

THE ACTIVANT *Opinion*

PyTorch Runs Everywhere. Who Gets Paid?

From new PyTorch tools to a new type of cloud

November 2026

AUTHORS:



Jono Vickery

Vice President, Research

About Activant

Activant is a research-led global investment firm that partners with high-growth companies. Since 2015, we have invested in category-defining businesses during their most critical phases of growth, partnering with founders who have won their initial battles and are ready for the next challenge.

Our approach pairs deep, proprietary research with patient, flexible capital and hands-on operational partnership. We work alongside founders and leadership teams to refine strategy, strengthen operations, and accelerate sustainable growth.

Activant Research is dedicated to uncovering the most exciting emerging technologies, sectors, and companies we believe will shape the future. Our research-driven perspective informs everything we do, helping us invest at meaningful inflection points and support founders in building enduring, category-leading businesses.

You can find out more about Activant and our research at https://activantcapital.com/.

TorchNeuron and TorchTPU: The CUDA moat is eroding

Nvidia, at over $5 trillion, is one of the two largest companies on Earth by market cap. It has spent much of the last year trading places with the current largest, Apple.[1] Last quarter it booked $96.2 billion of revenue at a 75% gross margin, a fine business.[2] Like Apple, it sells hardware, a category that competition has driven to commodity pricing time and again. However, both companies have avoided that fate in the same way: each maintains leadership through an improvement cycle, and each ties its hardware to a software ecosystem that makes the product better to use than any alternative.

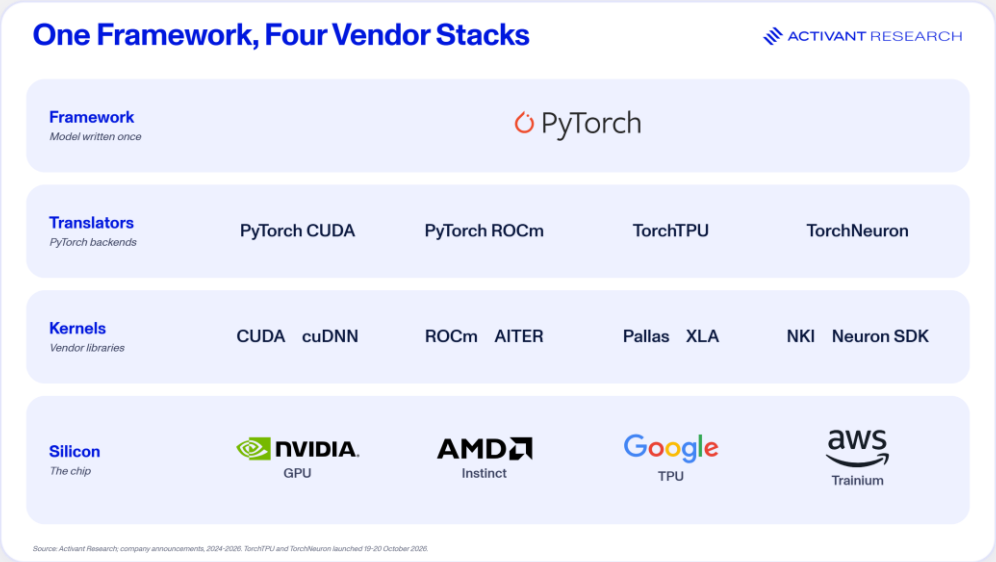
For Nvidia that software is CUDA, the programming layer that turns a model's math into instructions its GPUs can execute. Almost every model is written in PyTorch, and PyTorch grew up on CUDA, which gives Nvidia two moats. The first is compatibility: a new model runs on Nvidia the day it is released. The second is optimization: a decade of hand-written CUDA kernels, the small programs that do a model's heavy arithmetic, means most models also run fastest on Nvidia. In January we argued that inference providers optimizing on Nvidia would outperform the hardware independents for exactly this reason. In Inference Is Accelerating in July, we walked through the maturing inference software ecosystem, including automated kernel generation, and how that could erode the CUDA moat over the next 24 months, migrating profits out of silicon and into software.

That made last month's PyTorch Conference all the more interesting. AWS released TorchNeuron and Google open-sourced TorchTPU, and each registers its chip as a native PyTorch device.[3] A developer who changes the device name in their code from `cuda` to `neuron` or `tpu` can run the same model on Trainium or a TPU. For anything built in PyTorch, which is most models in use today, Nvidia's compatibility moat is gone.

We believe this confirms the direction of our July call and corrects it in a critical way. The CUDA moat is eroding as we said it would. But these tools are free, and they exist to sell Trainium and TPUs. That doesn't sound like profits migrating from silicon to software. The history of neutral software points the same way: of the 17 acquisitions of hardware-neutral compiler and orchestration companies we reviewed, 15 went to a chip vendor or a chip-adjacent buyer.[4] **Hardware orchestration and optimization software earns its margin when it is tied to silicon, not when it is sold on its own.** That means profit might be leaving silicon, but it is not going to software; it is going to whoever ties the two together to sell tokens. In the short to medium term that is the clouds. In the long run it is a neutral token seller that buys from every chip vendor, which no cloud can do.

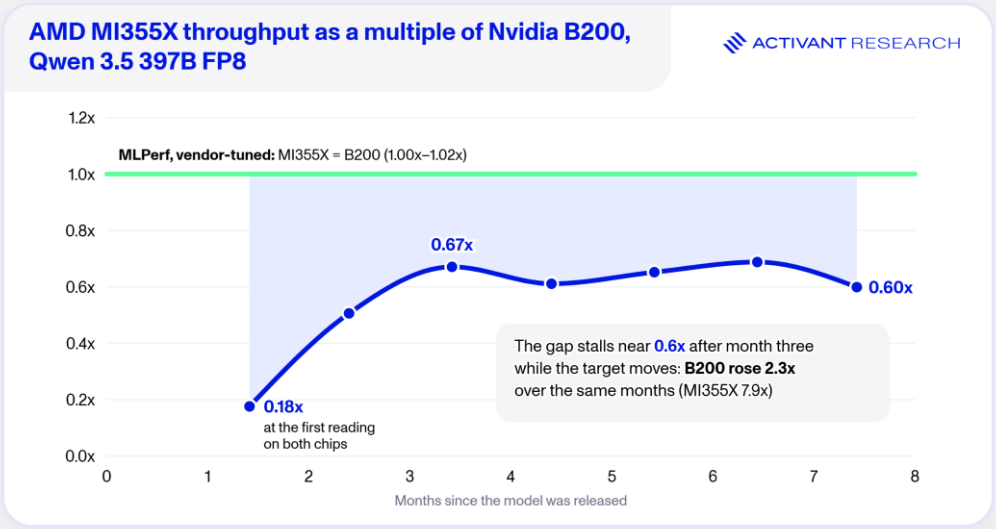
Compatibility is solved; optimization is not

Before these launches, running PyTorch on a TPU meant going through PyTorch/XLA, which made TPUs painful for PyTorch developers.[5] Trainium had a similar adapter, which ran only 46% of the 100 most-used text-generation architectures out of the box, while the new TorchNeuron runs 89%.[6] Google's new TorchTPU runs 98% of roughly 4,900 Hugging Face models, and trains Qwen3 235B within 5.7% of the throughput of Google's own JAX stack.[7]



But running is not the same as running well. TorchNeuron is a beta that skips about 22,000 of PyTorch's own tests and takes three to four minutes to compile its first training step.[8] Deepgram achieved about 60% more audio per dollar on Trainium, but that result came from a multi-month engagement with AWS.[9] If a hardware vendor doesn't offer an extensive library of kernels for optimizing model performance, customers need to hand-write them. That's what Cartesia still does for its latency-sensitive Trainium workloads.[10] Portability removed the compatibility moat. What remains is the optimization moat.

The data shows how much is left, and that it is a software gap rather than a silicon one. In MLPerf (an industry standard performance benchmark), where each vendor hand-tunes its own submission, AMD's MI355X matches Nvidia's B200 at 1.00 to 1.02x.[11] On the open-source SGLang engine that customers actually run, it delivers 0.59 to 0.79x. Over a model's life the gap is starker. On Qwen 3.5 397B, MI355X started at 0.18x of B200, reached 0.67x by its third month and has held near 0.6x since, while B200's own throughput rose 2.3x over the same period.[12]

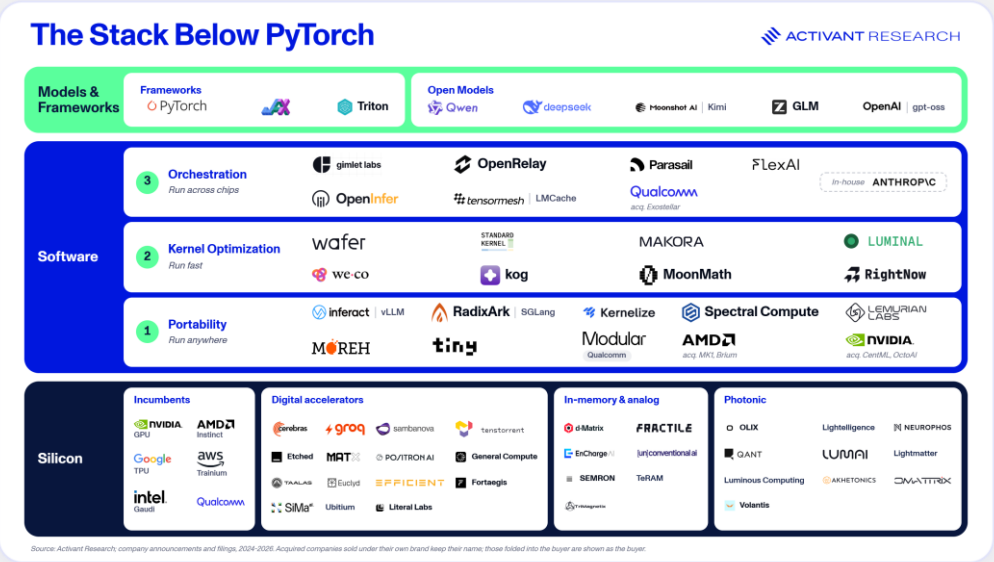


Software maturity is what closes that gap, and in a world where frontier AI is solving decades-old mathematics problems, writing good software for chips feels inevitable.

Google released an agent suite that writes and tunes TPU kernels on the same day TorchTPU went open source, and AWS's kernel agent produces correct Trainium kernels on 83% of tasks.[13],[14] No model yet writes kernels faster than well-configured stock PyTorch, and the best reaches 0.88x on KernelBench-Verified, but the chip owners are now funding the agents that will, which is why we hold the 24-month view.[15] Once a model both ports easily and loses nothing in performance, buyers run workloads on whatever hardware provides the best performance per dollar, or even split them across chips to find that optimum.

Anthropic already works this way. Its S-1 says Claude is trained and served natively across four chip families through a proprietary orchestration layer that took years to build, and that new chips now come online in a fraction of the time they once did.[16] Today only a lab with that density of engineering talent can build such a layer, but it shows where the rest of the market can go. The rest of the market gets there in three phases: first portability, which the clouds delivered last month; then kernel optimization, which agents are now automating; and finally orchestration across chips, which so far runs at scale only inside the frontier labs.

A cohort of startups is building each phase for everyone else. In portability, Spectral Compute's SCALE compiles existing CUDA code to run natively on AMD GPUs, Kernelize builds Triton compiler backends for GPUs, NPUs and TPUs, and Modular built Mojo and MAX to run one codebase across chips before Qualcomm bought it. In kernel optimization, Wafer, Standard Kernel, Makora and Luminal use AI models and search to write and tune kernels automatically. In orchestration, Gimlet splits a single inference workload across GPUs and SRAM-based accelerators, and OpenRelay routes inference across Nvidia, AMD, TPU and Trainium capacity. The first two phases are exactly the ones the chip owners are now giving away, which is where the business question begins.



Source: Activant Research, company announcements and filings, 2024-2026. Acquired companies sold under their own brand keep their name; those folded into the buyer are shown as the buyer.

Sell tokens, not infrastructure software

Taken to its extreme, silicon becomes a relative commodity and buyers match each workload to the hardware that optimizes its cost and performance. The first complication is that TorchNeuron and TorchTPU make PyTorch portable to Trainium and TPUs, but only to Trainium and TPUs.

Each cloud can combine its own chips with Nvidia's, but portability ends where the cloud ends, so this is bullish for the hyperscalers. The labs win alongside them, because they can move work between chips and play vendors against each other, as Anthropic does.

Independent silicon vendors like Fractile, Positron and OLIX gain: it becomes easier for them to win an initial contract as compatibility and performance become answerable questions, and they add heterogeneity that helps buyers manage their cost and performance objectives.

The second complication is the software layer that makes all this possible, and it is the harder case. A company selling an orchestration or compiler license sells to two types of customer: clouds and labs, and enterprises. The first can build it themselves or use open source, while the second isn't building its own AI infrastructure.

Past exits reflect that. Qualcomm bought Modular for about $3.1 billion in stock, Nvidia bought CentML and OctoAI, and AMD bought Nod.ai, Brium and MK1.[17] Attached to hardware, the software was worth a great deal, as in the case of Modular <> Qualcomm, yet none of these vendors reached scale as a standalone software business first.

So profit does migrate out of silicon, as we argued in July. It goes to whoever takes on the engineering that enterprises cannot, and sells the result by the token, attaching the software to the hardware.

That's the approach the newest cohort of software builders is following. Gimlet raised $300 million to build a multi-silicon inference cloud, starting with 100MW of Cerebras capacity.[18] Wafer moved from selling kernel optimization to selling inference, reaching an estimated $40 million of annualized revenue on OpenRouter alone, up from about $8 million across all channels in September.[19]

A neutral token seller can buy from every chip vendor, which a cloud cannot, and turn a cost gap into a price advantage. Wafer serves GLM-5.2 on MI355X at about 80% of B200's throughput for under half the cost, roughly 40% less per token,[20] and on OpenRouter it charges about half the average provider price for the same models.[21] But that lead is thin, Relace prices the same basket of models 8% below Wafer, and the whole field sells this traffic for cents per million tokens.[22] Kernel tuning wins one model at a time, and the chip owners are now automating it. The lasting cost advantage has to come from orchestration: routing each workload, or each phase of it, to the cheapest silicon that can serve it.

The CUDA moat is eroding, and profit is leaving silicon, but it is not moving to software sold as SaaS. The extreme scenario described above is at least two years away from feasibility, and arrives first in a narrower form, run by each cloud. The open form, where any model runs on any chip and the buyer never sees the silicon underneath, is the opportunity for a new type of inference cloud.

---

[1] InvestingCube, Largest Companies by Market Cap in 2026, 2026.

[2] Nvidia, Form 8-K, Q2 FY2027 Results, 2026.

[3] PyTorch, Introducing TorchNeuron, 2026; Google Cloud, TorchTPU Is Now Open Source, 2026.

[4] Company announcements and filings, including Qualcomm Form 10-Q; Activant Analysis, 2026. Seventeen acquisitions of hardware-neutral compiler, orchestration and inference-optimization companies, including Qualcomm's of Modular and Exostellar; Nvidia's of CentML, Run:ai, Deci and OctoAI; and AMD's of Nod.ai, Brium and MK1. The non-chip buyers include Red Hat (Neural Magic), which kept its target multi-vendor.

[5] PyTorch/XLA, Sources of Recompilation, 2026; SemiAnalysis, Amazon's Trainium Takes a Strike at the Training King, 2026. PyTorch/XLA compiled code into graphs at checkpoints the developer inserted by hand, so training loops had to be rewritten and line-by-line debugging broke. Every new input shape triggered an expensive recompilation; Trainium's adapter worked the same way.

[6] Hugging Face, TorchNeuron Model Support, 2026. Vendor-side measurement against Optimum Neuron v0.4.5.

[7] Google, TorchTPU (arXiv 2610.24161), 2026. 1,828 against 1,937 tokens per second on 64 v7 chips.

[8] SemiAnalysis, Amazon's Trainium Takes a Strike at the Training King, 2026. First step measured at 185 seconds eager and 256 seconds compiled.

[9] PyTorch, Introducing TorchNeuron, 2026. Deepgram figure is audio-seconds per dollar against GPUs, as reported by AWS.

[10] PyTorch, Introducing TorchNeuron, 2026.

[11] MLCommons, MLPerf Inference v6.0 Datacenter Results, 2026. Per accelerator in 8-chip systems, on Llama 2 70B and gpt-oss-120b.

[12] SemiAnalysis, InferenceX, Activant Analysis, 2026. Month-end MI355X throughput divided by B200 throughput on the same SGLang configuration; where one chip was benchmarked less often, its last reading is carried forward. The 0.18x launch point pairs MI355X's 20 February reading with B200's 3 March reading. Also the source of the 0.59 to 0.79x SGLang range in the preceding sentence (latest readings on DeepSeek R1 FP8 and Qwen 3.5 397B FP8 and FP4).

[13] Google, TPU Agent Suite, 2026.

[14] AWS, Stop Hand-Porting Kernels, 2026. 150 tasks; measures correctness, not speed.

[15] arXiv, KernelBench-Verified (2607.16241), 2026. Strict protocol against TF32 PyTorch on an H200; best model GPT-5.5.

[16] Anthropic, Form S-1, 2026, pp. 136, 205.

[17] Qualcomm, Form 10-Q, Q3 FY2026, 2026; GeekWire, Chip Giant Nvidia Acquires OctoAI, 2024; BetaKit, Nvidia Acquires Canadian AI Efficiency Startup CentML, 2025; AMD, AMD to Acquire Open-Source AI Software Expert Nod.ai, 2023; AMD, AMD Acquires Brium, 2025; AMD, AMD Acquires MK1, 2025. Modular consideration is about $3.1B at close, excluding about $700M of retention stock.

[18] Gimlet Labs, Gimlet Labs Raises $300 Million in Series B, 2026.

[19] OpenRouter, Wafer Provider Statistics, 2026; Activant Analysis, 2026. Wafer's daily tokens by model on OpenRouter over the ten days to 4 November 2026 (average 1.6 trillion a day), priced at Wafer's effective input and output price per model (after cache discounts) and each model's prompt-to-completion mix, give about $109,000 a day, or ~$40 million annualized on OpenRouter alone. September figure: Y Combinator, Wafer Raises $40M, 2026.

[20] Wafer, Wafer Raises $40M Series A, 2026. Self-reported, on GLM-5.2.

[21] OpenRouter, Effective Pricing by Provider, 2026; Activant Analysis, 2026. Each provider's blended price per million tokens after cache discounts, weighted by Wafer's model mix over the 30 days to 4 November: Wafer $0.086 against $0.155 volume-weighted across other providers.

[22] OpenRouter, Effective Pricing by Provider, 2026; Activant Analysis, 2026. Same method as footnote 21: Relace $0.079 per million tokens across 99% of Wafer's model mix.

Disclaimer: The information contained herein is provided for informational purposes only and should not be construed as investment advice. The opinions, views, forecasts, performance, estimates, etc. expressed herein are subject to change without notice. Certain statements contained herein reflect the subjective views and opinions of Activant. Past performance is not indicative of future results. No representation is made that any investment will or is likely to achieve its objectives. All investments involve risk and may result in loss. This newsletter does not constitute an offer to sell or a solicitation of an offer to buy any security. Activant does not provide tax or legal advice and you are encouraged to seek the advice of a tax or legal professional regarding your individual circumstances.

This content may not under any circumstances be relied upon when making a decision to invest in any fund or investment, including those managed by Activant. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by Activant. While taken from sources believed to be reliable, Activant has not independently verified such information and makes no representations about the current or enduring accuracy of the information or its appropriateness for a given situation.

Activant does not solicit or make its services available to the public. The content provided herein may include information regarding past and/or present portfolio companies or investments managed by Activant, its affiliates and/or personnel. References to specific companies are for illustrative purposes only and do not necessarily reflect Activant investments. It should not be assumed that investments made in the future will have similar characteristics. Please see “full list of investments” at activantcapital.com/companies/ for a full list of investments. Any portfolio companies discussed herein should not be assumed to have been profitable. Certain information herein constitutes “forward-looking statements.” All forward-looking statements represent only the intent and belief of Activant as of the date such statements were made. None of Activant or any of its affiliates (i) assumes any responsibility for the accuracy and completeness of any forward-looking statements or (ii) undertakes any obligation to disseminate any updates or revisions to any forward-looking statement contained herein to reflect any change in their expectation with regard thereto or any change in events, conditions or circumstances on which any such statement is based. Due to various risks and uncertainties, actual events or results may differ materially from those reflected or contemplated in such forward-looking statements.