compete with them, it taxes the off-ramp they're all racing toward.

That is the trap. Each hyperscaler talks about proprietary silicon as a way to weaken Nvidia's pricing power. But the same design partner keeps appearing underneath the story. Broadcom is the arms merchant for the rebellion. It supplies the IP that makes TPU, MTIA, and similar efforts manufacturable at scale. The hyperscalers diversified away from one GPU vendor and rebuilt dependence around one ASIC partner. Broadcom's own 10-K says "a significant portion of our revenue is derived from a small number of top-tier cloud providers"[10]. If you want a pure read on the custom-silicon buildout, Broadcom is sitting right in the middle of it.

## II. Interconnect Annexation: Nvidia's Socket-Agnostic Tax

The second toll booth gets far less attention. While everyone argued over compute sockets, Nvidia turned networking into a monster business. In Q1 FY25, net-

working revenue reached $3.2 billion, up 242% year over year[11]. One analyst put it plainly: Nvidia is now a networking company that also sells GPUs[12].

That is **interconnect annexation**. Nvidia is collecting money on the network fabric of AI clusters no matter whose accelerators fill the racks. The product is Spectrum-X, which Nvidia calls "the world's first Ethernet fabric built for AI," with generative AI network performance 1.7x higher[13]. The strategic intent is written right into Nvidia's own language: "Spectrum-X opens a brand-new market for us to bring large-scale AI to Ethernet-only data centers"[11]. Those Ethernet-heavy data centers are exactly where custom ASIC deployments show up. Nvidia built a product for customers that replaced its compute.

> *"In an inflationary world, a toll bridge would be a great thing to own because you've laid out the capital costs. You build it in old dollars, and you don't have to keep replacing it." Warren Buffett*

Buffett's toll bridge still fits. It just needs one update. AI infrastructure has two bridges, not one. Traffic pushed off Nvidia compute heads straight toward another paid crossing. A hyperscaler that swaps GPUs for Trainium still has to connect 32,000 accelerators into one working fabric. If it buys Spectrum-X, Nvidia still gets paid. The compute socket became socket-agnostic. The tax did not.

## III. The Two Toll Booths Collide at the Network Fabric

The center of the fight has moved from the chip to the wires between chips. At cluster scale, the network is the bottleneck. That makes the fabric the most valuable layer in the data center. It is also where Broadcom and Nvidia stop looking complementary and start looking like direct rivals.

Broadcom's answer is Jericho3-AI, a fabric that connects "up to 32,000 GPUs with perfect load balancing, zero-impact failover, and ultra-high radix"[14]. Nvidia's answer is Spectrum-X. Both are aimed at the same problem: keeping giant clusters fed with data so the accelerators do not sit idle. The contest is no longer about who sells the flashiest processor. It is about who owns the switching layer that every processor depends on.

| Toll booth | Owner | What it taxes | Key product | Scale signal |
|---|---|---|---|---|
| Compute off-ramp | Broadcom | Custom ASIC design and IP | TPU / MTIA silicon | ~$10B AI revenue, >35% of semis[3] |
| Network fabric | Both | Cluster interconnect | Jericho3-AI vs Spectrum-X | Nvidia networking $3.2B, +242% YoY[11] |

The architecture is pushing value into the fabric whether buyers like it or not. Google's TPU v4 uses optical circuit switches to reconfigure its interconnect topology on the fly[15]. Google's AI Hypercomputer pitch bundles "TPUs and GPUs with our AI-optimized network"[16] into one system. Jensen Huang said it clearly at GTC: "The data center is the new unit of computing"[17]. Once the unit of compute is the whole building, the company that controls the connective tissue gets paid on everything running inside it.

## IV. Why Marvell Is a Vassal Competitor, Not a Rival

Marvell keeps getting framed as the independent challenger that will break up the duopoly. That framing misses the scale and the structure. Start with the scale. $816 million versus a $10 billion AI run rate is not a head-to-head battle. It is barely the same weight class. Then look at the structure. Marvell's custom silicon and electro-optics business is growing fast, 87% is real growth[2], but it is growing inside an interconnect layer defined by Broadcom and Nvidia, not outside it. A third logo does not equal meaningful competition. Power sits with whoever sets the standard the third logo has to follow. In March 2026 the structure turned literal: Nvidia put $2 billion into Marvell and folded its custom XPUs into NVLink Fusion, so Marvell's silicon now plugs into Nvidia's own scale-up fabric rather than competing against it[20]. The would-be rival became a tenant inside the incumbent's network. That is the difference between a vassal competitor and a threat: Nvidia paid to keep Marvell racing on Nvidia's road.

**Figure 2:** Hyperscaler capex splits into two streams, but both must cross the same interconnect bridge the duopoly controls.

**Figure 3:** Absolute numbers, in millions. The media's "anti-Nvidia hero" earns a fraction of the two firms actually collecting the tolls. Sources: Broadcom, Nvidia, Marvell earnings[3] [11] [2].

## V. Where Frontier Lab Defection Proves the Thesis

The clearest proof that the Bifurcated Toll Road will stick comes from the labs. Custom silicon used to look like an internal hyperscaler project, justified by captive workloads and little else. That changed when Anthropic said it would "use AWS Trainium and Inferentia chips to build, train, and deploy our future foundation models"[18]. A frontier lab picked custom ASICs for training, not just inference. That matters. AWS says Trainium2 delivers up to 4x faster training and 3x more memory than the first generation[19]. That kind of jump starts to close the gap with Nvidia's roadmap.

So hyperscaler capex is splitting for good. One stream still goes to Nvidia for frontier training at the edge of the curve. Another growing stream goes to Broadcom-backed ASICs for inference and internal jobs. Both streams still cross a network fabric controlled by the same two incumbents. Leaving the compute socket is not an escape. It is a lane change.

## VI. What Operators and Investors Should Watch Next

The hyperscaler revolt was never about wiping out Nvidia. Jassy and Zuckerberg want enough alternatives to improve their bargaining position, not a market with one less supplier. That objective fits perfectly with Broadcom and Nvidia both winning, because pricing pressure on compute says nothing about who owns the wires.

Three signals matter. Open standards come first. UALink and Ultra Ethernet are the only real threats to the interconnect tax because they turn the fabric into a commodity. The next signal is true in-house ASIC design. If a hyperscaler absorbs the full design flow and no longer needs Broadcom's IP, the compute booth starts to weaken. The third is concentration risk turning against Broadcom, whose dependence on a few giant cloud customers becomes a weakness the moment one leaves. Until one of those breaks, the revolt will keep financing the tolls it was supposed to

eliminate. The best place to look in the AI buildout is not the chip that gets all the headlines. It is the booth almost everyone ignores.

KEY FINDINGS

Broadcom expects AI revenue above $10 billion, over 35% of its semiconductor business and more than ten times Marvell's $816 million data center figure.

Nvidia's Q1 FY25 networking revenue reached $3.2 billion, up 242% year-over-year, faster than much of its compute story.

Spectrum-X taxes Ethernet-only data centers running custom ASICs, so Nvidia collects on the network fabric even when it loses the compute socket.

Anthropic committed to AWS Trainium and Inferentia to train its foundation models, which proves custom silicon now reaches frontier training, not just inference.

Only UALink and Ultra Ethernet genuinely threaten the interconnect tax that both Broadcom and Nvidia depend on.

REFERENCES


[1] Meta Q1 2024 Earnings Call, capex guidance of $35-40 billion driven by AI infrastructure. https://investor.fb.com/investor-events/event-details/2024/Q1-2024-Earnings/default.aspx https://s21.q4cdn.com/399680738/files/doc_financials/2024/q1/META-Q1-2024-Earnings-Call-Transcript.pdf

[2] Marvell Technology Q1 Fiscal 2025 Earnings Call, data center revenue $816M, up 87% YoY. https://investor.marvell.com/events/event-details/q1-fiscal-2025-marvell-technology-inc-earnings-conference-call https://investor.marvell.com/events/event-details/q1-fiscal-2025-marvell-technology-inc-earnings-conference-call

[3] Broadcom Inc. Q1 FY2024 Financial Results Call, AI revenue expected to exceed $10B, over 35% of semiconductor revenue. https://investors.broadcom.com/events/event-details/broadcom-inc-first-quarter-fiscal-year-2024-financial-results-conference-call https://investors.broadcom.com/events/event-details/q1-2024-broadcom-earnings-conference-call

[4] Amazon Q1 2024 Earnings Event, Andy Jassy on Trainium and Inferentia price-performance. https://ir.aboutamazon.com/events/event-details/2024/Q1-2024-Amazon-Earnings-Event/default.aspx https://ir.aboutamazon.com/events/event-details/2024/Q1-2024-Amazoncom-Inc-Earnings-Conference-Call-/default.aspx

[5] MLPerf Training v3.1 Results, TPU v5e performance-per-dollar against GPUs. https://mlcommons.org/en/training-normal-31/ https://mlcommons.org/2023/11/mlperf-training-v3-1-hpc-v3-0-results/

[6] Reuters, Broadcom wins custom chip business with Meta. https://www.reuters.com/technology/broadcom-wins-custom-chip-business-with-meta-exec-says-2024-03-07/ https://www.reuters.com/technology/broadcom-wins-custom-chip-business-with-meta-exec-says-2024-03-07/

[7] SemiAnalysis, Google TPU v5 Brings Broadcom to $3B+ AI Revenue. https://www.semianalysis.com/p/google-tpu-v5-brings-broadcom-to https://newsletter.semianalysis.com/p/broadcoms-google-tpu-revenue-explosion

[8] Meta Engineering, Introducing our next-generation MTIA. https://engineering.fb.com/2024/04/10/data-center-engineering/meta-training-inference-accelerator-mtia/ https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/

[9] The Information, Microsoft's New AI Chip Could Reduce Its Reliance on Nvidia. https://www.theinformation.com/articles/microsofts-new-ai-chip-could-reduce-its-reliance-on-nvidia https://www.theinformation.com/articles/microsoft-develops-ai-server-gear-to-lessen-reliance-on-nvidia

[10] Broadcom Inc. 2023 Form 10-K, revenue concentration among top-tier cloud providers. https://investors.broadcom.com/financial-information/sec-filings/sec-filings-details/default.aspx?FilingId=17145713 https://www.sec.gov/Archives/edgar/data/0001730168/000173016823000096/avgo-20231029.htm

[11] NVIDIA Q1 FY25 Financial Results, networking revenue $3.2B, up 242% YoY; Spectrum-X commentary. https://investor.nvidia.com/events-and-presentations/events-and-presentations/event-details/2024/NVIDIA-1st-Quarter-FY25-Financial-Results/default.aspx https://www.sec.gov/Archives/edgar/data/0001045810/000104581024000113/q1fy25pr.htm

[12] The Next Platform, Nvidia Is Now A Networking Company That Also Sells GPUs. https://www.nextplatform.com/2024/05/23/nvidia-is-now-a-networking-company-that-also-sells-gpus/

[13] NVIDIA Developer Blog, Supercharging AI Workloads with NVIDIA Spectrum-X. https://developer.nvidia.com/blog/supercharging-ai-workloads-with-nvidia-spectrum-x/

[14] Broadcom Blog, Jericho3-AI: The Industry's Highest Performance Fabric for AI Networks. https://www.broadcom.com/blog/jericho3-ai-the-industry-s-highest-performance-fabric-for-ai-networks

[15] TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning. https://arxiv.org/abs/2304.01433 https://arxiv.org/abs/2304.01433

[16] Alphabet Q1 2024 Earnings Call, AI Hypercomputer architecture. https://abc.xyz/investor/earnings/ https://abc.xyz/investor/events/event-details/2024/2024-q1-earnings-call/

[17] NVIDIA GTC 2024 Keynote, Jensen Huang, "The data center is the new unit of computing." https://www.nvidia.com/en-us/gtc/keynote/ https://www.nvidia.com/en-us/on-demand/session/gtc24-s62542/

[18] Anthropic, Anthropic and AWS collaborate to advance generative AI. https://www.anthropic.com/news/anthropic-amazon-trainium https://www.anthropic.com/news/anthropic-amazon-compute

[19] AWS, Trainium2 chips designed for high performance deep learning training. https://aws.amazon.com/blogs/aws/aws-trainium2-chips-designed-for-high-performance-deep-learning-training/ https://aws.amazon.com/blogs/aws/amazon-ec2-trn2-instances-and-trn2-ultraservers-for-aiml-training-and-inference-is-now-available/

[20] NVIDIA Newsroom, NVIDIA AI Ecosystem Expands as Marvell Joins Forces Through NVLink Fusion; $2 billion investment announced March 31, 2026. https://nvidianews.nvidia.com/news/nvidia-ai-ecosystem-expands-as-marvell-joins-forces-through-nvlink-fusion https://nvidianews.nvidia.com/news/nvidia-ai-ecosystem-expands-as-marvell-joins-forces-through-nvlink-fusion

---