THE HIDDEN LAYER SERIES | ARTICLE 7

*By Joe Batista | November 2026 | 15-minute read. Find the series here at the IdeaVault*

# The Layer Above You: What It Costs to Leave CUDA

*Why the software already written for an AI chip can matter as much as the chip itself*

*This is Article 7 of The Hidden Layer Series. Article 1 laid out the AI Value Map. Article 2 showed how bottleneck power migrates through the AI stack. Articles 3 to 5 followed it into packaging capacity, the energized megawatt and the rules that ration grid access. Article 6 found that at Layer 5 the price of money can matter as much as the price of chips. Article 7 moves down to Layer 2 and follows the code that runs on the chips.*

**CORE INSIGHTS**

1. **At Layer 2, the moat may be the software above the chip as much as the chip itself.** Nvidia reports more than 7.5 million developers using CUDA and its other software tools. The real switching question is how much production code, performance tuning and operational expertise a rival must help customers reproduce.
2. **The value is in the tuning.** By Nvidia's account, 400 of its engineers spent four months raising DeepSeek-R1's performance about fourfold on the same GB200 systems. A rival has to match work like that, not just run the code.
3. **Winning a second source now comes with visible financial incentives.** AMD granted OpenAI and Meta warrants for up to 160 million AMD shares each and committed to invest up to $5 billion in Anthropic. Nvidia invests in customers too.
4. **Switching takes three things: owning the software, being able to move it and having the engineers to tune it.** Several of the largest buyers adding other chips control part of that software. For most enterprises, the dependency may sit one layer higher, with their cloud or model provider.
5. **The code moat may be eroding from above.** Shared frameworks, open compilers and AI-written kernels lower the cost of moving code. Whether that ends the moat is a hypothesis, and we test it separately.

## The Toolkit Aimed at the Moat

Buying another AI chip is relatively straightforward. Getting the software that runs your business to perform equally well on it is another matter. The difference between those two decisions is where Nvidia's software advantage becomes an economic advantage.

On September 30, 2026, DeepSeek and Huawei took a step toward reducing that difference. DeepSeek announced software infrastructure for Huawei's Ascend AI chips spanning six components: TileLang, DeepGEMM-Ascend, DeepEP-Ascend, TileKernels, FlashMLA and DeepSelect. The announcement combined new repositories with updates to existing projects, and mirrored tools DeepSeek had already published for Nvidia GPUs. According to DeepSeek, as reported by Reuters, Huawei gave "full support," and the two companies jointly advanced a "supernode" solution based on 128 Ascend 950 chips. In Reuters' translation, DeepSeek said: "To build a new generation of independent, self-controlled GPU software ecosystems, the first priority is establishing a high-level language..."

One relevant project is TileLang, a language for writing kernels, the small programs that do the heavy mathematical work on a chip. Its main repository records native Ascend 950 support as of September 30, 2026, separate from earlier Ascend adapters published in an external project in September 2025.

The design choice was compatibility. DeepGEMM-Ascend, the Ascend version of DeepSeek's matrix-math library, is released under the MIT open-source license, describes itself as API-compatible with the original CUDA-based DeepGEMM implementation and uses the same package name, though hardware-specific differences in data layout remain. TechRadar's assessment: "the goal isn't a performance claim, but switching-cost reduction." Attention is early: at the time of its October 7 report, TechRadar counted 515 stars on GitHub for the Ascend library and 8,522 for the CUDA original.

Nvidia had already named the risk. In its quarterly report for the period ended July 26, 2026, it said its exclusion from China's data center market had "helped our competitors build larger developer and customer ecosystems to challenge us worldwide." Nvidia does not describe the threat only as a rival chip. It describes it as rival code.

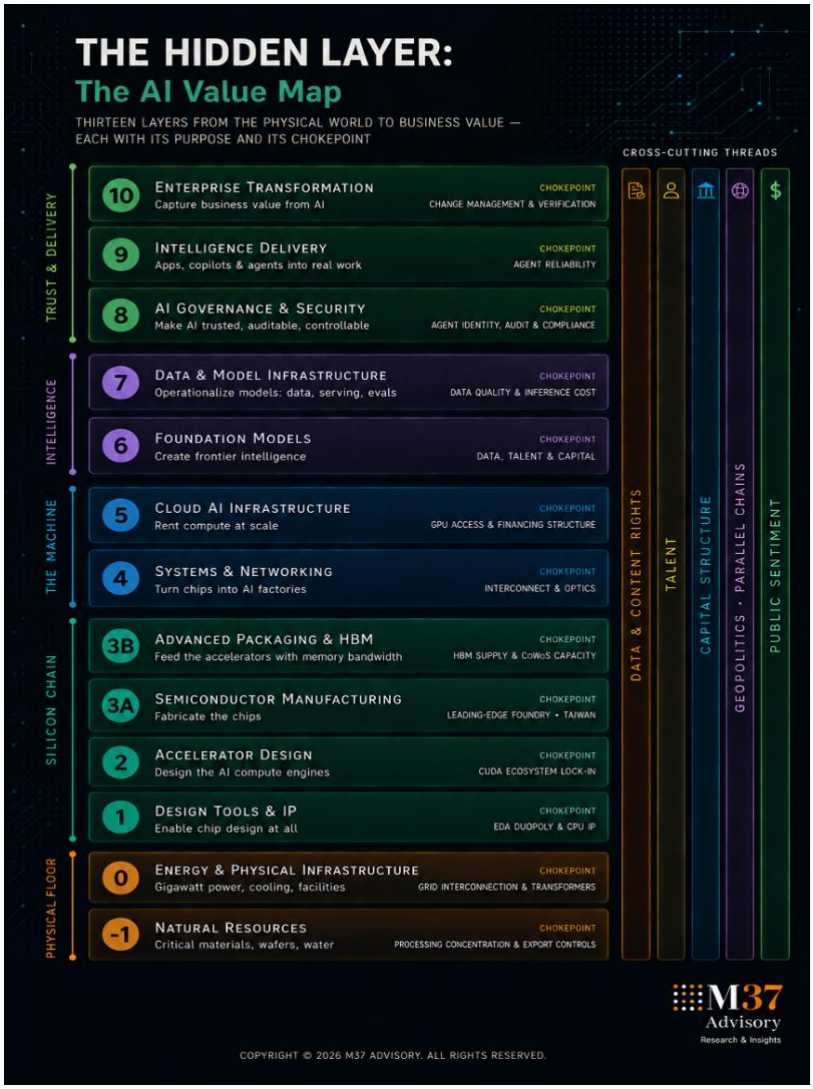
**The chips did not change. The cost of leaving the code behind them may have.**

That code is large. Nvidia's annual report describes CUDA as "the foundational NVIDIA CUDA development platform that runs on all NVIDIA GPUs," with "over 7.5 million developers worldwide using CUDA and our other software tools" and "hundreds of domain-specific software libraries." In the quarter ended July 26, 2026, Nvidia reported Data Center revenue of $89.0 billion, up 117% from a year earlier.

There is no defensible single price for leaving CUDA. This article examines what makes up that price, who pays it, and why getting code to run is different from restoring its performance.

> ***"A chip is sold once. The code written for it is a reason to buy the next one from the same company."***

## Where the Lock Sits




*The Hidden Layer: The AI Value Map. 13 layers, each with its purpose and its chokepoint.*

**ON THE MAP: Layer 2 — Accelerator Design | Chokepoint: CUDA Ecosystem Lock-In | This article: the software side of the chokepoint**

Article 2's working model says bottleneck power grows with Constraint × Market Structure × Customer Urgency, enabled by Capital Capacity. Switching cost is one of the four Market Structure dimensions. Articles 2 to 5 found the stack's tightest physical constraints in memory, packaging and power; Article 6 found the money. This article looks at the part of Layer 2's strength that isn't physical: what it costs a customer to take its work elsewhere. Two of the map's cross-cutting threads run through it: **Talent**, because in our reading much of that cost is specialist engineering time, and **Geopolitics**, because export controls have pushed one large market to pay it.

"The code" is not one thing. AI software reaches the chip through four layers, and each locks in differently:

- **The model code:** written in frameworks such as PyTorch or JAX. In principle it is portable; PyTorch, for example, also runs on AMD GPUs through AMD's ROCm software.
- **The kernels:** the small programs that do the heavy math, often hand-tuned for one chip architecture. Much of a model's speed depends on them.
- **The libraries and communication software:** vendor libraries such as Nvidia's cuBLAS, cuDNN and NCCL, and the software that moves data between thousands of chips.

- **The people and tools:** profilers, debuggers, cluster operations, and the engineers who know them.

**A note on "lock-in."** Nothing stops a customer from buying another chip. Lock-in here means the cost of making existing code run as well somewhere else. One element is contractual: Nvidia's CUDA license (section 1.2) prohibits reverse engineering, decompiling or disassembling specified output generated with its software in order to translate it to a non-Nvidia platform. That limits one route out, translating compiled output. It is not a prohibition on rewriting source code for another chip.

## Three Examples That Show the Code

### ① The Value Is in the Tuning

Switching has two costs. The first is getting code to run on a new chip. The second is getting it to run as fast as it did before.

At Nvidia's GTC conference in March 2026, Ian Buck, its vice president of hyperscale and high-performance computing, told Computer Weekly that 400 Nvidia engineers spent four months optimizing the open-weight DeepSeek-R1 model on GB200 systems, making 38 major optimizations, including kernel fusion, tensor parallelism and a four-bit number format. "We increased the performance of DeepSeek-R1 by four times on the same GPU infrastructure," he said. "We just increased the revenue of every GB200 by four times, all in software."

Both lines are Buck's characterization. We have not been able to confirm which performance measure he meant, whether throughput per GPU, tokens per second or something else, so we treat "four times" as Nvidia's account rather than a measured fourfold increase in business output. The revenue remark describes potential output, not reported revenue.

The example is optimization on Nvidia's own platform, not a move away from it, and it does not measure the cost of leaving CUDA. It shows how much of a chip's output can depend on software work done after the chip ships. The relevant switching cost is the incremental effort and lost output involved in achieving comparable business performance on the alternative platform. That cost varies by workload and cannot be inferred from the effort spent optimizing the incumbent. Keeping the same package name and interface, as DeepSeek did, lowers the cost of getting code to run. It does not, by itself, deliver the tuning.

> ***"The chip sets the ceiling. The software decides how close you get."***

**THE SWITCH IN BRIEF**

- **Mechanism:** tuning on one platform sets the performance a rival has to match. Until it does, the customer gives up output.
- **Limit:** the main public example is one vendor's account of optimizing its own hardware. It shows the value of tuning, not the cost of switching.
- **Implication:** price the parity, not just the port.

### ② Incentives for a Second Source

Between October 2025 and July 2026, AMD offered equity and engineering support to three of the largest AI buyers as they committed to its chips. In our reading, part of that support helps offset switching costs; part buys deployment commitments and helps finance expansion.

**OpenAI.** On October 6, 2025, AMD announced a 6-gigawatt agreement to supply OpenAI with multiple generations of Instinct GPUs, starting with 1 gigawatt of MI450 GPUs in the second half of 2026, and a warrant

for up to 160 million AMD shares. Tranches vest as deployments scale, and also depend on AMD's share price and on OpenAI reaching technical and commercial milestones. AMD's chief financial officer, Jean Hu, said the partnership was "expected to deliver tens of billions of dollars in revenue for AMD," a forecast, not recognized revenue.

**Meta.** On February 24, 2026, AMD and Meta announced a multi-year agreement for up to 6 gigawatts, starting with a custom MI450-based GPU running on ROCm, AMD's answer to CUDA. Meta received a warrant for up to 160 million AMD shares with conditions structured similarly to OpenAI's. AMD's announcement gave no dollar value, but AMD said it would sell up to $60 billion of chips over five years, Reuters reported that day.

**Anthropic.** On July 22, 2026, AMD announced that Anthropic would deploy up to 2 gigawatts of MI450-series GPUs from the first half of 2027, building on its existing use of AMD's MI355X GPUs. This time the equity went the other way: AMD "committed to make a strategic equity investment of up to $5 billion in Anthropic in the future." The teams "will use Claude to optimize workloads for AMD Instinct GPUs and accelerate ROCm software development."

Two cautions. First, equity is not unique to the challenger. In September 2025, Nvidia announced a letter of intent contemplating investment of up to $100 billion in OpenAI; by March 2026 its investment had instead been reported by Reuters at $30 billion, and Jensen Huang said the $100 billion opportunity was "probably not in the cards." In November 2025, Nvidia announced an investment commitment of up to $10 billion in Anthropic, with engineering work on "optimizing future NVIDIA architectures for Anthropic workloads." Equity now moves to win customers and to keep them, and some of it finances the buildout, the subject of Article 6.

Second, none of these customers is leaving Nvidia. A week before the AMD agreement, Reuters reported that Nvidia had signed a multiyear deal to sell Meta millions of its current and future AI chips. OpenAI and Anthropic buy from Nvidia too. These buyers are purchasing a second source and the bargaining power that comes with it, and paying the parity cost only for the workloads that move.

For context, Nvidia reported a 75.0% gross margin, GAAP and non-GAAP, for the quarter ended July 26, 2026, and AMD reported 56% non-GAAP (54% GAAP) for the quarter ended June 27. Both are company-wide figures for businesses broader than AI accelerators. They do not establish how much room either has to price its accelerators, and they are not a measure of the CUDA premium.

***"AMD isn't only selling chips. It is using equity and engineering support to help customers adopt another platform."***

**THE SWITCH IN BRIEF**

- **Mechanism:** the challenger offers equity and engineering support to secure volume; the incumbent invests to keep customers close.
- **Limit:** equity also finances the buildout, and every buyer named here still buys Nvidia. The deals show second-sourcing, not abandonment.
- **Implication:** assess any incentive against the full obligations you accept in return.

## ③ The Buyers Who Own the Layer Above

Several of the largest buyers adding alternative chips also control or co-develop important parts of the software layer between their models and the hardware. Meta created the PyTorch framework and moved it to the independent PyTorch Foundation in 2022. OpenAI released Triton in July 2021, an open-source language for

writing GPU kernels without writing CUDA C/C++ code. DeepSeek publishes its own kernel libraries and, since September 30, Ascend versions of them.

Anthropic is a different kind of example: a company developing and optimizing workloads across several platforms. In October 2025 it agreed to expand to up to one million Google TPUs, describing at the time "a diversified approach that efficiently uses three chip platforms": Google's TPUs, Amazon's Trainium and Nvidia's GPUs. It has since added AMD.

These cases point to three separate conditions for switching, and they are not interchangeable:

- **Software ownership:** the company controls some of the software connecting its models to the hardware.
- **Software portability:** that software can move between chip architectures.
- **Engineering capacity:** the company has enough specialist talent to tune the new platform to parity.

Owning software does not make it portable, and portable software does not automatically run efficiently on another chip. In our reading, the buyers best placed to add a platform are those with all three, because they can spread the parity cost across very large fleets. Most enterprises rely on others for them.

China is the forced test. Export controls restrict Nvidia's sales there, so Chinese developers face the switching cost whether or not they choose it, and the September 30 release is part of paying it.

A newer force works on the code itself. On February 27, 2026, researchers at ByteDance Seed and Tsinghua University submitted CUDA Agent, a system trained to write CUDA kernels. On the KernelBench benchmark, its kernels beat those of PyTorch's torch.compile on 100%, 100% and 92% of tasks at the three levels: shares of tasks won, not speed improvements, on small self-contained tasks. Tech Wire Asia reported in early October that it had found no independent replication. Buck sees the trend as a strength: "It's actually accelerating CUDA adoption." AMD and Anthropic are betting the same tools can build ROCm faster. Both can be true. Which effect is larger is the open question.

***"The buyers adding alternative chips at scale often control part of the software layer above them."***

**THE SWITCH IN BRIEF**

- **Mechanism:** ownership, portability and engineering capacity together let a buyer pay the parity cost once and use it across a large fleet.
- **Limit:** the pattern rests on a few very large buyers. They may switch because they are large, not because they own software.
- **Implication:** before you commit, ask whether you have all three conditions, or who supplies them for you.

## Why This Pattern Keeps Repeating

**History doesn't repeat, but it rhymes.** The closest precedent is the contest for 64-bit computing in the early 2000s. Intel and HP's Itanium, launched in 2001, introduced a different instruction set. Strong native performance required software porting and optimization, while running legacy x86 code was a weakness. AMD took the other route: its Opteron, launched in April 2003, extended x86 to 64 bits while keeping compatibility with existing 32-bit code. In June 2004, Intel introduced Xeon processors with a compatible x86-64 extension.

The software followed the installed code. In March 2011, Oracle announced it would stop new software development for Itanium, saying "Intel management made it clear that their strategic focus is on their x86

microprocessor"; Intel disputed that, saying its Itanium work "continues unabated." Intel's discontinuance notice later set January 30, 2020 as the last order date and July 29, 2021 as the last shipment date.

The cases differ. In 2001 the new architecture asked customers to port their code; in 2026 the incumbent owns the installed code and the challengers are asking. AI code is also younger and more concentrated in a few frameworks. The analogy supports a mechanism, not an outcome: the platform that keeps existing code running has an advantage. When the software moves, the hardware can follow. It does not tell us whether frameworks and AI-written kernels will make CUDA code as easy to carry as AMD made x86 code.

## The Switching-Cost Equation

Article 2's model needs a real shortage, a concentrated market and urgent customers, enabled by capital. At Layer 2, CUDA's contribution is Market Structure, through switching cost, and unlike a shortage, switching cost isn't the same for every buyer.

> **THE M37 LENS: THE ECONOMICS OF SWITCHING (A SIMPLIFIED MODEL)**
>
> **Net value of switching = the present value of the alternative platform's savings over the life of the commitment**
>
> **minus: porting the code + re-tuning to parity + output lost below parity + people and tools + running two platforms + execution risk**
>
> **The decision:** does the lifetime economic advantage exceed the full cost and risk of moving?
>
> The savings depend on the commercial terms and the alternative's delivered performance. The costs are set largely by the code already written and by who owns the layer above it. **Running two platforms** matters most for buyers that second-source rather than switch: duplicated engineering, validation, monitoring, support and operational complexity. The model also mixes two clocks: the savings arrive over years, while the cost of leaving is paid up front.
>
> Three parties can help bear the cost: **the buyer**, with its own engineers; **the new supplier**, with warrants, investment or engineering help; and **the community**, with open-source tools and AI-written code. Buyer-specific help may benefit one customer first; improvements to shared frameworks and libraries can spread more widely, whoever funds them. The model leaves out supply, rack-scale systems and the bargaining value of a second source.

| Who bears the switching cost | How | Example | What it costs | Who keeps the value |
|---|---|---|---|---|
| **The buyer** | Its own engineers and its own software layer | Meta with PyTorch and AMD; OpenAI with Triton | Sustained engineering; needs scale | The buyer, as lower cost and bargaining power |
| **The new supplier** | Warrants, equity investment, joint engineering | AMD warrants to OpenAI and Meta; up to $5 billion committed to Anthropic | Dilution or cash | Shared: the buyer gets equity, the supplier gets volume |
| **The community** | Open-source languages, libraries and AI-written kernels | DeepSeek's Ascend release; OpenAI's Triton; CUDA Agent | Slower, uneven progress toward parity | Every buyer, including those using it only to bargain |

The same chip can sit in two market structures. Article 2's rubric scores switching cost in dollars per design, and the public evidence does not let us place either kind of buyer in those bands, so we do not score them. In our

reading, the lock-in may be weakest at the top of the market, where the largest purchases are made, and strongest with buyers that have less purchasing scale and less control over their software.

> ***"Nvidia sells the same chip to everyone. It doesn't sell the same lock-in."***

**WHAT WE MEAN BY "THE CODE IS THE MOAT"**

**Observed:** by Nvidia's account, software tuning raised one model's performance on the same hardware about fourfold; DeepSeek built its Ascend tools for interface compatibility with its CUDA tools; each of AMD's three multi-gigawatt GPU agreements came with warrants or an equity commitment; and several of the buyers adding non-Nvidia chips control software above the chip.

**Inferred, the switching-cost thesis:** the cost of moving code, above all regaining performance, is a real input to accelerator decisions, borne in engineering work and commercial incentives. In our reading it supports Nvidia's position for as long as rivals must offer such incentives.

**Hypothesized, the code-moat hypothesis (unproven):** the moat is the code already written, not just the chip. If so, it holds most firmly where buyers lack the three conditions, and it erodes as shared frameworks, compilers and AI-written kernels make code cheaper to move. Chip performance, networking, supply and Nvidia's pace of new generations may matter as much.

**What would weaken both:** smaller buyers without their own software adopting rival chips at scale with no incentive; or Nvidia's position holding firm even among buyers that can move their code, which would point to the hardware and systems, not the software, as the moat.

## The Switching-Cost Test: Four Questions You Should Be Asking

1. **Where does your code live?** How much of your AI workload sits in portable framework code, and how much in hand-tuned kernels, vendor libraries and tools tied to one chip?
2. **Port or parity?** What would it cost to make your workloads run elsewhere, what would it cost to make them run as fast, and what does running two platforms cost in between?
3. **Do you own the layer above you, can it move, and can you tune it?** Are your frameworks, compilers and inference software yours, open, or your vendor's? What do their license terms allow?
4. **Who is paying you to switch, or to stay?** If a supplier offers equity, credits or engineers, what do you commit in return, and how much of the support actually offsets the cost of moving your code?

Most enterprises will answer these questions one layer higher. They may not own the model, the inference framework, the compiler, the kernel libraries, the cloud infrastructure or the engineers needed to tune any of it. Many consume AI through a cloud or model provider. For them, the dependency may not be CUDA at all, but the provider's abstractions, APIs, tools, data connections and operating environment. Know whether your prompts, fine-tuned models, integrations and workflows would work on another provider's stack, and what it would cost to rebuild them if they wouldn't.

## The Prediction

Same rules as Articles 2 to 6: checkable by anyone, marked in public in October 2028, including where we were wrong. One indicator tracks the **switching-cost thesis**. A separate test checks the buyer pattern predicted by the

**code-moat hypothesis**. Three supporting indicators track the business context; they inform the reading but don't decide it.

### Switching-cost thesis: indicator of the mechanism

| Prediction (by Oct 2028) | Supports the thesis | Weakens it | Status |
|---|---|---|---|
| **Suppliers keep offering financial adoption incentives** | **Supports:** at least three new publicly disclosed agreements, from the date of publication to June 30, 2028, in which a non-Nvidia accelerator supplier grants warrants or equity, commits an equity investment, or provides financing support to a customer as part of a commitment to deploy at least 1 gigawatt or $10 billion of its accelerators. The three AMD agreements above predate publication and don't count. This indicator tracks supplier-supported adoption; it does not isolate the portion of support attributable to software switching costs | Fewer than three qualifying agreements | Open |

### Code-moat hypothesis: separate test of the predicted buyer pattern

| Prediction (by Oct 2028) | Supports the predicted buyer pattern | Weakens the predicted buyer pattern | Status |
|---|---|---|---|
| **Adoption concentrates among the largest AI builders** | **Universe:** publicly disclosed commitments to deploy non-Nvidia AI accelerators of at least 500 megawatts or $5 billion, announced from the date of publication to June 30, 2028. **Classification:** each buyer is (A) one of the largest AI builders on the fixed list below, or (B) any other buyer: enterprises, GPU clouds, sovereign and government programs, telecom operators. **Supports:** at least 75% of qualifying commitments by count are from group (A) | Group (B) accounts for more than 25% of qualifying commitments by count. Untestable if fewer than four commitments qualify. The test shows where adoption happens, not why | Open |

### Supporting indicators (business context; they inform the reading but don't decide it)

| Prediction (by Oct 2028) | What it would show | What would weaken it | Status |
|---|---|---|---|
| **The CUDA base keeps growing** | Nvidia's fiscal 2028 annual report states at least 10 million developers using CUDA and its other software tools (7.5 million in its fiscal 2026 report) | Below 10 million, or the figure is no longer reported | Open |
| **Rivals reach scale** | AMD's Data Center segment revenue for fiscal 2027 is at least 12% of Nvidia's Data Center revenue for fiscal 2028. In the quarters ended June 27 and July 26, 2026, it was about 7.6% (our calculation) | Below 12% | Open |
| **Nvidia's company-wide profitability holds** | Nvidia reports a company-wide non-GAAP gross margin of at least 70% for fiscal 2028 (75.0% in the quarter ended July 26, 2026). Margins move with product mix, pricing, manufacturing costs and systems revenue, so this is business context, not evidence for or against the CUDA thesis | Below 70% | Open |

**Measurement rules, fixed now**

- **Non-Nvidia accelerator supplier:** any company other than Nvidia that sells or leases AI accelerators it designs or has designed for it, including AMD, Intel, Alphabet, Amazon, Microsoft, Qualcomm, Huawei, Cambricon and Cerebras, and custom-chip partners such as Broadcom and Marvell when they are party to the agreement.
- **Qualifying agreements and commitments:** counted at announcement, from company releases, filings, or two independent major news reports. "Up to" amounts count at the stated maximum. Memorandums of understanding, letters of intent and frameworks don't count. A company deploying chips it designs itself, for its own use, doesn't count; chips it sells or leases to another company do.
- **Why two thresholds:** the incentive indicator uses 1 gigawatt or $10 billion because it targets the largest deals, where supplier support has so far been disclosed; all three AMD agreements above were 2 gigawatts or more. The buyer-pattern test uses a lower bar, 500 megawatts or $5 billion, so it captures enough commitments to classify. Both are design choices fixed now.
- **Counting once:** an expansion, amendment or follow-on announcement under an existing agreement or commercial program between the same supplier and buyer counts as the same agreement, unless it adds a new site or a new commitment that on its own meets the size threshold. Separate announcements of the same deployment count once, at the larger disclosed size.
- **Supplier support:** warrants, equity grants, equity investments or commitments to invest, guarantees, backstops or financing provided by the accelerator supplier or its affiliates, disclosed as part of the same agreement or announced together with it.
- **Group (A), the largest AI builders:** Alphabet, Amazon, Microsoft, Meta, Oracle, Apple, OpenAI, Anthropic, xAI, Alibaba, ByteDance, Tencent, Baidu and DeepSeek, including subsidiaries, fixed at publication. The list is a proxy for the buyers most likely to have software ownership, portability and engineering capacity; it cannot separate those conditions from scale. A purchase by a third party for a named group (A) end customer counts as group (A).
- **Gross margin:** Nvidia's company-wide, full-year non-GAAP gross margin for fiscal 2028 as reported in its fourth-quarter results release. If Nvidia stops reporting non-GAAP figures, we use GAAP.
- **Developer count:** as stated in Item 1 of Nvidia's Form 10-K for fiscal 2028.
- **Revenue ratio:** AMD's Data Center segment revenue for fiscal 2027 and Nvidia's Data Center revenue for fiscal 2028, as reported in each company's Form 10-K. AMD's segment includes server CPUs, and the fiscal years end at different times (late December and late January); the ratio is a broad revenue-scale comparison, not a market-share estimate. Figures are as first reported.

How we'll read the result: the thesis indicator tracks whether rivals still need financial incentives to win adoption; it does not measure switching cost directly. The buyer-pattern test can support or weaken the code-moat hypothesis by showing where adoption happens, but not, on its own, why. We will publish the results, including any that challenge our original view.

**What did your last port cost? How long did parity take? Show me the numbers. The map improves when you do.**

**THE BOTTOM LINE**

Articles 3 to 6 found constraints in capacity, power, rules and capital. This article finds one in accumulated software. The chip is only part of the commitment. Every model optimized, every kernel tuned, every library integrated and every engineer trained adds to the cost of moving elsewhere.

For the largest AI buyers, that cost may be manageable: those that own enough of the software layer, and have the engineers, can spread the effort across enormous workloads. Most enterprises don't have that advantage. Their dependency may sit one layer higher, with the cloud or model provider that controls the software between them and the hardware.

Nvidia's moat isn't necessarily permanent. But neither is it measured only by the performance of the next chip. It is measured, in part, by how much work a customer must reproduce before another chip can do the same job.

**So, ask yourself: If you had to run your AI on another chip next year, what would it cost to run it as fast, and who would pay?**

## How This Article Was Tested

An independent fact-check team reviewed every statistic and quotation on the draft's 42-item fact-check list. We deleted claims that could not be verified at their original source and corrected those that could. Before publication, the team re-verified all 32 checkable claims in the final manuscript against primary sources and found no material errors. What remains falls into three groups.

**Checked against the original document:** Nvidia's fiscal 2026 Form 10-K (Item 1), its second-quarter fiscal 2027 results release and Form 10-Q; AMD's OpenAI announcement (SEC-filed), its Meta and Anthropic announcements and its second-quarter 2026 results release; Nvidia's OpenAI letter-of-intent announcement; Anthropic's announcements on Google TPUs and on its partnerships with Microsoft and Nvidia; Nvidia's CUDA license; the TileLang and DeepGEMM-Ascend repositories; the CUDA Agent paper; OpenAI's Triton and Meta's PyTorch Foundation announcements; Intel's Itanium discontinuance notice; and Article 2 of this series.

**Reporting, attributed in the text:** the DeepSeek and Huawei announcement and DeepSeek's quotation (Reuters); TechRadar's assessment and GitHub counts; Ian Buck's remarks (Computer Weekly); AMD's $60 billion figure for the Meta deal (Reuters); the Nvidia–Meta deal (Reuters); Nvidia's later OpenAI investment and Jensen Huang's remarks (Reuters); Tech Wire Asia on CUDA Agent replication; and Oracle's 2011 statement and Intel's response (The Register, eWeek).

**Our calculations and judgments:** the AMD-to-Nvidia Data Center revenue ratio, labeled as ours where it appears.

The review changed the article. We updated Nvidia's September 2025 OpenAI letter of intent with the reported $30 billion investment and Huang's March 2026 remarks, so the $100 billion figure is no longer presented as a current commitment. We removed the engineering-cost comparison and its dollar estimates: Nvidia's optimization of DeepSeek-R1 on its own hardware and engineers' estimates for moving other models to Huawei's chips measure different work, and the porting estimates could not be verified at their original source. For the same reason we removed a report on which chips Chinese developers use for training. We described DeepSeek's release as new and updated components rather than six new projects; cited Reuters, Nvidia's quarterly filing and Nvidia's license directly; dropped an unverified ownership percentage for OpenAI's AMD warrant; corrected the CUDA Agent date and described its results as shares of tasks won; and made the Itanium history and the margin and revenue comparisons more precise.

The team then reviewed the full manuscript against the fact-check record. We removed wording that went further than the evidence: supplier incentives are no longer described as a measure of switching cost; statements about

customers "leaving" Nvidia now describe second-sourcing; categorical claims about which buyers can move, and who can lower costs, were qualified; prediction labels were matched to what each test can show; and the counting windows now start at publication. A final editorial check tightened four passages and removed the Market Structure scores, which could not be tied to Article 2's measures with the evidence available. A last editorial pass reframed the article around three conditions for switching, added the cost of running two platforms and the enterprise view, moved the margin indicator to business context, and stopped describing DeepSeek's whole release as open source: only DeepGEMM-Ascend's license was confirmed.

Practitioners who have ported AI workloads between chip platforms, and those who buy or price accelerators: if you see where the evidence is thin, say so in the comments. Challenges that change the analysis will be credited by name in the next update.

## About the Author

Joe Batista advises boards, C-suite executives, and go-to-market teams on technology value creation, growth strategies, and the opportunities hidden beneath major technology shifts. He has spent more than 40 years in the technology industry, guiding companies on growth and technology decisions, spanning the minicomputer era at Digital Equipment Corporation (DEC) through client-server, cloud, virtualization, and now AI, with executive experience at Dell Technologies and Hewlett Packard Enterprise.

Through M37 Advisory, Joe helps leadership teams recognize recurring patterns, uncover second-order effects, find the opportunities others overlook, and identify the strategy worth betting on. His work is grounded in a simple belief: **What if the biggest obstacle to your growth isn't the competition, but the way you think?**

> *"Most executives have been bombarded with 'theories' about what might work. Based on our research, Joe's extensive playbook is rooted in creating alignment of interest and business outcomes that leaders should take notice of."*
>
> — Jim Dickie, Research Fellow & Sales Mastery Founder, CSO Insights

**Author Contact**

Email: Chief.Creatologist@M37-Advisory.com

Website: https://www.m37-advisory.com

Idea Vault: https://www.m37-advisory.com/idea-vault

**NEXT IN THE SERIES**

**Article 8: Governance Is the New Supply Chain** — We move up the map to Layer 8, AI Governance & Security, where the chokepoint is agent identity and audit. Every AI system built on the chips discussed here will run work that someone, or some agent, authorized. Proof of who authorized what an agent did may become the scarce input. You can only defend AI where you can prove what it did.

**To subscribe to the series,** complete the form

## Sources

### Primary documents

- NVIDIA: Form 10-K, fiscal 2026 (fiscal year ended Jan 25, 2026): CUDA description, 7.5 million developers using CUDA and other software tools, "hundreds of domain-specific software libraries"

- NVIDIA: Form 10-Q, quarter ended July 26, 2026: "larger developer and customer ecosystems" quotation
- NVIDIA: Financial results for second quarter fiscal 2027 (Aug 26, 2026): Data Center revenue, 117% growth, 75.0% gross margin
- NVIDIA: CUDA Toolkit End User License Agreement (section 1.2): translation limitation
- AMD and OpenAI announce strategic partnership to deploy 6 gigawatts of AMD GPUs (Form 8-K, Exhibit 99.1) (Oct 6, 2025)
- AMD and Meta announce expanded strategic partnership to deploy 6 gigawatts of AMD GPUs (Feb 24, 2026)
- AMD and Anthropic strategic partnership (Jul 22, 2026)
- AMD: Second quarter 2026 financial results (Aug 4, 2026)
- NVIDIA: OpenAI and NVIDIA announce strategic partnership to deploy 10 gigawatts of NVIDIA systems (Sep 22, 2025)
- Anthropic: Microsoft, NVIDIA and Anthropic announce strategic partnerships (Nov 18, 2025)
- Anthropic: Expanding our use of Google Cloud TPUs and services (Oct 23, 2025)
- DeepGEMM-Ascend repository (GitHub): MIT license, API compatibility, same package name
- TileLang project repository (GitHub): developers, native Ascend 950 support (Sep 30, 2026), earlier external Ascend adapters (Sep 29, 2025)
- ByteDance Seed and Tsinghua AIR: CUDA Agent, arXiv 2602.24286 (v1) (submitted Feb 27, 2026)
- OpenAI: Introducing Triton, open-source GPU programming for neural networks (Jul 28, 2021)
- Meta: Announcing the PyTorch Foundation (Sep 12, 2022)
- PyTorch: PyTorch for AMD ROCm platform now available as Python package
- M37 Advisory: Article 2, The Accidental Kingmaker, The Hidden Layer Series: working model and switching-cost rubric

**Reporting (the claims that rely on it)**

- Reuters: DeepSeek partners with Huawei to develop chip programming tools (Sep 30, 2026): "full support," 128-chip supernode, DeepSeek quotation
- TechRadar: Huawei and DeepSeek release open source AI tools (Oct 7, 2026): release components, "switching-cost reduction," GitHub star counts as of Oct 7
- Tech Wire Asia: Nvidia's CUDA moat was built on engineers (Oct 2026): CUDA Agent replication status
- Computer Weekly: CUDA at 20 (Mar 23, 2026): Ian Buck quotations, 400 engineers, four months, 38 optimizations
- Reuters: Nvidia to sell Meta millions of chips in multiyear deal (Feb 17, 2026)
- Reuters: AMD clinches second mega chip supply deal, this time with Meta (Feb 24, 2026): AMD's $60 billion, five-year figure
- Reuters: Nvidia will not be able to invest $100 billion in OpenAI due to IPO, CEO Jensen says (Mar 4, 2026): $30 billion investment, Huang quotation
- The Register: Oracle stops Itanium development (Mar 23, 2011): Oracle statement
- eWeek: Oracle ends Itanium support, draws ire of Intel, HP (Mar 23, 2011): Intel's response
- The Register: Intel kills Itanium again (Feb 1, 2019): last order and shipment dates (Intel PCN 116733-00)


*Keywords: AI infrastructure, Layer 2, accelerator design, CUDA, switching cost, ROCm, AMD, Nvidia, Huawei Ascend, DeepSeek, TileLang, AI Value Map, M37 Advisory*