citrini · 2026-07-20

All Along the AI Watchtower

Qualcomm’s Card, the Builder’s Toolkit, Enterprise Spending, and… PEMDAS

Bob Dylan wasn’t musing about artificial intelligence related capital expenditures when he wrote “All Along the Watchtower”. Still, he managed to summarize the AI value-accrual debate pretty well…

> *“Businessmen, they drink my wine, plowmen dig my earth. None of them along the line know what any of it is worth”*

Everyone along the AI value chain thinks they can price the piece directly in front of them. The value of the whole, and who gets to keep it, is where there’s less confidence. Against the backdrop of a historic momentum correction, it’s even harder to see through fog.

But in our view, many of the current narratives – open source alternatives, cost competition, Kimi3 – fail to adequately explain the market’s selloff. Even with optimizations, the pie of aggregate token spend is growing rapidly. Chinese lab competition is not new, and Kimi’s K3 open source model is both massive and now compute-constrained. Even the renewed Iranian conflict doesn’t explain why the industries relatively insulated from an energy shock have underperformed the broader market.

Rather, it’s more likely that a blistering, leverage-fueled, breakout of the most crowded expressions of the AI and other thematic trades creates its own fragility – the unwind generally mirrors the vigor of the ascent.

Still there is a lesson to be learned. In the throes of the uptrend it’s easy to neatly agree on beneficiaries or casualties, usually by looking at which stock price is going up or down. But as volatility increases, so does the value of finding AI upside in places that aren’t unanimously agreed upon.

On one hand, the washout over the past several weeks helps clear the deck and could lead to a more durable leg among well known winners. ***But today, we’re trying to find the names that have been put in the wrong category...***

A mobile-chip designer that turns constraints into advantage; the tools that will be called more even with fewer seats; and the biggest winners of IBM’s earnings miss; and… PEMDAS (you’ll have to read to find out).

We begin with the most literal example of an old label obscuring a new business…

1\. Qualcomm

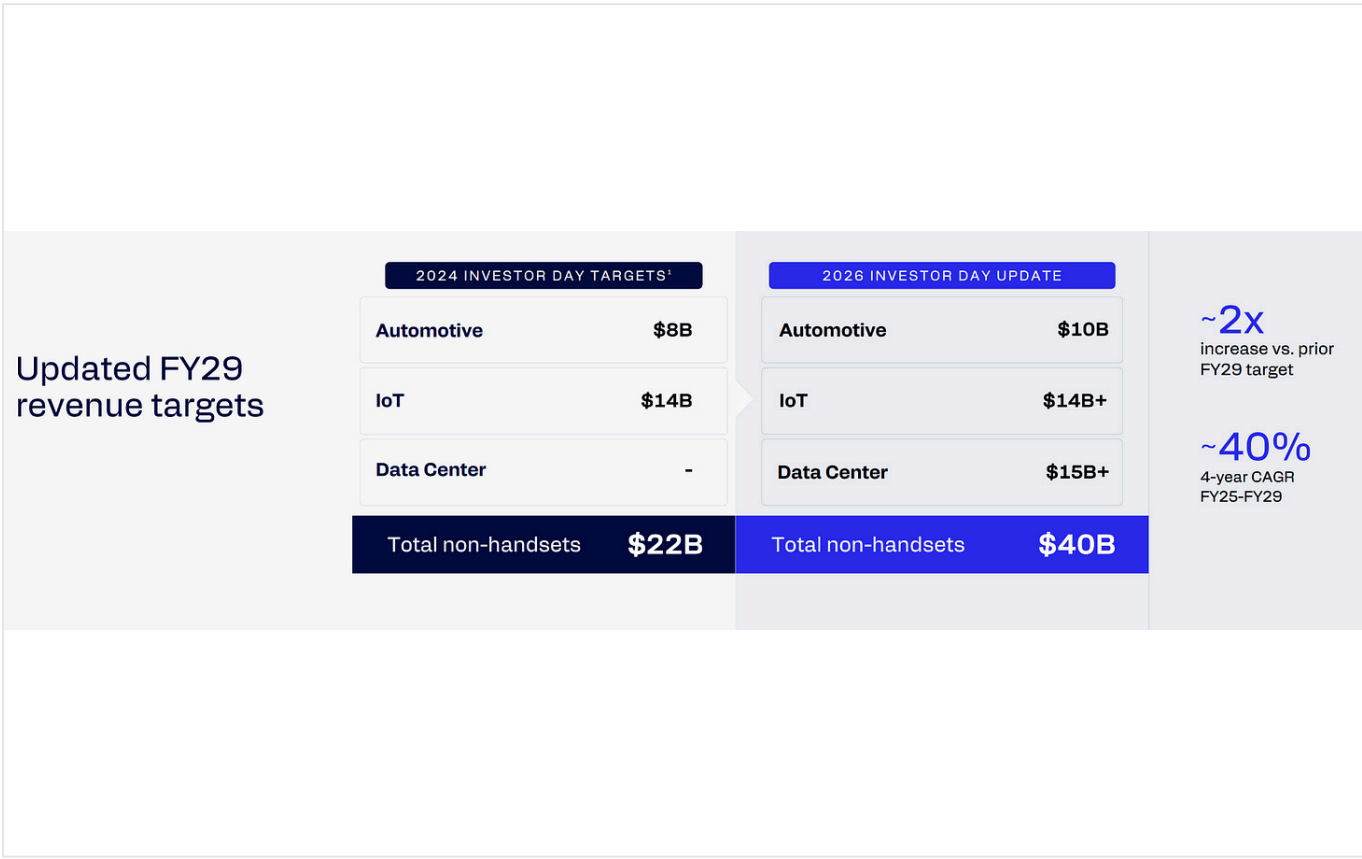
-

Last week, we *thought* we had a big scoop on **Qualcomm (QCOM US)** – hidden HTML on their site pointing to a yet-to-be-announced partnership with Anthropic. Unfortunately, sources close to both Anthropic and Qualcomm denied this emphatically.

So, we killed the article.

But over the past few days we've felt a pull to speak about what QCOM is doing anyway, because even if they don't have a partnership now… it doesn't mean they won't have one in the future.

Qualcomm's Investor Day on June 24 was a big event. The company made an incredibly ambitious raise to their guidance for FY29 that nearly doubled their non-handset revenue expectations from $22 billion to $40 billion.



The major driver is in data center revenue – and while we recognize investor reluctance to underwrite QCOM data center revenues after its failed attempt to enter the market back in 2017 with the Centriq 2400 – there are reasons to take this attempt seriously.

The largest constraints in AI processing today, power and memory bandwidth, are areas a mobile chip designer has unique experience in tackling. In particular, their approach to get around the "memory wall" could be the biggest selling point.

Qualcomm announced a new architecture for "High Bandwidth Compute" that uses LPDDR and something similar to "near-memory processing". The bigger implication, beyond just QCOM's revenue guidance, is something we've been talking about since last December.

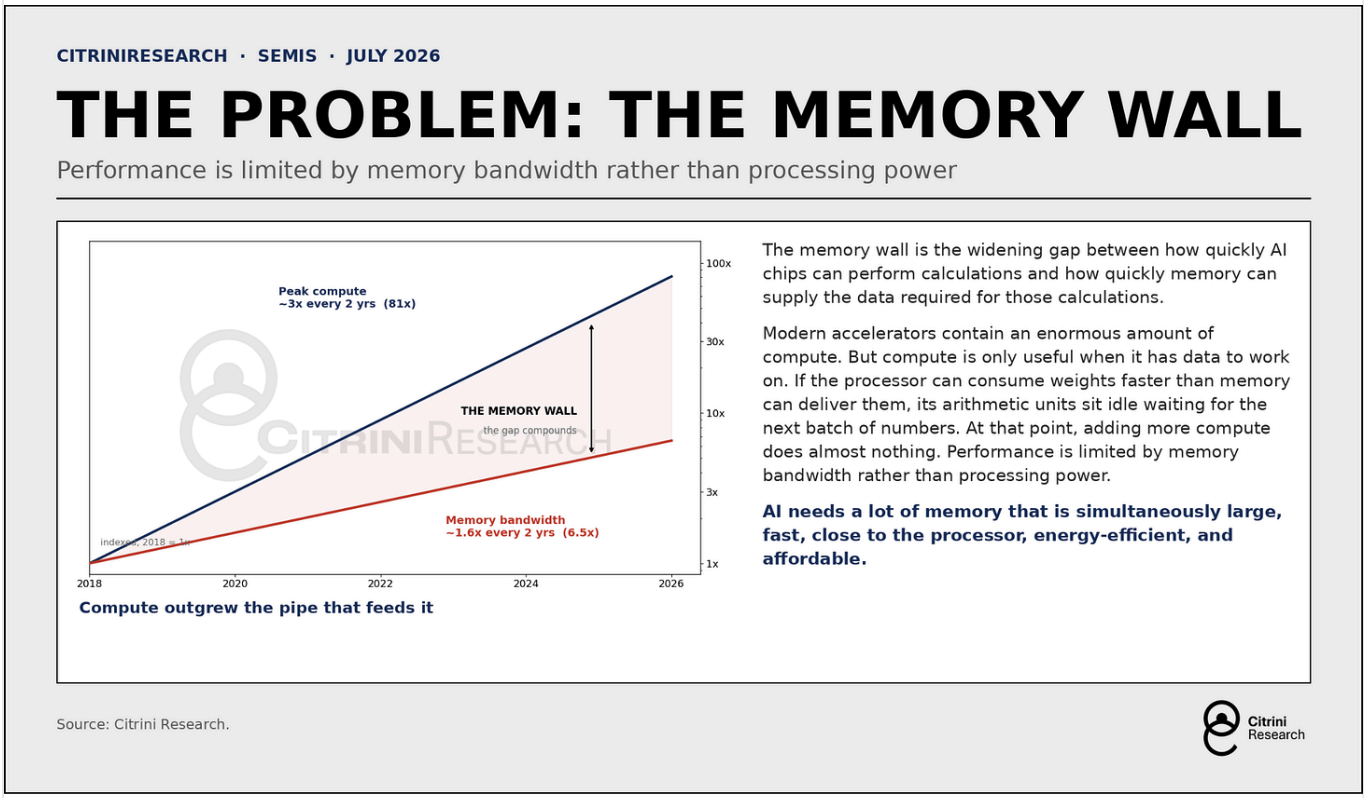
The oldest line in economics is “the cure for high prices is high prices”. In most markets, the cure comes around via supply. In technology, it’s not just supply that can cure high prices but also innovation. **The higher DRAM prices go, the more incentive there is to find ways around HBM.**

HBM isn’t an immutable requirement, it’s just the industry’s current answer to the cost of moving enormous amounts of data back and forth between memory and the accelerator. As that answer becomes more expensive, engineers have more incentive to attack the problem elsewhere: move simple computation closer to the memory, keep colder weights and KV cache in cheaper tiers, and send the accelerator only the data it actually needs.

Near-memory compute is one expression of that shift. The irony is that every increase in HBM pricing strengthens near-term memory revenues while also strengthening the case for architectures designed to use fewer HBM bytes per token.

Back in **26 Trades for 2026**, we discussed some approaches from the software side that could allow memory to be used more efficiently. But all of them are very subject to Jevons’ Paradox if the memory used has the same margin profile. It’s a change to the hardware and architecture that has more impact on the memory trade.

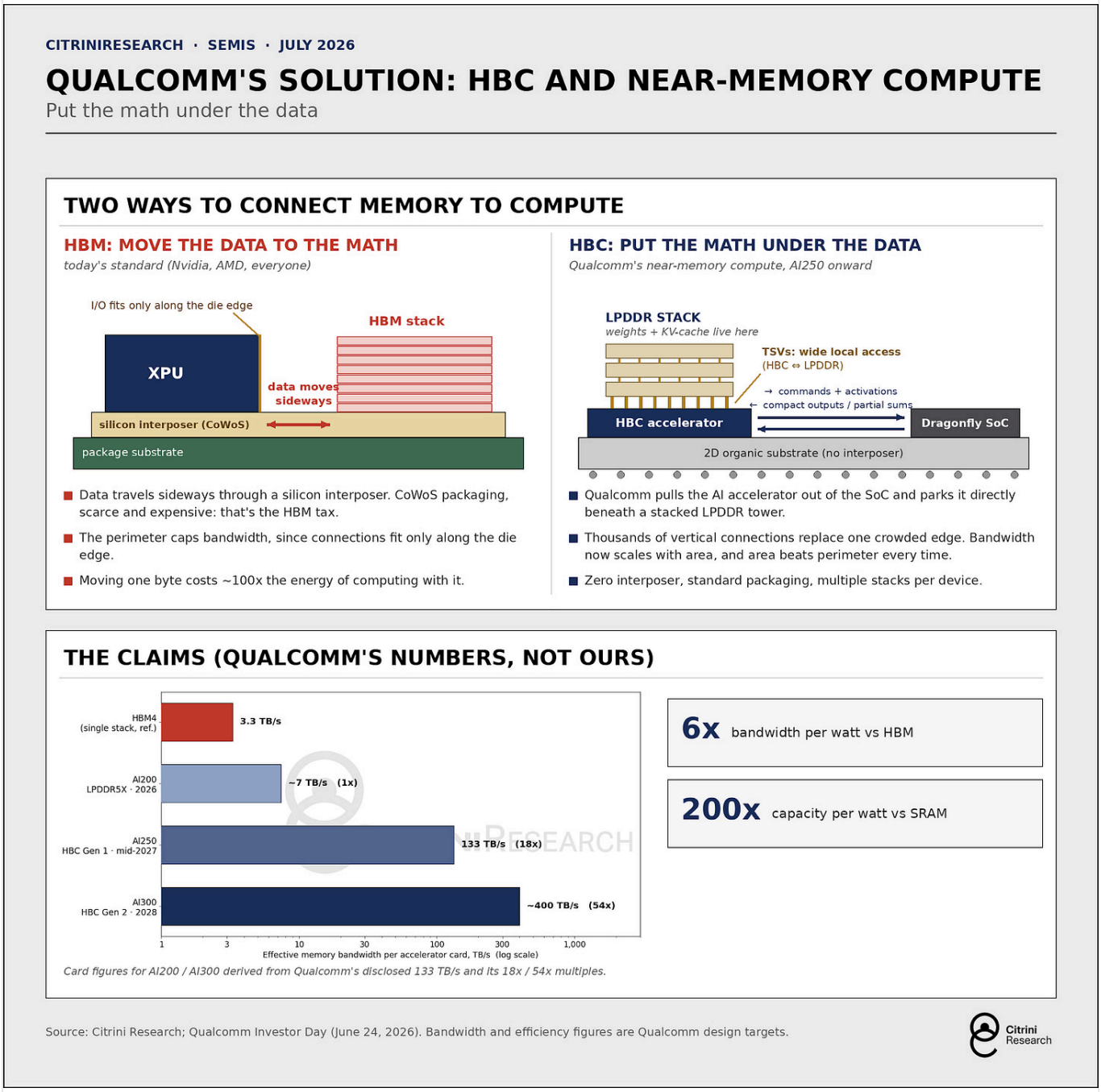
You’re all probably familiar with the memory wall, and the fact that memory-bound inference is the largest contributing factor to it. Just in case, though, let’s refresh.



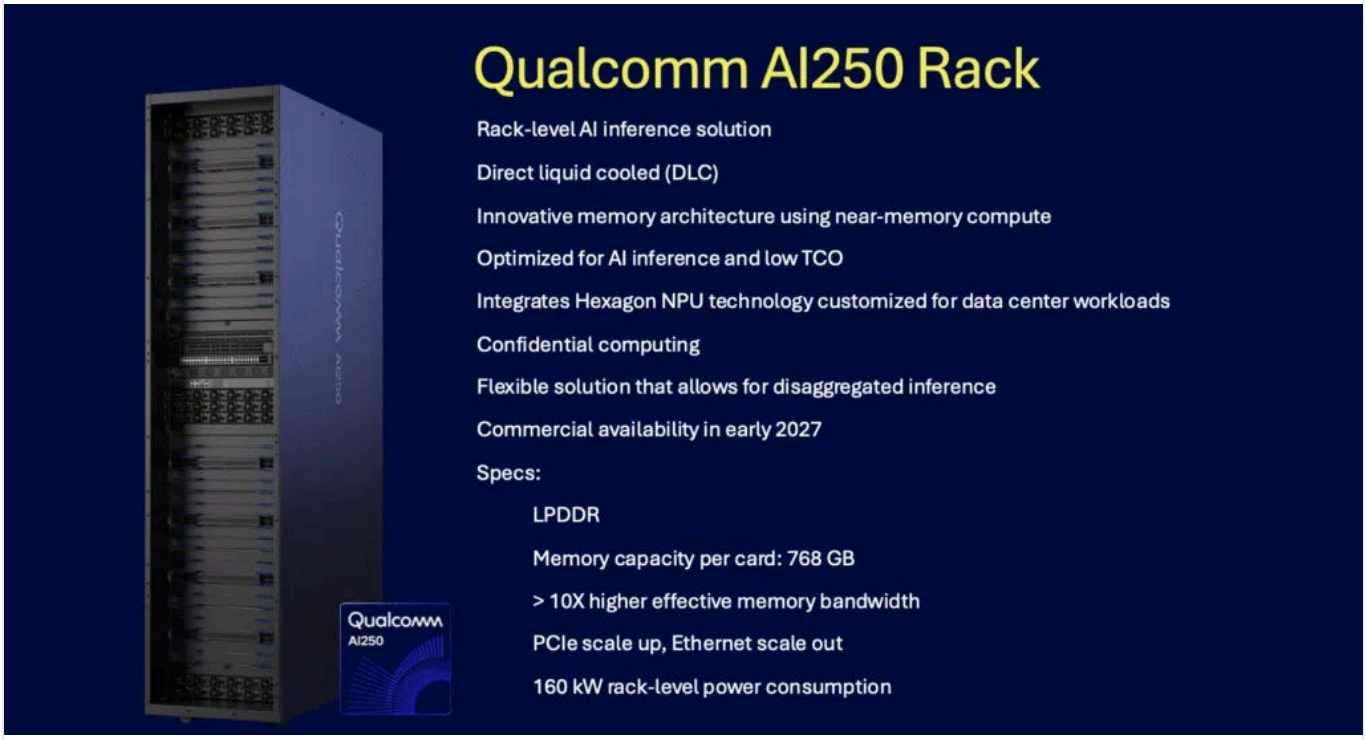
Simply put, compute capacity scaled faster than the ability to feed it. Decode-heavy inference is bound by bytes per second – by stacking DRAM with advanced packaging (silicon interposers) these needs are met in HBM (high bandwidth memory). This has resulted in a market that is extremely short DRAM but is exploring several potential solutions that involve ways around it.

While there's been significant focus from the software side on using memory more effectively, the hardware side is taking detours around the HBM toll booth. Primarily, this boils down to keeping model data on-chip, replacing DRAM with cheaper media, moving compute closer to the data, matching memory to each inference phase, and pooling capacity across accelerators.

Qualcomm's version, High Bandwidth Compute (HBC), which stacks the processor directly beneath the memory, will arrive in 2027.



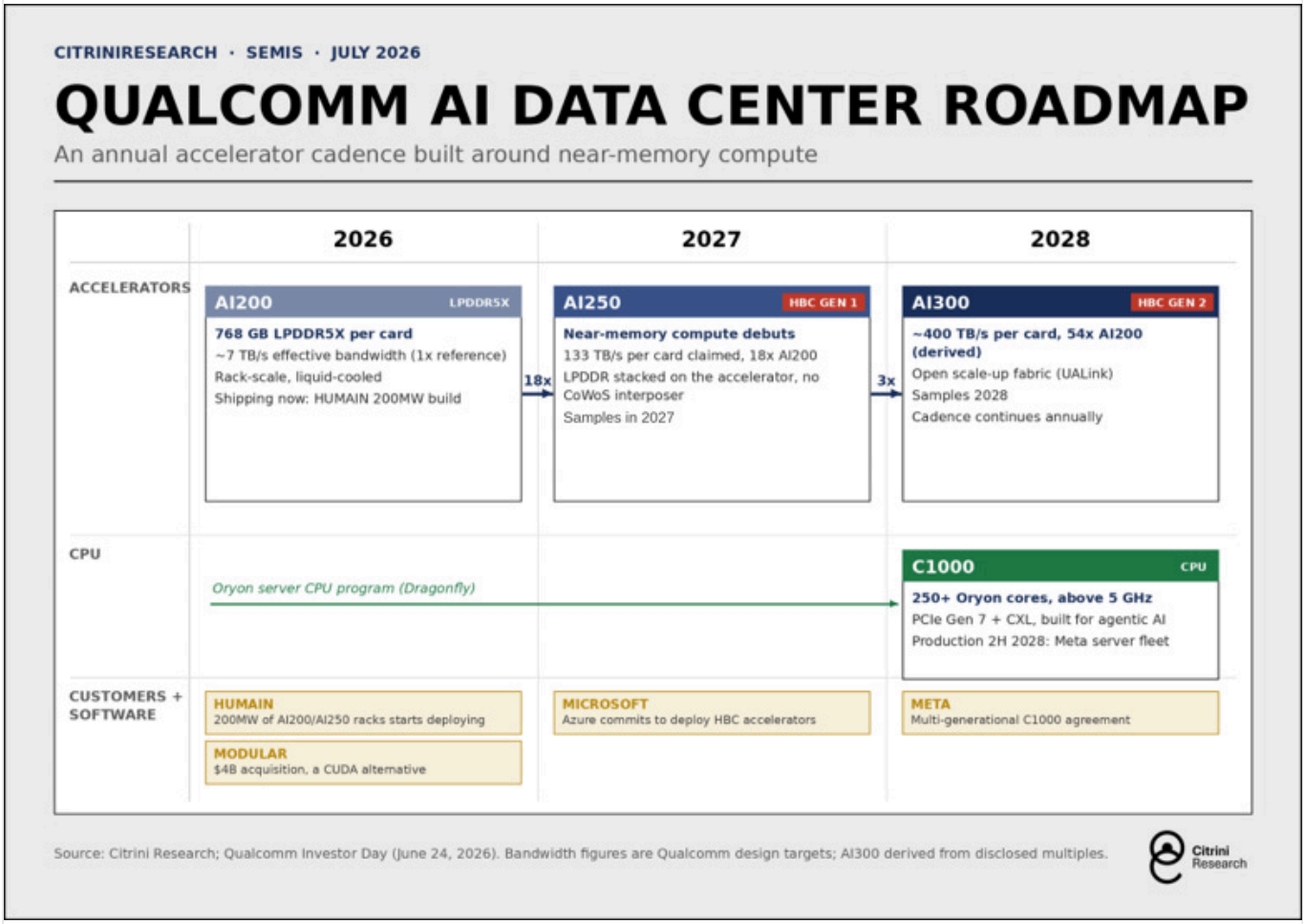
HBC takes the AI accelerator from the SoC and places it directly under an LPDDR DRAM stack, connected by through-silicon vias, allowing bandwidth to flow vertically across the whole die rather than out on the edges like HBM. Tony Pialis, QCOM's EVP of Data Center, claims this gets SRAM-like performance with stacked-memory density, eliminates the interposer (*and expensive CoWoS scale constraints*) and allows them to tile multiple HBC stacks in one device without advanced packaging. This is **near memory compute (or what Qualcomm describes as "similar to near memory compute")** – high capacity per stack, low power, and stackability.



The result, they claim, is a chip that yields 6x as much bandwidth per watt compared to HBM with 200x the capacity per watt compared to SRAM. Not just that, they claim 4-8x better decode performance per watt and on a TCO basis. It uses all the things that Qualcomm was forced to get good at with mobile devices (low power in a small area).

Note – the claim of 133 TB/s most likely describes aggregate effective bandwidth between LPDDR and the HBC accelerator during memory bound decode operations. The sustained end-to-end decode will be lower than this headline number.

Qualcomm also laid out its AI roadmap at their Investor Day, showing an annual AI200 -> AI250->AI300 accelerator cadence along with a CPU offering in 2028.



Even without the Anthropic partnership we believed was suggested by this secret code hiding on their website (that disappeared after we reached out for comment), QCOM's commercial data center entry is **much more credible than it was six months ago:**

- Alphawave contributes SerDes, optical DSPs, CXL/UCIe, chiplets and custom-silicon execution
- Qualcomm says two hyperscaler custom-silicon programs begin contributing in FY27
- Meta has signed a multi-generation CPU agreement, with C1000 production expected in the second half of 2028
- The pending Modular acquisition supplies the open software layer needed to make models portable across Qualcomm and third-party accelerators

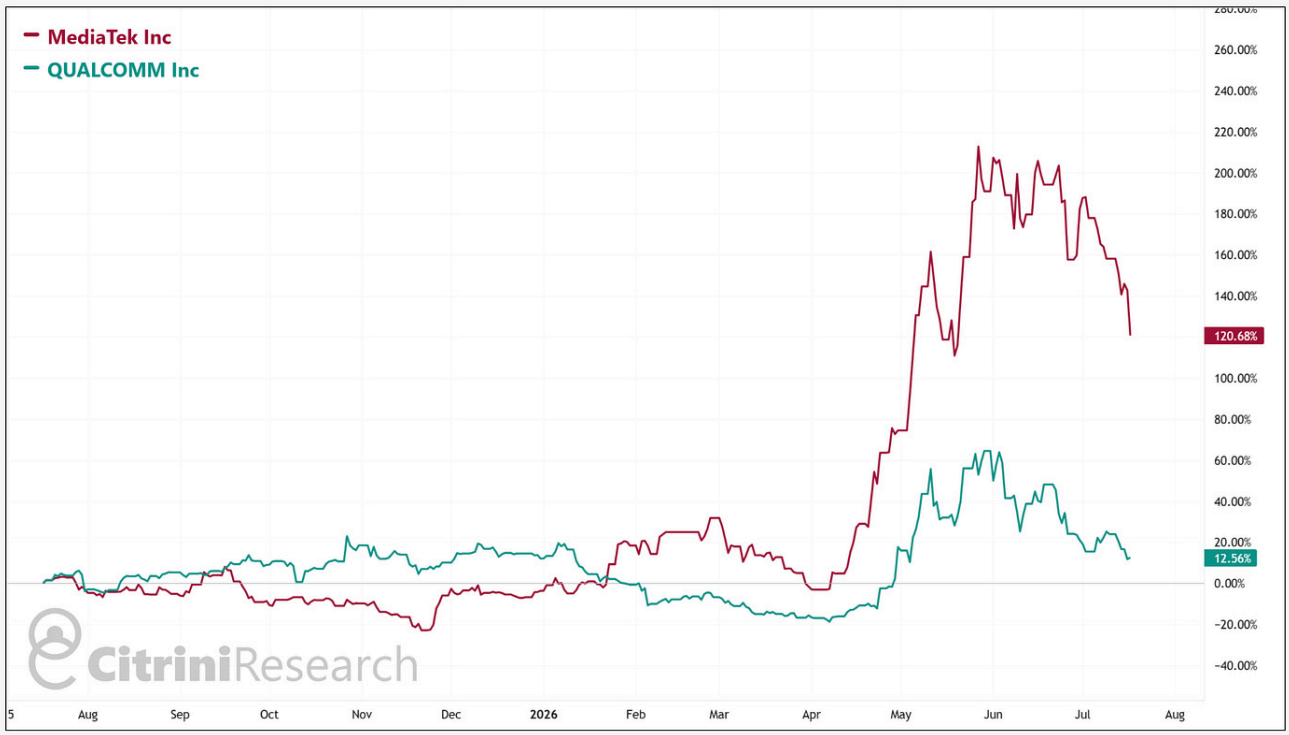
And to people who say *"sure, but this won't have an impact until 2028/29"*, I'd just remind them: that is ***precisely the year we are putting multiples on this late into the game.***

And it's certainly not a far-fetched idea – Anthropic is aggressively diversifying compute. Its announced stack already includes Google TPUs, Nvidia GPUs, and Amazon Trainium. Qualcomm, meanwhile, is offering

exactly the kind of alternative architecture Anthropic has been accumulating: complete, rack-scale systems designed specifically to run trained AI models.

Qualcomm has been in the proverbial doghouse for semis for pretty much the entire cycle. Citrini Research came into 2026 bullish on the fabless mobile chip names, both **MediaTek (2454 TT)** and Qualcomm. Despite their exposure to a rough handset cycle, we thought they still could pull off a re-rating.

MediaTek came into the year trading just shy of 20x NTM earnings…



Source: Citrini Research

MediaTek has re-rated hard as TPU upside brought it out of the handset cold and into the warm embrace of the data center. **Qualcomm has recently been very vocal about capturing data center market share,** so… can it pull off the same feat?

The Path

Its product and revenue timeline remain early. The first AI200 racks arrive this year, while Meta's C1000 deployment doesn't begin production until the second half of 2028.

That makes each new customer validation unusually important. In its coverage of the announcement, TD Cowen wrote "its FY29 $15 billion target could prove conservative, as it only includes its current existing customers across its four product families (connectivity, ASICs, merchant XPUs, and merchant CPUs); any additional customers would present potential upside."

The math to make these projections requires a lot of faith in a management team that hasn't earned it quite yet. We're watching for early indications that management can pull off the required pivot to data center without failing to capitalize on the areas where the rest of the business could catch a tailwind: automotive through robotaxi deployments, mobile robotics, and the AI device constellation.

CITRINIRESEARCH · SEMIS · JULY 2026

WHAT QUALCOMM COULD BECOME

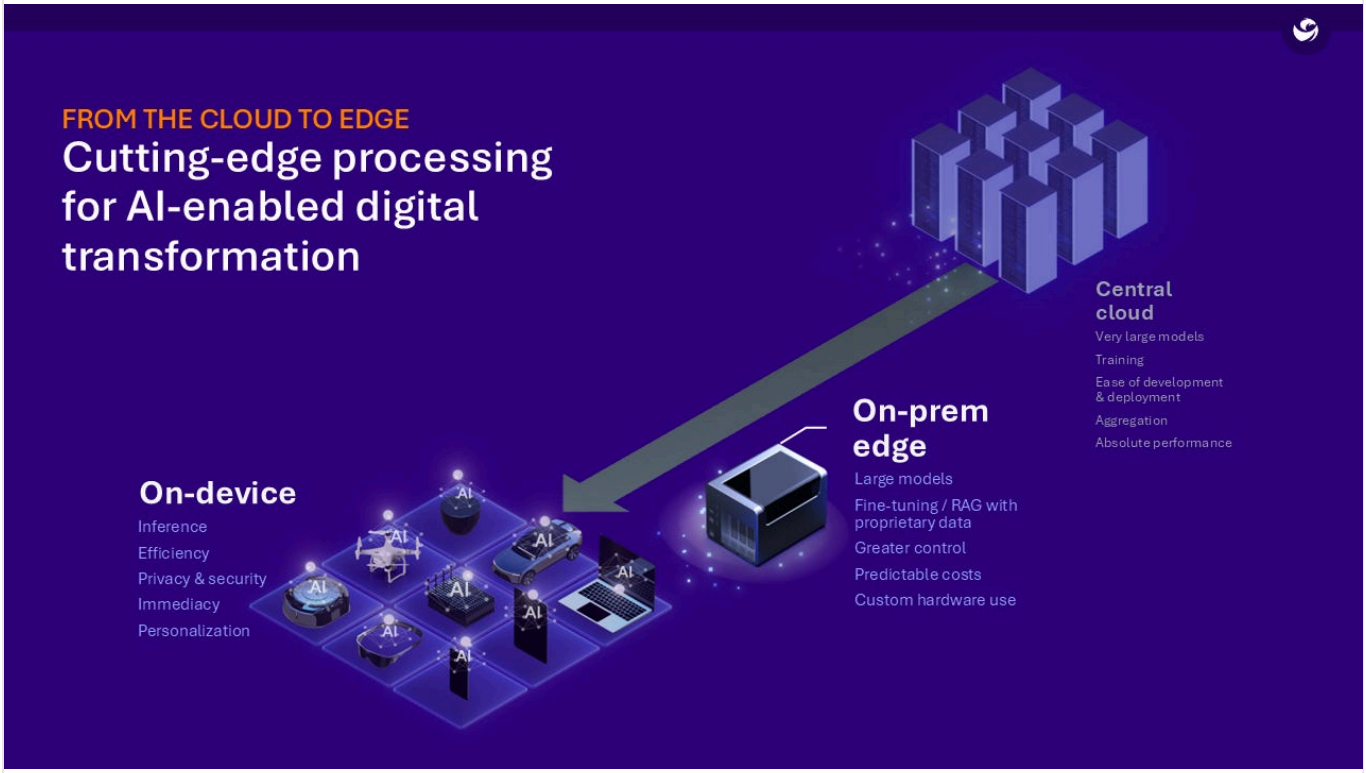
Three opportunities, ranked, and the FY29 targets attached to them

	OPPORTUNITY	WHAT QUALCOMM COULD BECOME	FY29 MANAGEMENT TARGET
1	Data-center inference and the memory wall	*The lowest-cost token engine for agentic inference*	>$15B Data center
2	Personal AI device constellation	*Merchant silicon for every non-Apple personal agent*	$6B Personal AI + Compute
3	Robotics and physical AI	*The brain and nervous system of robots*	$8B Industrial + Networking + Robotics

Source: Citrini Research; FY29 targets per Qualcomm Investor Day (June 24, 2026).

Citrini Research

However, we are actually more bullish on the automotive and IoT segments than the company itself, and these are the areas that can make up for hiccups. The automotive segment did $1.33 billion in the March quarter, up 38%, and management says content per vehicle rises eightfold between its third- and fifth-generation platforms (the $65 billion pipeline is cumulative expected program revenue over the life of the wins, so conversion rate is the variable worth watching).



The IoT target, meanwhile, did not move higher: remaining at ~$14 billion. Any pickup in the constellation of AI devices between now and 2030 that Qualcomm naturally fits into means significant upside here, and we think that's more likely than not.

Qualcomm already has a solid foothold in the consumer end of the constellation via Meta's smart glasses. The company's **Snapdragon AR1** sits in every pair of Ray-Ban Metas, pairing a 12-megapixel camera with an image signal processor that lets users deploy multimodal edge AI in everyday situations. These are the most successful smart glasses shipped to date: EssilorLuxottica sold more than 7 million pairs of Meta AI glasses in 2025 alone (including both Ray-Ban and Oakley), and is on track to hit its 10 million pair annual capacity target ahead of schedule. EssilorLuxottica is seeking to double capacity to 20 million pairs or more, citing robust demand.

What matters more for Qualcomm is the silicon content: the AR1 is estimated at roughly one-third of the bill of materials, or $50-60 per pair. In the upside case of 10 million pairs this year and 16 million next, that's a collective ~$1.4 billion of revenue attributable to consumer AI across the next two years.

Finally, robotics is the most significant upside optionality in QCOM's forecast. QCOM's Dragonwing IQ10 is a deployment-ready platform that could become a standard for startups in emerging robotics, which demonstrates how QCOM's robotic content is the whole perception stack: SoC, vision, sensor processing, connectivity, safety silicon, motion control interfaces and the software wrapped around them. Guidance embeds about $700 million, anchored by the Figure collaboration and Arduino (which could be significant if hobbyist robotics gains momentum). Any surprise that approaches our expectations for robotics volume over the next four years makes current guidance a rounding error.

2\. The Builder's Toolkit

-

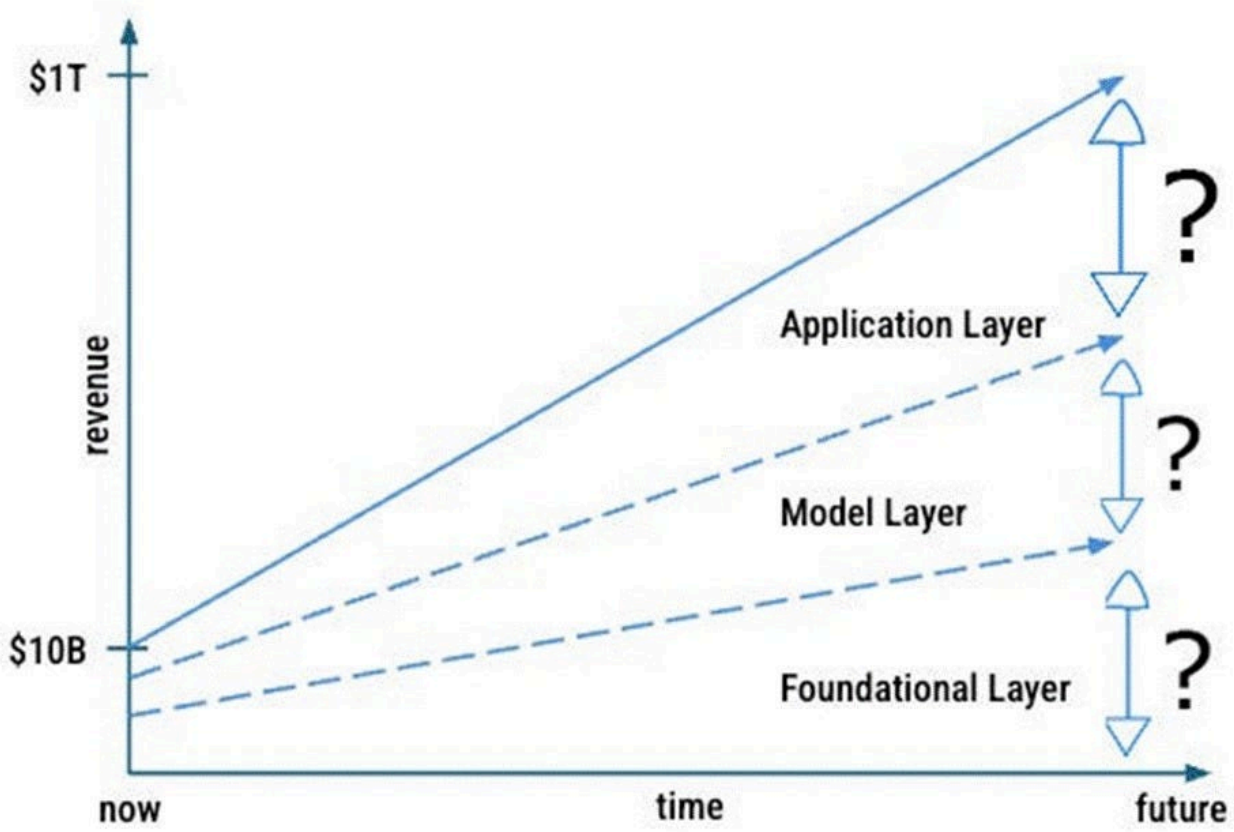
If you've been following the market for the past two years, you might think "AI" is an incredible economic revolution where the world's largest companies bludgeon each other senseless building the world's biggest thing; walled in metal, filled with silicon, circulated with electricity, cooled by refrigerants, connected by fiber.

But at some point, doesn't value ***need*** to accrue outside the infrastructure in order to justify the buildout?

Back in **April 2025**, we refreshed our "Phase 2" AI framework, arguing:

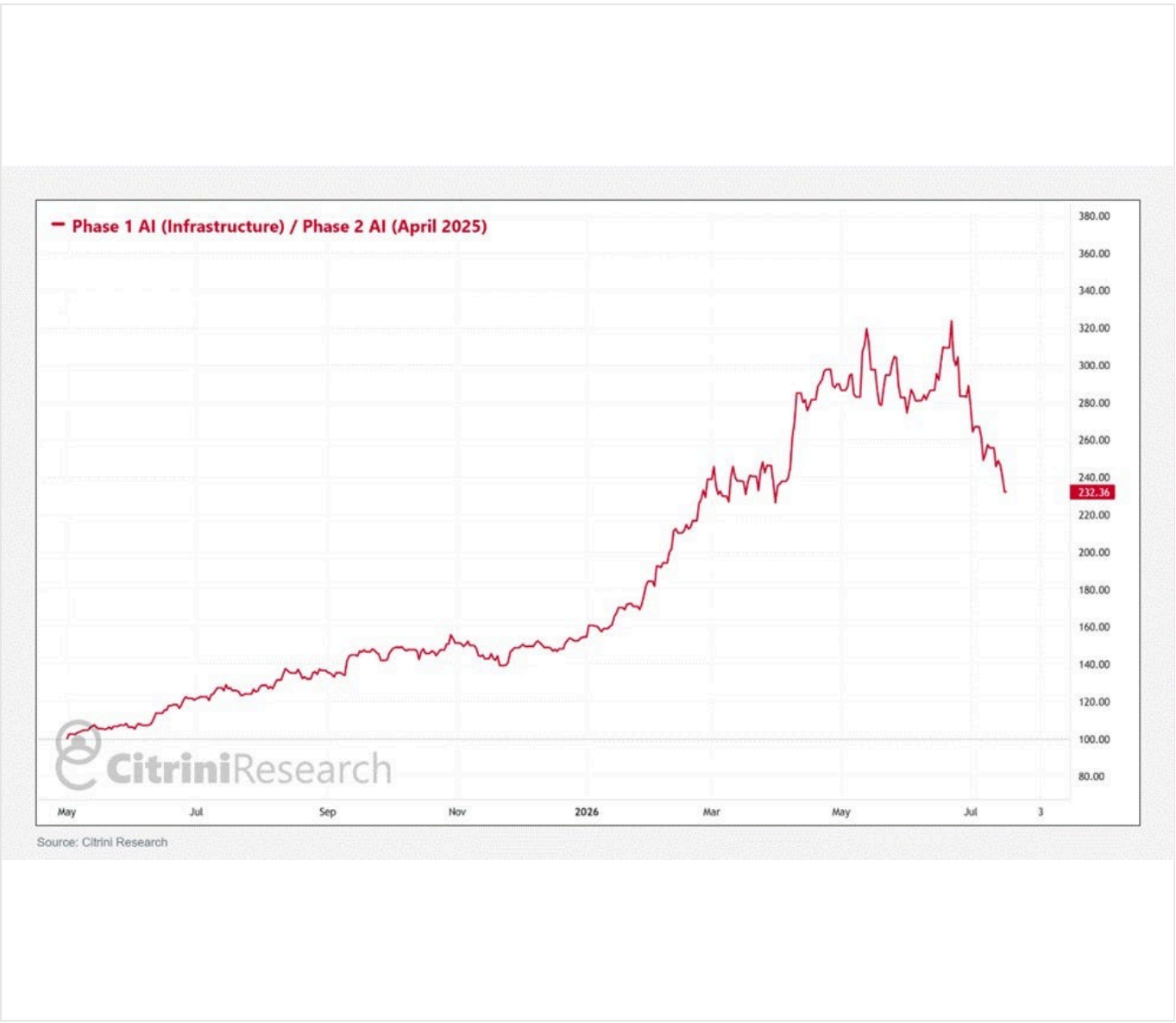
> *As we transition to Phase 2, these traditional moats face potential commoditization or even erosion… This shift is gradually going to force investors to rethink how they view AI upside. The question is no longer simply "who has the best/biggest model?" but rather "who can create the most effective systems around these models?," "where do my unit economics actually work?," and "who's actually going to benefit from using AI?".*

Or, how is value going to accrue between different layers; foundational (infra), models, and applications.

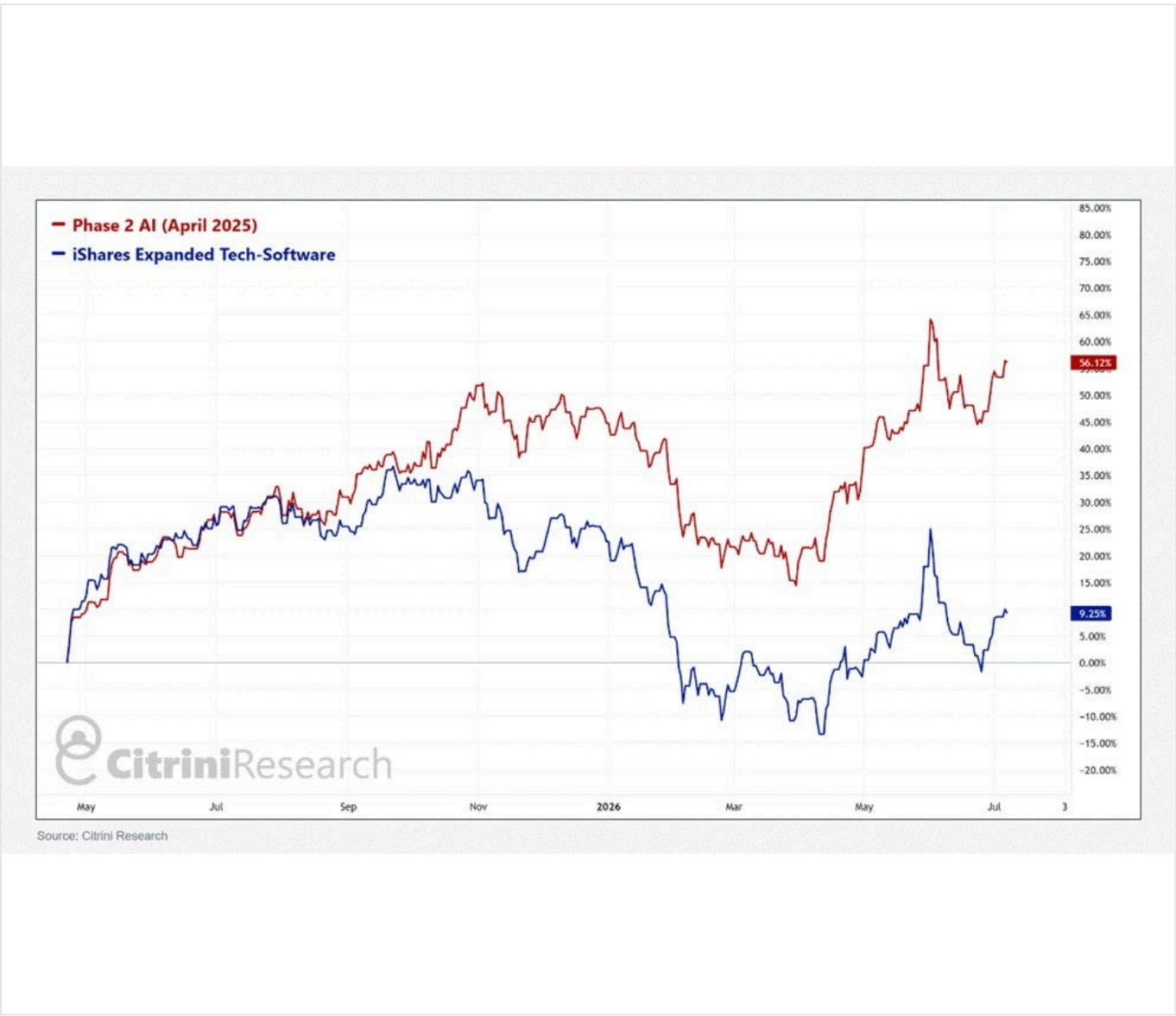


And without a clear answer, we cast a fairly wide net in our revised “Phase 2” basket, following four pillars of differentiation: ***Data, Design, Deployment, Distribution.*** The group primarily consisted of application and infrastructure software names, along with “data” businesses like FICO and FactSet.

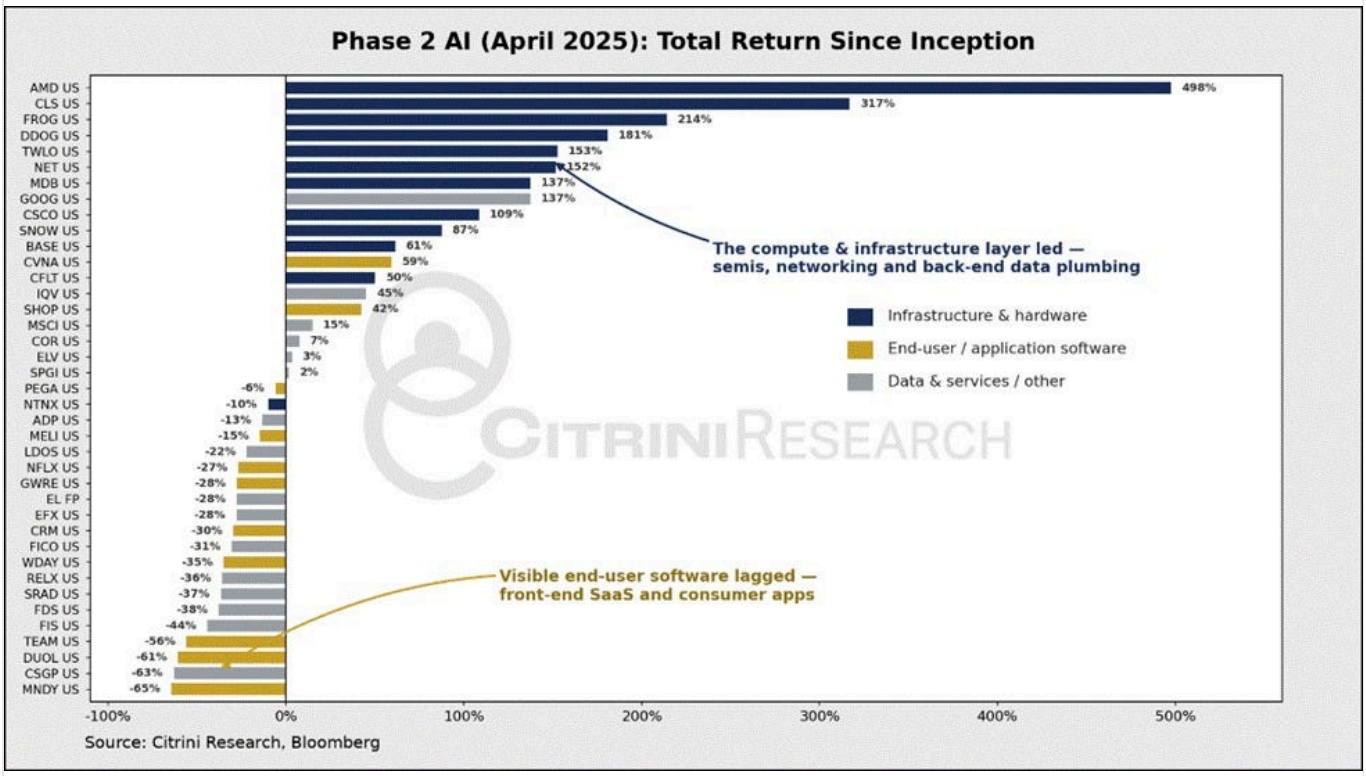
With hindsight, the smartest thing would have been to sit on the beach in the infrastructure trade. But the exercise has still proven useful, and the past several weeks suggest at least some sort of rebalancing of expectations.



Notably, this software-heavy Phase 2 basket significantly outperformed the broader industry through the SaaSpocalypse (a comparison vs. IGV ETF is below) and beat the SPX, meaning we were probably onto *something*.

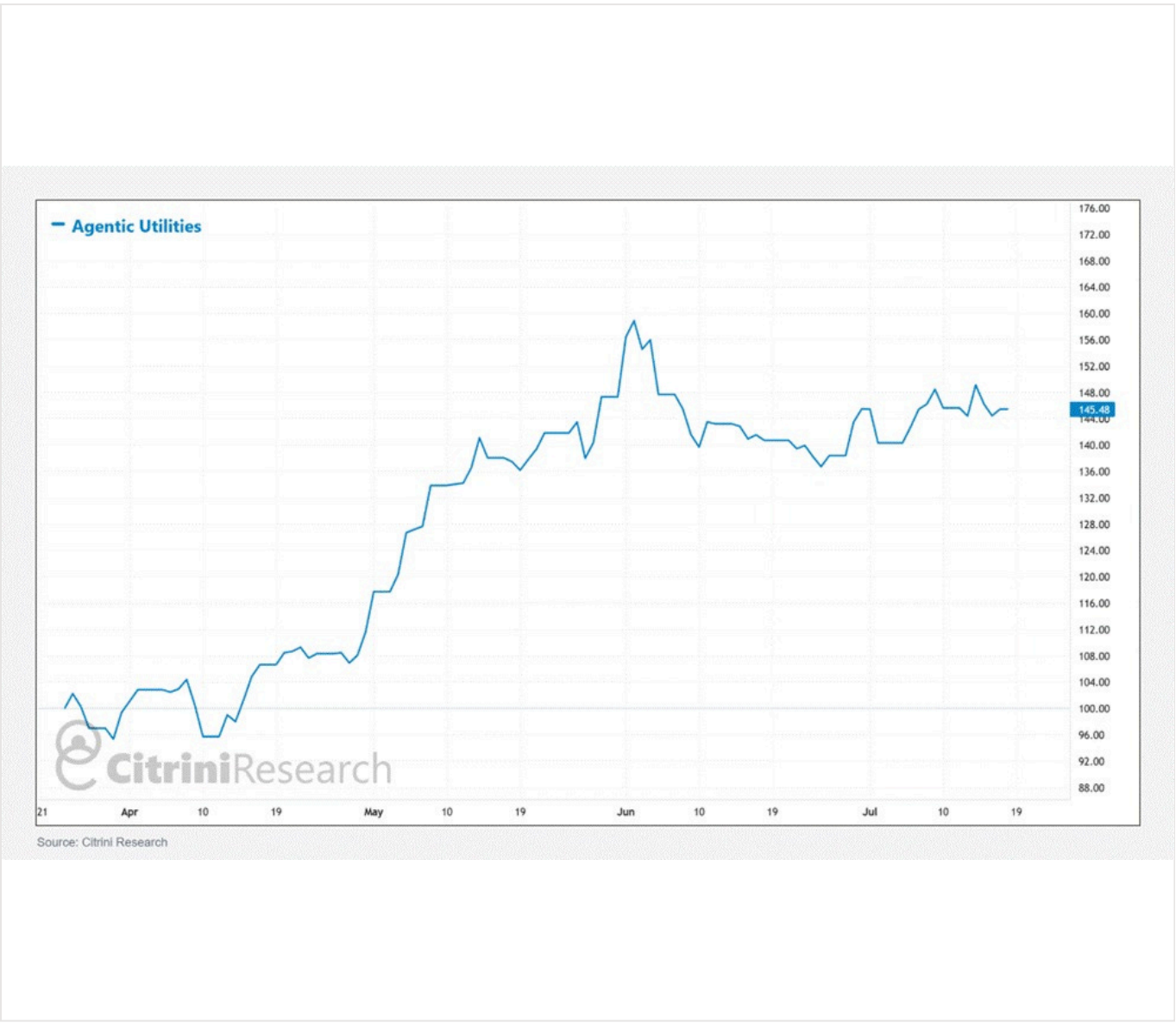


But under the hood is the real story. Simply — infrastructure software has ripped, while application software and data businesses have gotten clobbered. Far more than any idiosyncratic factor (or valuation), performance has been driven by categorical classification, with very few exceptions.



In other words, the market has been willing to underwrite some value accrual beyond the data center, but only so far as “hard software” — the FROGs and DDOGs. This makes sense — in the same way that NVDA is largely agnostic to the lab winners, infrastructure software is largely agnostic to the application winners.

Our **Agentic Utilities** piece, published in March, essentially argued this point. We ***know*** that AI is going to massively increase and reshape traffic, and so the “backbone” players are going to do well.

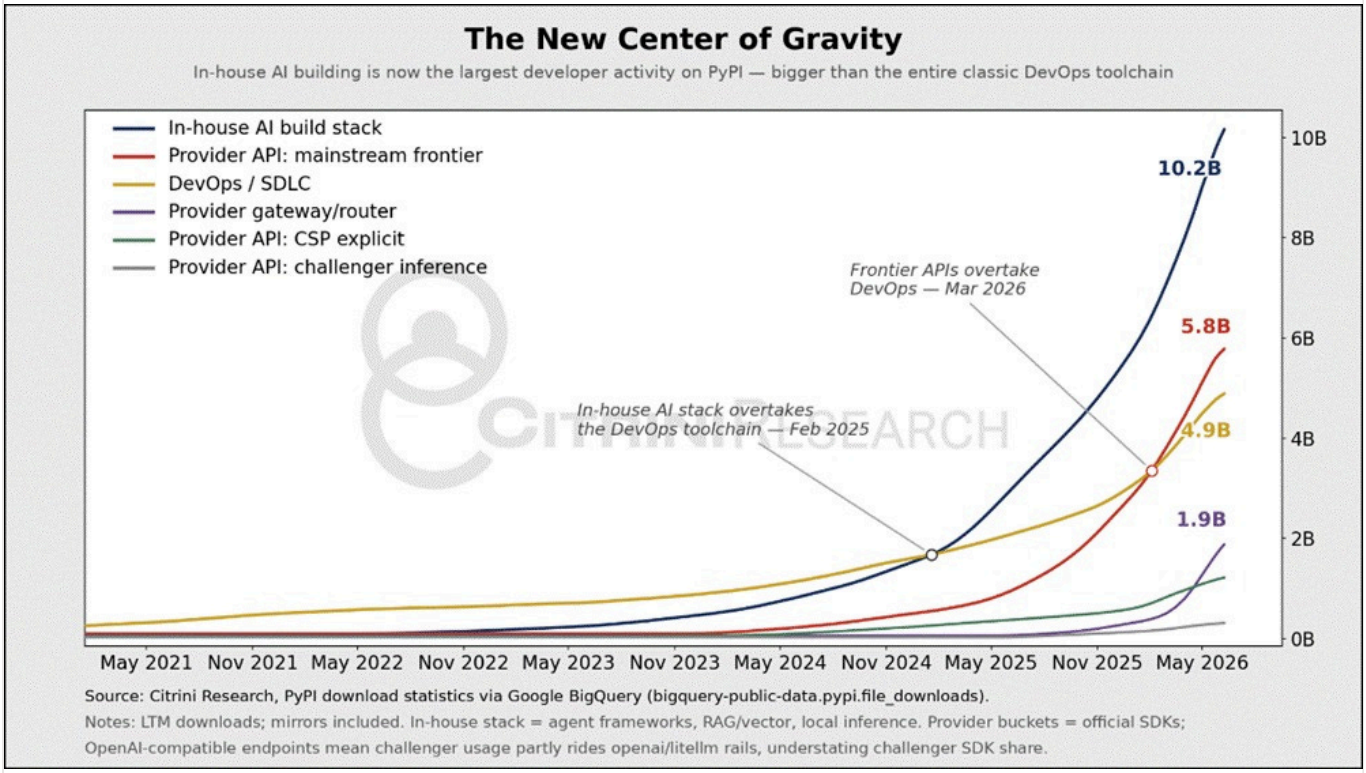


Now, there's certainly an argument that the black sheep of the industry — ***application software*** — could be on the verge of outperformance, either as a relative-value reversal, or on a more fundamental narrative shift around distribution advantage, internal cost efficiencies, and AI integration.

But judging risk on balance, we think the best approach is not a blanket "buy software because it's cheap" call. After all, there are still going to be some big speedbumps ahead. With the wide-spread explosion of agentic coding really only occurring in the past 6 months, we could very well see a single enterprise software miss send the entire cohort tumbling further. Instead, we continue to favor names that have direct AI-driven tailwinds, rather than companies whose existential threats may be "less bad" than the market expects.

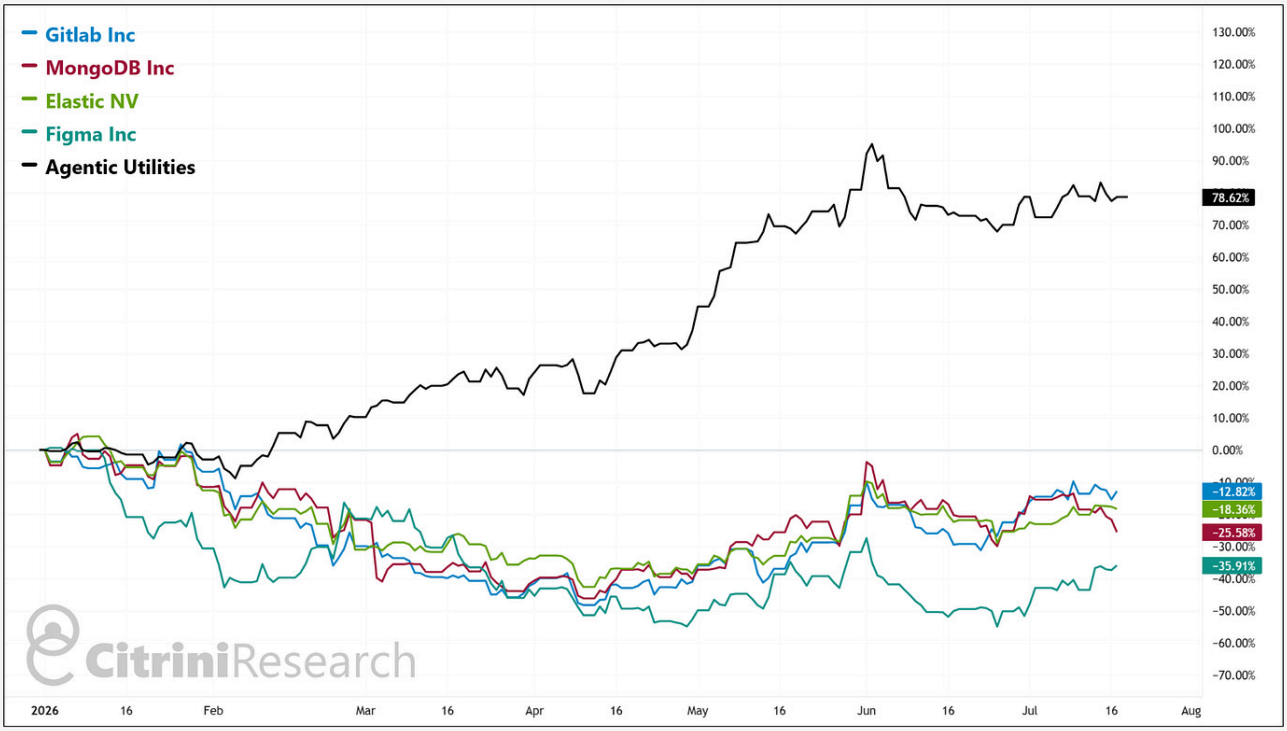
Here's one obvious trend. As AI allows more people to build software faster, a whole lot more software is going to get built.

The new center of gravity in the developer world has shifted toward in-house AI building tools, which have already eclipsed the entire classic DevOps toolchain in activity on PyPI.



So far, many traditional developer tools have fallen into the “loser bucket”, as it’s been assumed that human crutches disappear in the agentic era.

But do more builders mean more toolkits?



Source: Citrini Research

Software Repositories (GTLB US)

GitLab Public Software Comparables

Company / Positioning			Market Setup			Growth & Profitability			Valuation / Dilution					Analyst Notes	
Ticker	Comp Tier	Area	Last Price ($)	1M (%)	YTD (%)	EV ($mm)	Est. Comp Sales Growth (%)	3Y Sales CAGR (%)	FCF Margin (%)	EV/S (x)	EV/EBITDA (x)	FCF Yield (%)	5Y Avg Basic Share CAGR (%)	Valuation Z-Score (5Y)	Analyst Notes
GTLB	Anchor	DevOps platform	$33.31	19.9%	(11.2%)	4,332	24.9%	31.1%	23.2%	4.3x	29.5x	4.7%	27.1%	(0.52)	**Peer-level growth + FCF at a valuation discount; dilution and seat-to-consumption pricing shift drive debates.**
FROG	Direct workflow / DevOps	DevOps / software supply chain	$91.01	17.1%	45.7%	10,300	25.0%	23.8%	26.8%	18.3x	80.0x	1.4%	20.1%	(0.42)	Closest public DevOps comp with greater usage and consumption exposure.
TEAM	Direct workflow / DevOps	Collaboration / DevOps	$94.48	6.7%	(41.7%)	24,120	24.7%	23.0%	27.1%	3.9x	11.4x	4.9%	1.3%	(1.25)	Useful seat-based workflow benchmark, but broader collaboration exposure reduces purity.
DDOG	Infrastructure adjacency	Observability	$256.26	11.5%	88.4%	87,705	29.5%	26.9%	26.7%	23.8x	77.1x	1.1%	2.9%	(0.20)	Premium usage-based infrastructure benchmark with strong workload expansion.
DT	Infrastructure adjacency	Observability	$44.04	8.1%	1.6%	11,774	18.8%	20.3%	26.2%	5.9x	16.3x	4.0%	1.4%	(1.13)	Lower-growth observability benchmark with hybrid subscription and consumption economics.
MDB	Infrastructure adjacency	Database	$339.27	(1.0%)	(19.2%)	24,926	23.6%	24.3%	20.3%	9.5x	38.6x	2.2%	6.6%	(0.16)	Consumption-oriented data platform benchmark; useful for workload-driven expansion.
ESTC	Infrastructure adjacency	Search / observability	$62.13	2.9%	(17.6%)	5,685	17.3%	17.6%	18.5%	3.3x	15.1x	4.9%	3.8%	(0.05)	Search and context-infrastructure adjacency with cloud consumption exposure.
PANW	Premium platform benchmark	Security platform	$326.21	16.7%	77.1%	260,899	19.5%	18.8%	37.6%	24.7x	57.9x	1.7%	2.6%	3.13	Broader platform benchmark; premium multiple reflects scale and consolidation economics.
CRWD	Premium platform benchmark	Security platform	$186.60	9.3%	59.2%	186,490	23.2%	29.0%	25.8%	36.7x	99.4x	0.8%	2.8%	0.20	High-quality module-expansion benchmark rather than a direct DevOps comp.
WDAY	Mature enterprise benchmark	Enterprise applications	$143.85	10.0%	(33.0%)	34,974	13.3%	15.4%	29.1%	3.5x	9.3x	7.9%	2.3%	(1.49)	Mature seat-and-module benchmark that helps frame cash generation and lower growth.
All Peer Median (ex-GTLB)				**9.3%**	**1.6%**	**24,926**	**23.2%**	**23.0%**	**26.7%**	**9.5x**	**38.6x**	**2.2%**	**2.8%**	**(0.20)**	
Direct Peer Median (FROG, TEAM)				**11.9%**	**2.0%**	**17,210**	**24.9%**	**23.4%**	**26.9%**	**11.1x**	**45.7x**	**3.1%**	**10.7%**	**(0.83)**	

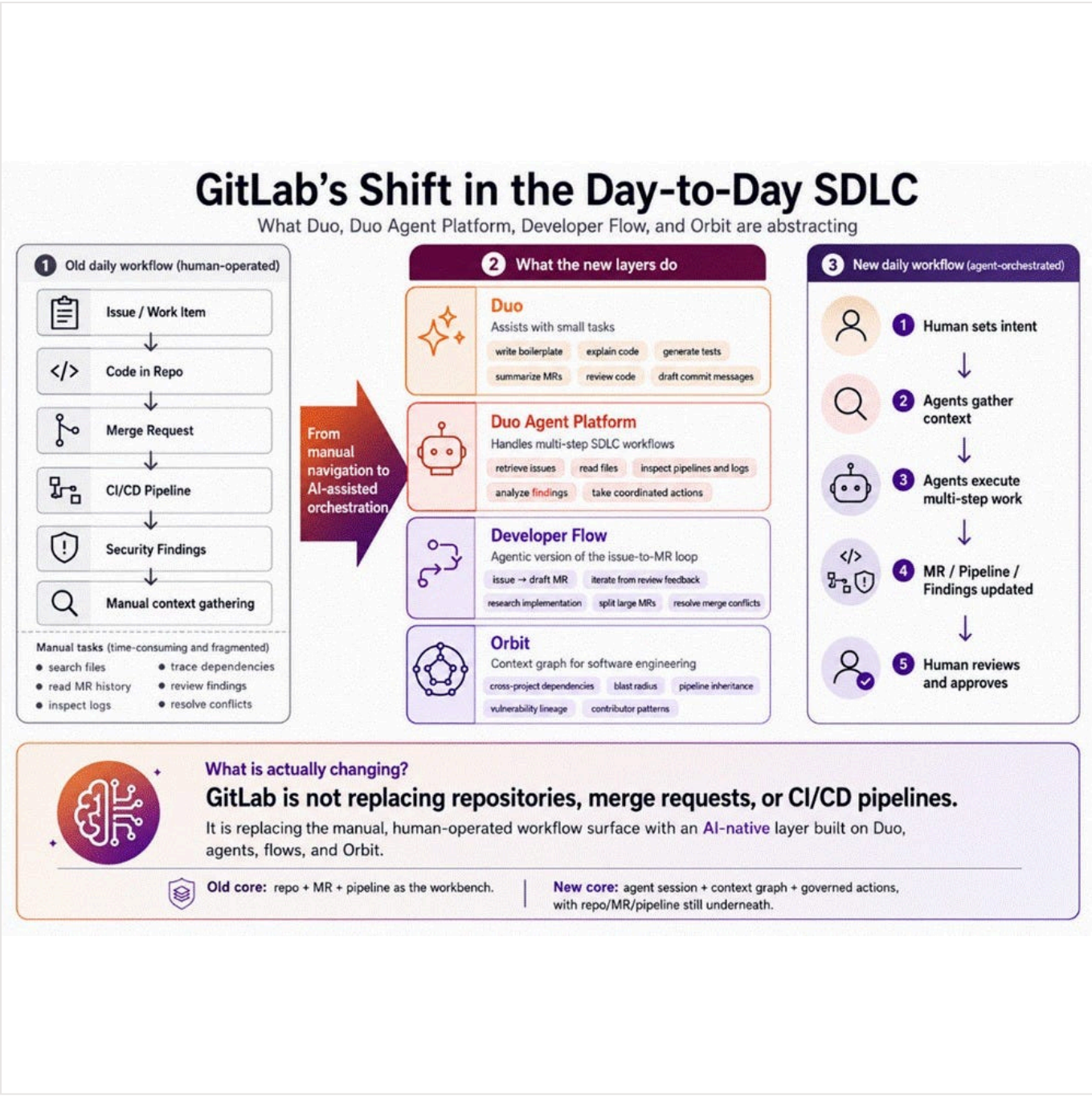
Software repositories began as systems of record: a durable history of what changed, who changed it, and how the project could return to an earlier state. AI is turning that record into a live operating environment. As more software work is delegated to agents, the valuable layer increasingly becomes the one that can map how code, tests, permissions, incidents, and dependencies relate and make that context available at runtime. **GitLab's (GTLB US)** opportunity is to move from preserving the development record to governing the work performed against it.

First: what is Git? Why does it need 'hubs' and 'labs'?

Simply put, Git is a ledger for tracking and attributing progress and changes in a software development project. GitHub (owned by Microsoft since 2018) is the world's code clearing house, allowing developers to share code and notes freely, but note how the essence of the platform is about sharing and distributed version control rather than driving improvement through iteration. GitHub does have paid features, but Microsoft certainly derives more value from the digital 'real estate' as a System-of-Record rather than from the underlying cash flows the asset produces.

GitLab, on the other hand, uses the open-source Git as a substrate for supplying developers with tools they can actually use to make better software, faster. Take GitLab Orbit, for example. Where GitLab previously was a workbench for scouring code to push urgent patches, it has taken this access and built a 360° Mission Control center.

If the DevOps of five years ago was *"what is live, and where is the leak?"*, the DevOps of the agentic era is *"how does everything here relate — and what breaks if this changes?"*



The importance of codebase intelligence becomes magnified as apps and user experiences are increasingly compiled just-in-time and just-for-you. GitLab's recently unveiled Orbit platform answers questions like…

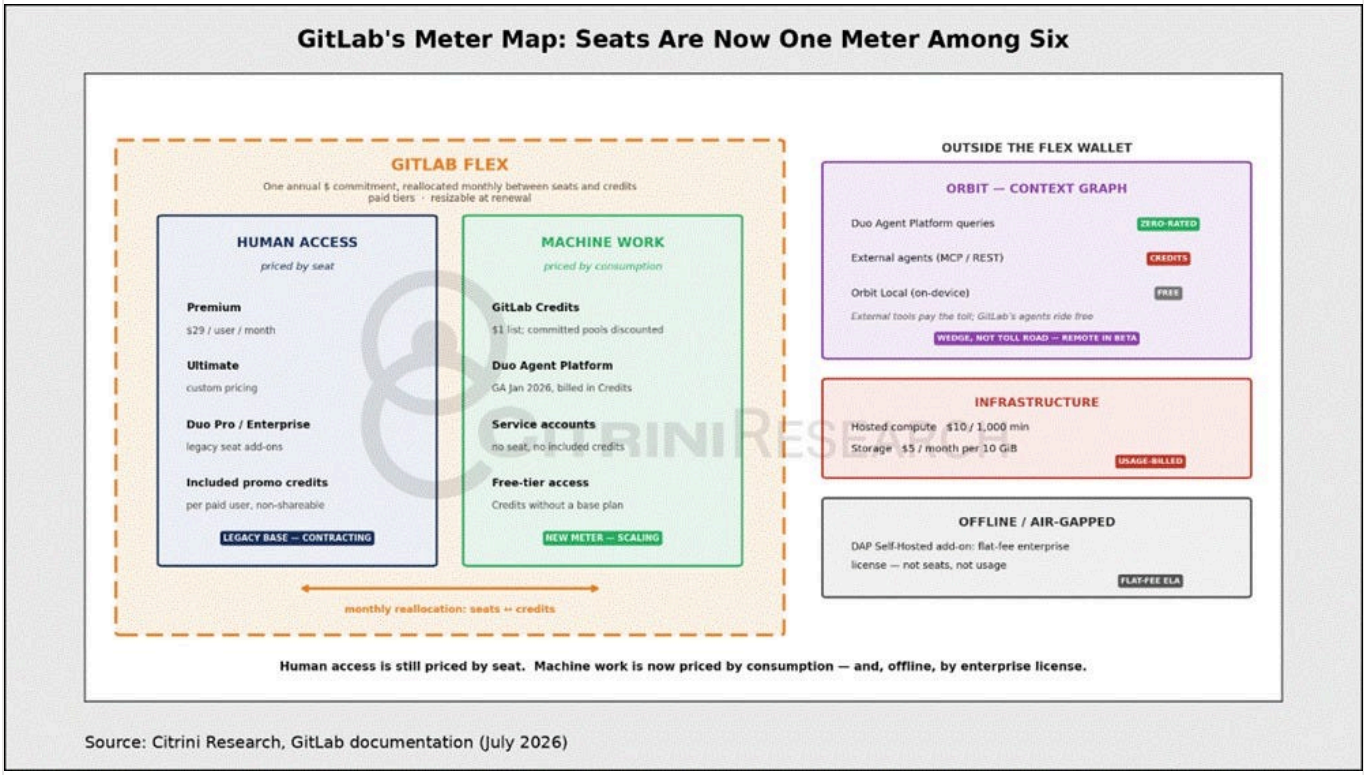
"How does X relate to Y in the context of a broad codebase?"

"Are these tests aligned with real customer behaviors?"

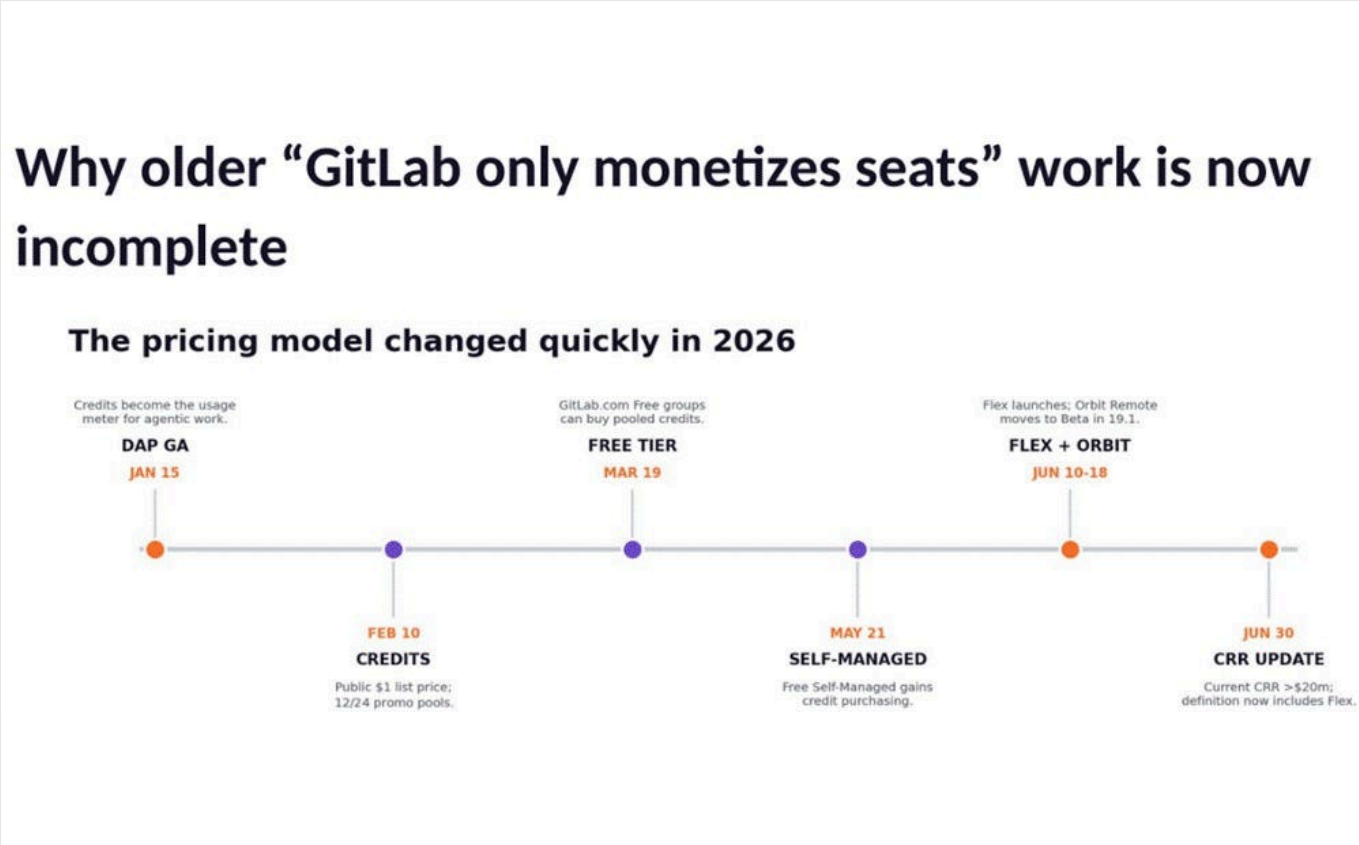
"How dependent is this app on the user having appropriate vector-math hardware?"

While this may seem like a "nice to have" for human developers that historically know their code inside-and-out, even the most advanced agents will have to reconstruct context at the start of a session for the foreseeable future. Even when they appear to have truly long-lived context at some point in the quickly approaching future, this will still be based on codified, formulaic relationships. Based on graphs.

GitLab appears uniquely situated to benefit from the proliferation of open-source and ‘generic’ AI models as the barriers to entry for their customer base compared to JFrog, for example, are much lower. Fortune 500-focused JFrog has benefited from enterprise-scale engineering budgeting strength and the improved economics of consumption-heavy pricing models, two areas the market has fairly penalized GitLab for being weak in. While we already touched on the factors that drive the market for DevOps tooling well past the Fortune 500, the shift to more value-based pricing should brighten the mood as well.



We believe that the GTLB team is aware of its stock’s performance as well as Street commentary surrounding the pricing model that has weighed on it, so we find it hard to believe this has been an error of omission or oversight. Rather, as we said in Agentic Utilities, the tools and economics that allow for this level of billing control have hardly existed until now. As such, GitLab is finally implementing new payment options as capabilities take form.



GitLab Flex is the commercial bridge between those meters. A paid customer makes one annual dollar commitment and can change the monthly allocation between seats and credit-based capabilities without a new contract amendment.

Orbit extends the same logic to the context layer, but its economics should be described precisely. External agents that query Orbit Remote through MCP or REST consume GitLab Credits, giving GitLab a potential toll on Claude Code, Codex, Cursor, and custom tools operating against the GitLab graph.

However, queries made by GitLab’s own Duo Agent Platform are currently zero-rated and Orbit Local is free. Orbit Remote remains in beta and is not yet production-ready, while GitLab has not published a per-query credit multiplier. Today, Orbit is therefore both a future direct API meter and an indirect way to improve DAP quality, speed, and token efficiency. Despite this early stage, we see broadly increasing billable surface area as the Orbit platform ramps to critical mass, enabled by improvements in payment rails like what Cloudflare described in their blog on July 1 of this year:

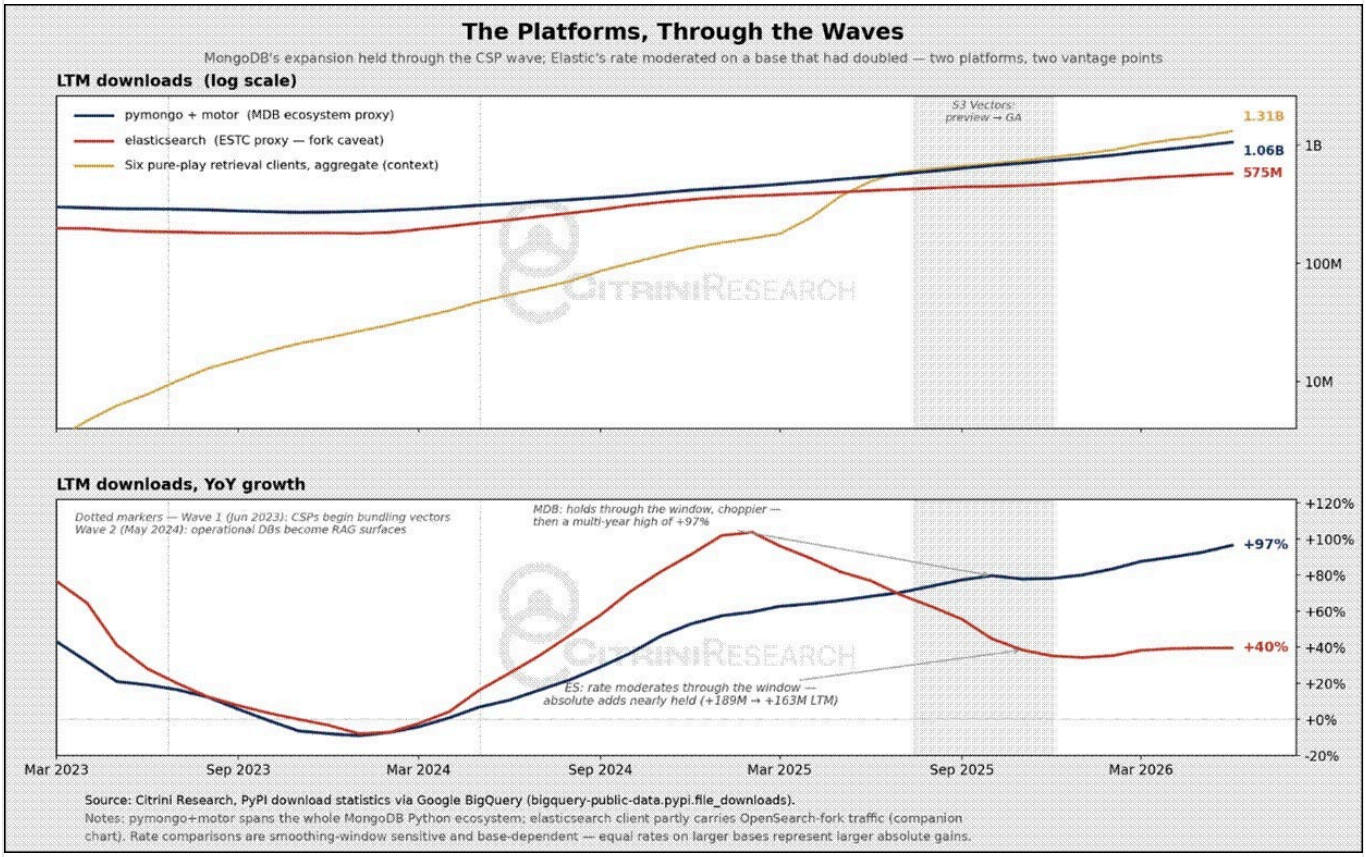
> *“This reality demands a new model: usage-based pricing for everything \[...\] The natural unit of payment for software is the request, the token, or the outcome, not the seat or the month.”*

DBaaS

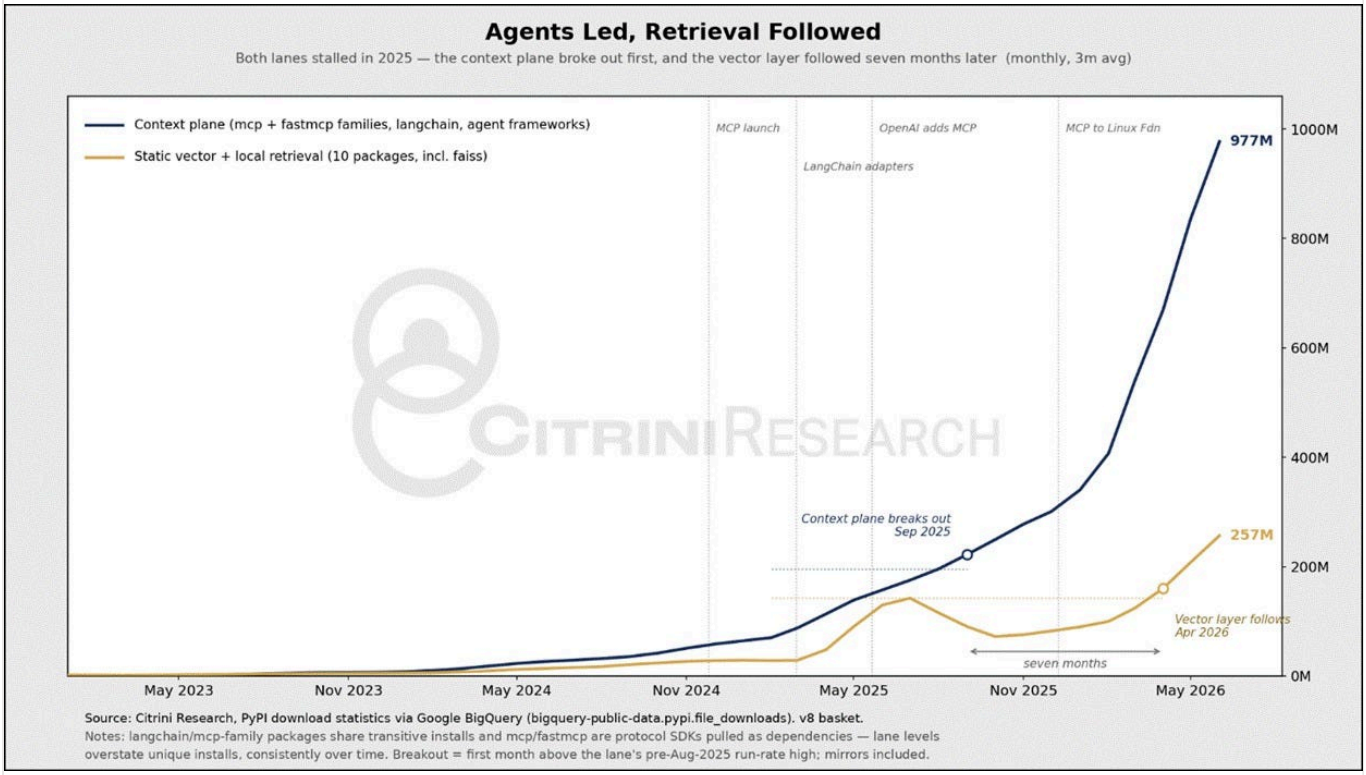
Meanwhile… while the debate over software has shifted from ‘good’ vs. ‘bad’ to ‘infrastructure’ vs. ‘application,’ one group within the ‘good x infrastructure’ camp has conspicuously not been invited to the party: DBaaS (database-as-a-service) providers like **MongoDB (MDB US)** and **Elastic (ESTC US)**.

The irony in this sector: vector stores are critical to AI — LLMs don't work without embeddings, and agents don't work well without a way to reference them — but the labs knew it, the market knew it, and the CSPs (cloud service providers) certainly knew it. AWS, Oracle, Google, and Microsoft all bundled 'good enough' vector search into the compute products enterprises were already buying, and even the neoclouds that never built one simply point customers to open-source Milvus on their clusters or a Pinecone hookup. If the AI-facing feature is a bundled checkbox, the standalone database vendor loses.

Bundling did real damage to the narrative and it landed at an awkward moment: right as retrieval augmented generation (RAG) went from a context-building technique to a fundamental part of agentic systems. Some in AI circles proclaimed we had hit "peak RAG".



But… tool calls, goal seeking, image *re*-generation — all of it relies on what look an awful lot like RAG pipelines.



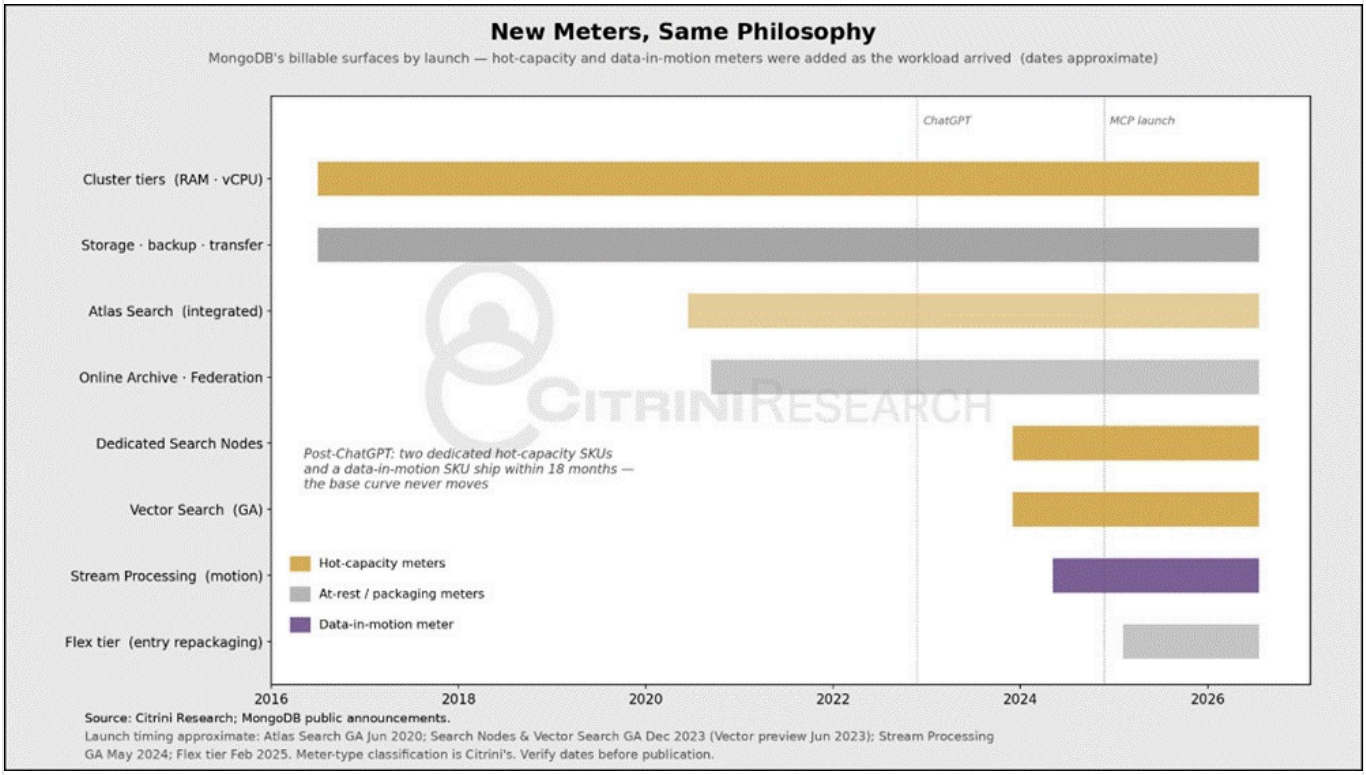
MongoDB: Heating Up the Working Set

Early RAG systems placed large, static document collections into retrieval pipelines. Information was then chunked, embedded, stored, searched, and inserted into a prompt. Of course, while database activity mattered, retrieval often occurred once or a few times per model interaction.

As we've learned, agentic systems behave differently. A single task generates repeated reads and writes across session state, application data, tool results, permissions, events, intermediate outputs, and prior actions. The payloads are smaller, but the interactions are constant – the working set is continuously changing. Retrieval stops being a preprocessing step and becomes part of the execution loop itself.

This is the MDB thesis: Atlas does not monetize inexpensive bytes sitting at rest. Its valuable meters are the memory, CPU, throughput, and availability required to keep operational data hot and immediately queryable. Vector search is one access method bolted onto an operational database whose economics improve every time an application touches its own state — and agents touch state compulsively.

Critically, Atlas list pricing hasn't moved since at least 2020. What MongoDB has done in the post-ChatGPT window is attach new hot-capacity meters — dedicated search nodes, vector search, stream processing — to the same unchanged base curve. The toll booth was already built. The AI cycle's only job is to send traffic through it.



Atlas growth re-accelerated from roughly 26% to over 29% through fiscal 2026, crossing a $2 billion annualized run rate. The quarter ending April 2026 showed Atlas up 29% with guidance raised for the full-year. Consumption pricing means the transmission from workload to revenue is automatic — no repricing negotiation, no packaging decision, no sales cycle. MDB's meters have been running for a decade.

Two risks stand out:

First, the marginal new application is increasingly written by a coding agent, and coding agents default to Postgres — Supabase, Neon, pgvector — because that's what saturates their training data. MDB's agentic thesis is therefore a thesis about the installed enterprise estate, where agents amplify interaction with systems that already exist, not about winning the next million vibe-coded apps.

Second, the modern form of the bundling risk — the scariest competitor for the context workload is not a CSP database clone. It's the agent platforms internalizing memory. If session state, tool results, and agent memory end up living inside the hyperscalers' agent primitives or the labs' own memory APIs, the database never sees the incremental workload at all.

Elastic: The Index Is a Copy, Not a Pointer

If MongoDB monetizes an application interacting with its own state, **Elastic (ESTC US)** monetizes an organization finding anything at all. But the common shorthand "databases of databases" is too simple, too.

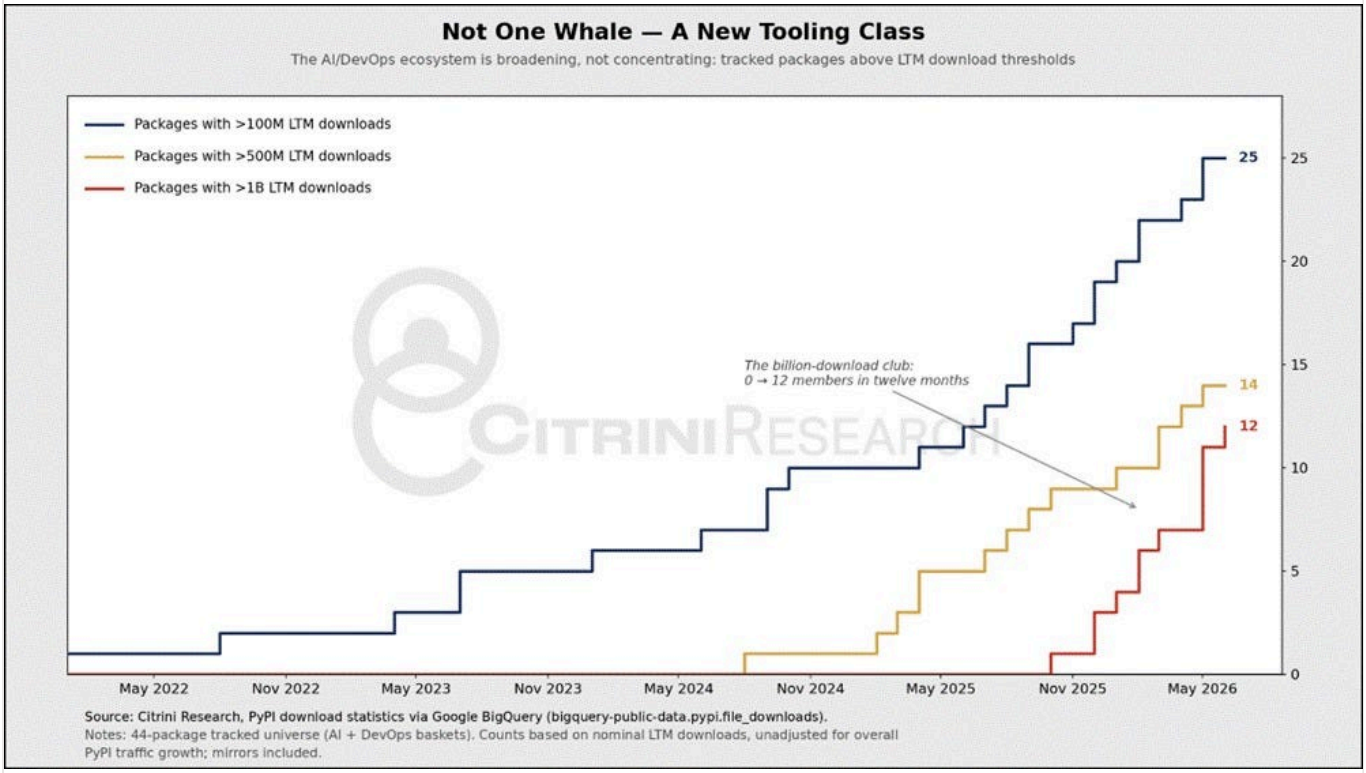
Elastic does not search other systems where they live. It ingests them — connectors, pipelines, permission mappings — into its own indices, and sells you the queryable copy. Once an enterprise has fully plumbed its repositories, logs, and document stores into one indexed, permission-aware plane, ripping it out means rebuilding every pipe.

This becomes more valuable as agents operate across a larger environment. A human developer knows which repository, dashboard, or incident log contains the answer. An agent must discover it programmatically, determine which result is current and authorized, and retrieve it fast enough to stay inside its execution loop.

The workload is broader than vector similarity: agentic retrieval combines semantic search with keywords, metadata, permissions, recency, and operational signals. The more heterogeneous the environment, the more valuable hybrid retrieval and ranking become relative to the basic vector index that came free with your compute.

The bear case is that MCP-style direct tool calling lets an agent hit Jira's API, then GitHub's, then the wiki's — with no unified index required. Hybrid search itself is commoditizing as every database aims to bolt on BM25-plus-vectors-plus-rerank.

The counter argument is: *"find the relevant thing among fifty million documents, under this user's permissions, in two hundred milliseconds"* is not a tool-calling problem, it is an index problem, and it gets harder as the environment gets bigger.



Fiscal 2026 revenue grew 17% and Q4 grew 16%, but FY27 was guided to ~14% — a decelerating revenue line. ***But,*** cRPO accelerated to +20%, total RPO to +28%, over a third of Elastic's $100,000+ customers are now using its AI capabilities, and monthly (SMB) cloud grew just 3%.

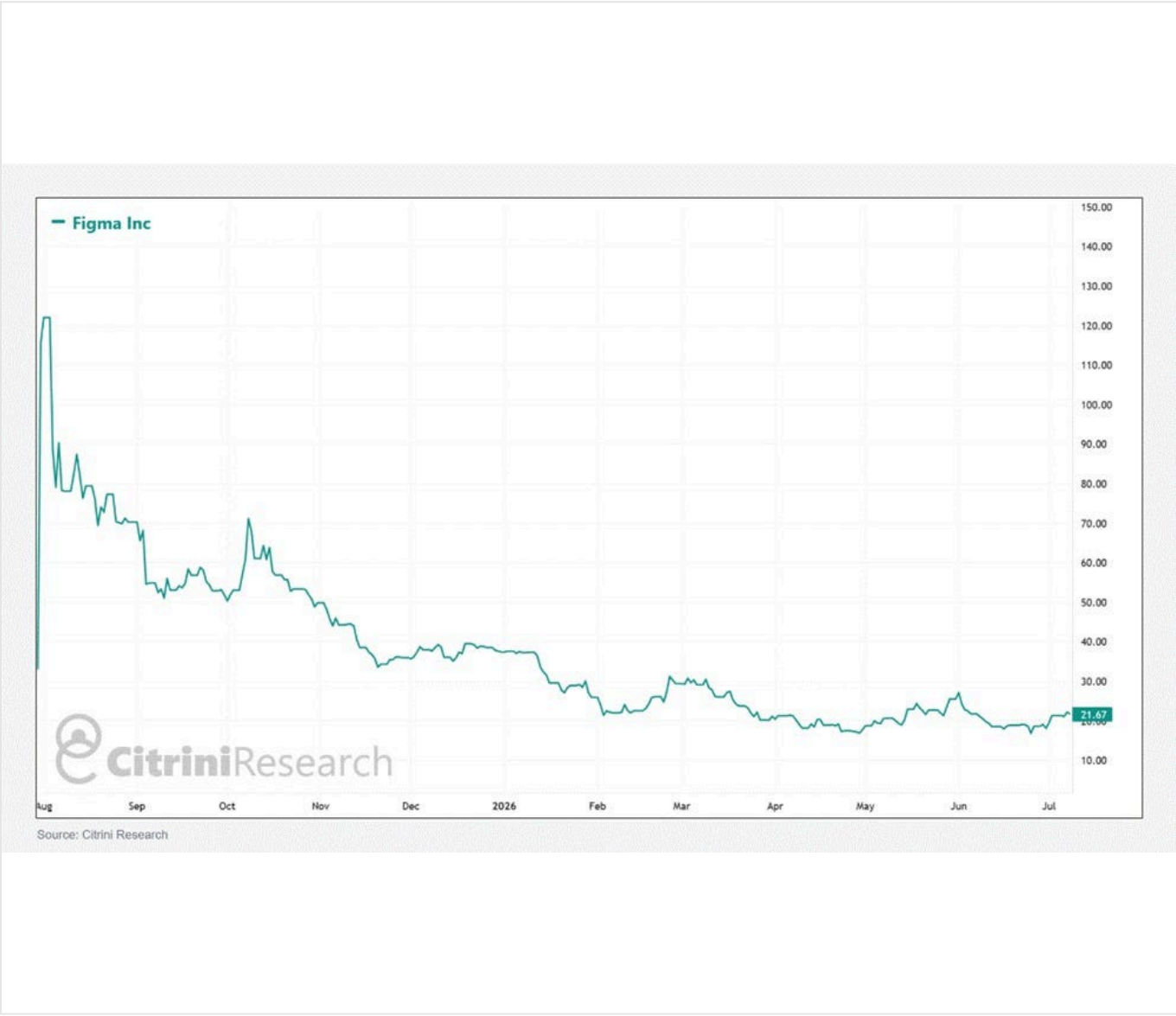
So large customers are committing faster while recognized revenue slows and the low end leaks. Either enterprises are standardizing on the search plane ahead of consumption (bullish, and exactly what the thesis predicts) or contract duration is simply stretching (cosmetic).

But unlike MDB, you are not paying for the market's agreement while you wait. ESTC is priced as ex-growth, guiding ~19% operating margins with a working buyback, so the search-plane thesis comes packaged as a cheap option rather than an embedded assumption.

The catch-up trade, then, is not that vector stores regain their novelty. It is that AI moves from occasional retrieval to continuous context maintenance, and that these three names monetize that shift through three different transmissions: GitLab must convert seats into meters (a pricing transition, still executing), MongoDB already meters the working set (a transmission that works today), and Elastic sells the queryable copy of everything else (an option on enterprise standardization).

Design

As a matter of taste, design is fundamentally subjective. What **Figma (FIG US)** promises is *artisanal AI* — vibecoded organs still need human skin. Since the coding agents hit the mainstream, its revenue growth and net dollar retention have ***accelerated*** sequentially each quarter, now reaching 46% YoY with 139% NRR. Its stock price… hasn't.



Human designers amass an internal repository of design elements, what works together, and what doesn't. Each unique case receives specific treatment, but our cumulative experience certainly informs our decision making in real-time. Professional designers, though, also need to follow strict rules such as shape ratios, color palettes, and brand guidelines — putting creativity and determinism in constant tension. What this creates, from the perspective of the designer, **is a specification that sounds a lot like an LLM chatbot context**. You start with a baseline system prompt that has concrete rules, preferred tools, and the like.

This is important to **Figma (FIG US)** because, even if this design loop reduces human involvement, the underlying framework maintains a remarkably consistent geometry. It doesn't matter if humans or machines

are aligning pixels, localizing text, or testing for color-blindness — the same rules and guidelines need to be clearly codified and enforced.

This sounds a lot like the **ask → retrieve → reason → act → update** loop that forms the backbone of effective coding agents. If anything, trends such as personalized ads, just-in-time compilation, and dynamic campaigns or regulations highlight the value of a smooth iterative process. As each ad becomes more and more personalized, the unit value of that ad increases and makes the execution of higher-touch efforts even more vital. Generating 500 variations of the same Hero image in real-time means that each one is a targeted strike that carries higher risk and higher reward.

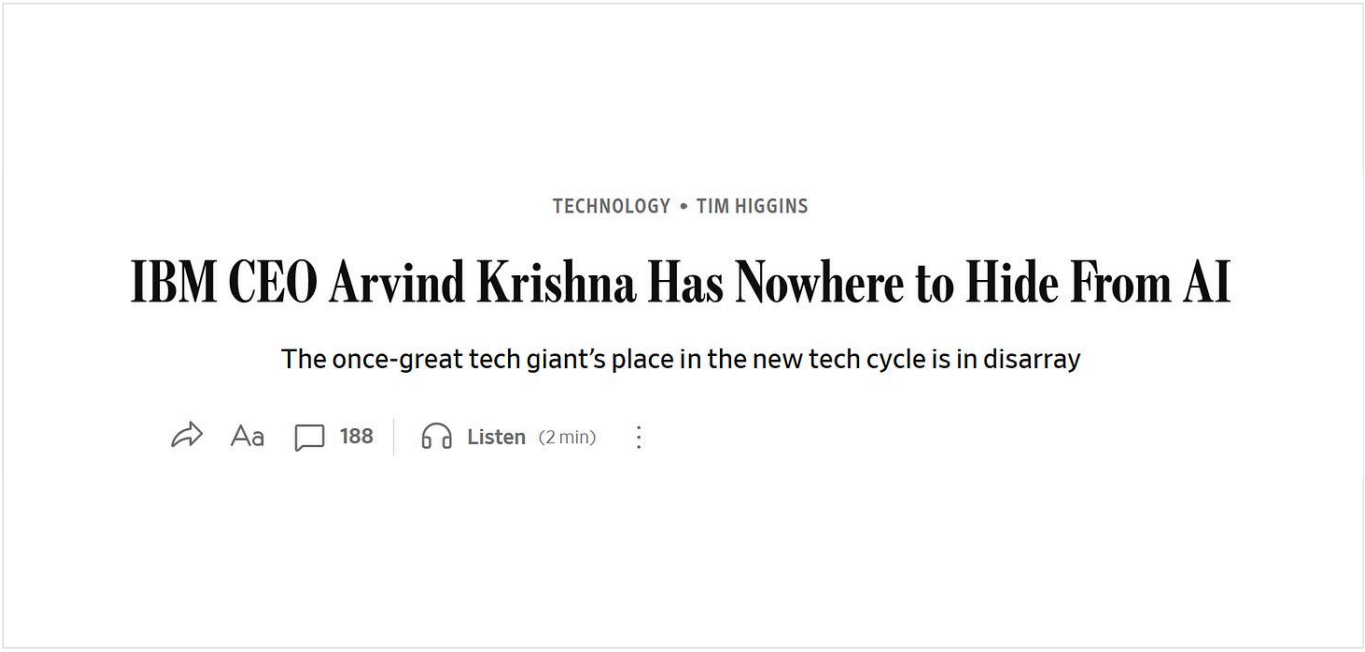
We actually don't think the specifics of these design trends matters nearly as much as the economic direction of travel. The debate shouldn't be *'man vs. machine,'* rather an axis of *'model vs. system'* seems more appropriate. Again, the echoes of agentic coding here are ringing in our ears. If there were a frontier LLM that could take an input and create a publishable product all 'under the hood,' then, yes, products such as Figma would be very challenged.

However, the AI industry has clearly realized that massive base model training can efficiently get you about 90% of the way there, while the remainder (the stuff that we pay for) needs a bit more handholding. Think: humans writing tools or harnesses for agents generating outsized gains in specific, complex verticals. This makes sense. The real world usually has a very steep learning curve, and it bodes well for the durability of hybrid human-machine systems rather than the one-click-and-done panacea that recent price action suggests LLMs have become in the realm of design.

3\. Enterprise Spend: Secure and Sovereign

-

Before the market opened on July 14, IBM pre-announced a dismal quarter, with a big miss to its mainframe hardware business and attached software.



But the much more important signal in our view is where this spending was being redirected to. Per IBM's letter:

> *"In the last few weeks of June,* ***we saw clients shift their quarterly capex spend toward servers, storage, and memory purchases*** *to secure supply-constrained infrastructure ahead of expected price increases. This dynamic impacted client buying patterns. While we anticipated some supply chain related impact in our expectations, we did not anticipate the magnitude of the capex reprioritization.* ***In addition, clients were distracted with rapidly-evolving, industry-wide cybersecurity concerns in the quarter.****"*

In a follow-on interview with CNBC's Sara Eisen, CEO Arvind Krishna shared more color:

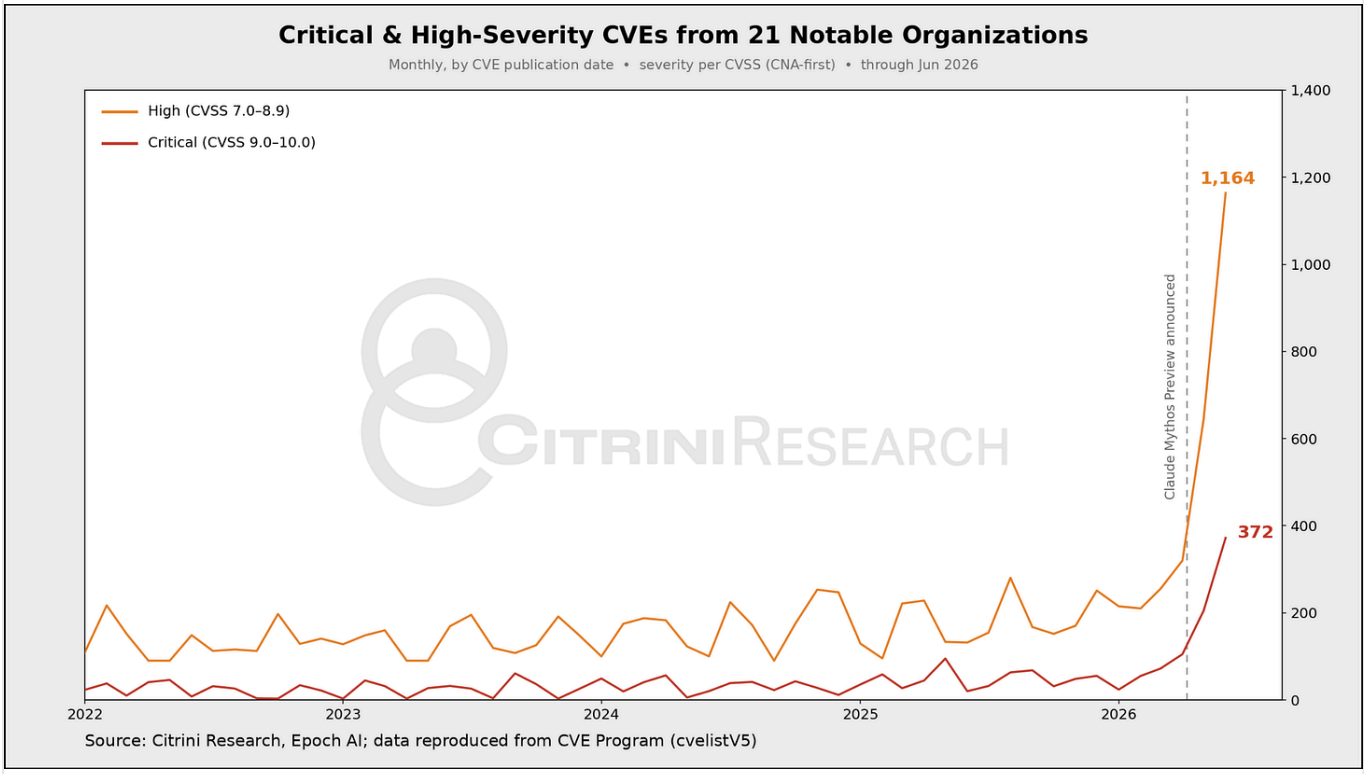
> *"****Mythos is making people pause to say, wait, how much do I need to spend on cyber?****"*

In short, enterprise capex suddenly began to shift away from mainframes (at the tail end of a very strong product cycle, to be fair) in favor of four named things: servers, storage, memory, and cybersecurity. This is some of the strongest evidence yet of two interrelated and accelerating trends in enterprise AI spending: ***security and sovereignty*****.**

Security

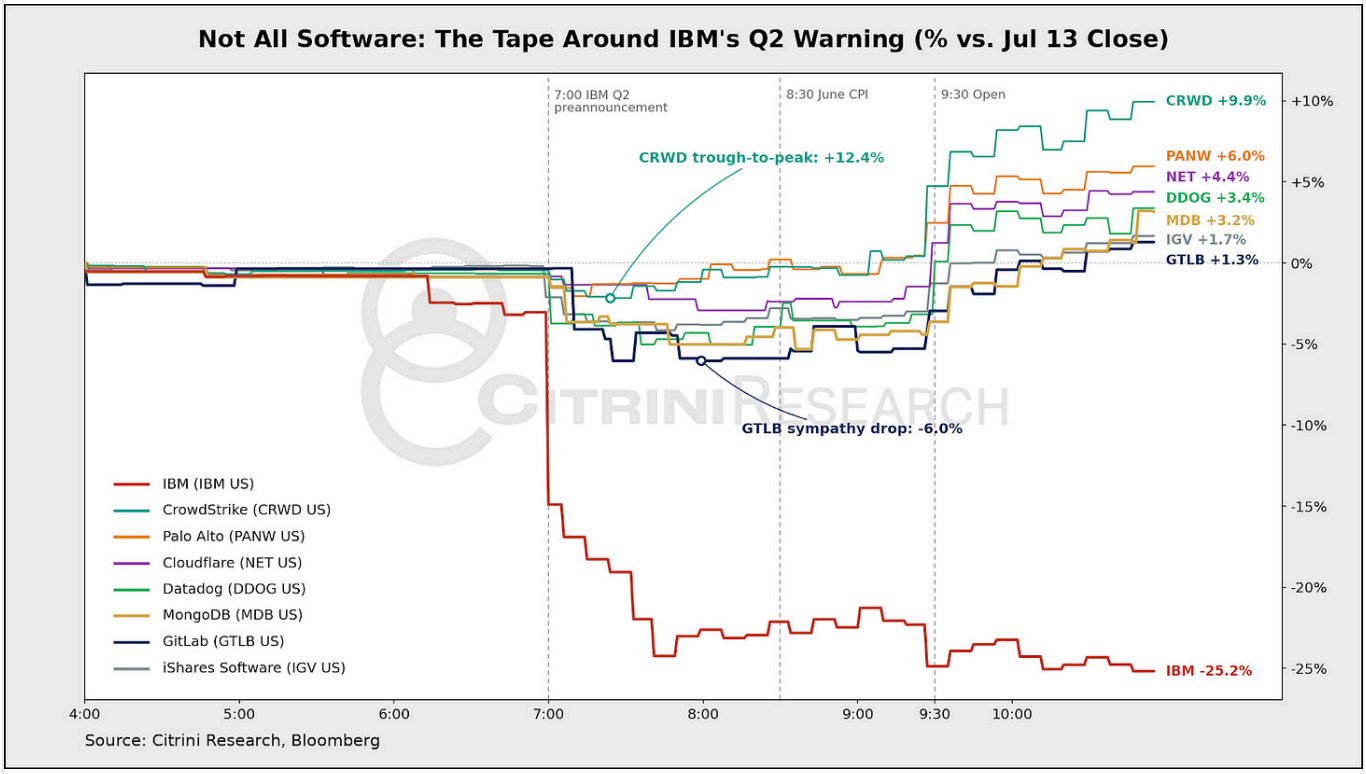
The cybersecurity angle is straightforward. In the weeks that followed Mythos Preview, thousands of zero-day vulnerabilities were unearthed, including a 27-year-old bug in OpenBSD (purportedly a *security-focused* operating system), a 16-year-old flaw in FFmpeg, and a 29-year-old memory leak in the Squid web proxy. *The release of a highly-capable and open-source Kimi3 arguably turbocharges these risks.*

To paraphrase Palo Alto Networks' product chief Lee Klarich: *there will be more attacks, faster attacks, and more sophisticated attacks.*



Mythos identifying vulnerabilities is not the same as Claude spinning up a CRM from scratch. Instead, the widening complexity and surface area of cyberattacks simply expands the use case for cybersecurity services.

As we argued in **Phase 2** and again in Agentic Utilities, no CISO is willing to risk the stability of their entire digital infrastructure on a single point-of-failure – least of all the same vendor providing the models capable of cyberattacks. There is assurance (and job security) to be found in employing sophisticated, third party cybersecurity services – these names rallied hard after the market recognized what the IBM letter actually meant.



It's worth repeating that Anthropic did not aim to circumvent the cybersecurity complex with the release of Mythos. In fact, they partnered with several cybersecurity companies within **Project Glasswing**. The businesses within Glasswing span the gamut of the cybersecurity stack: **Palo Alto Networks (PANW US)**, **Cisco (CSCO US), CrowdStrike (CRWD US)**, and **Cloudflare (NET US)** were highlighted as key cybersecurity and networking partners to assist in this endeavor. Then, in June, Anthropic widened the scope of Glasswing to include more partners from across the cybersecurity arena, including names like **Okta (OKTA US)**, **Rubrik (RBRK US)**, and **Netskope (NTSK US)**.

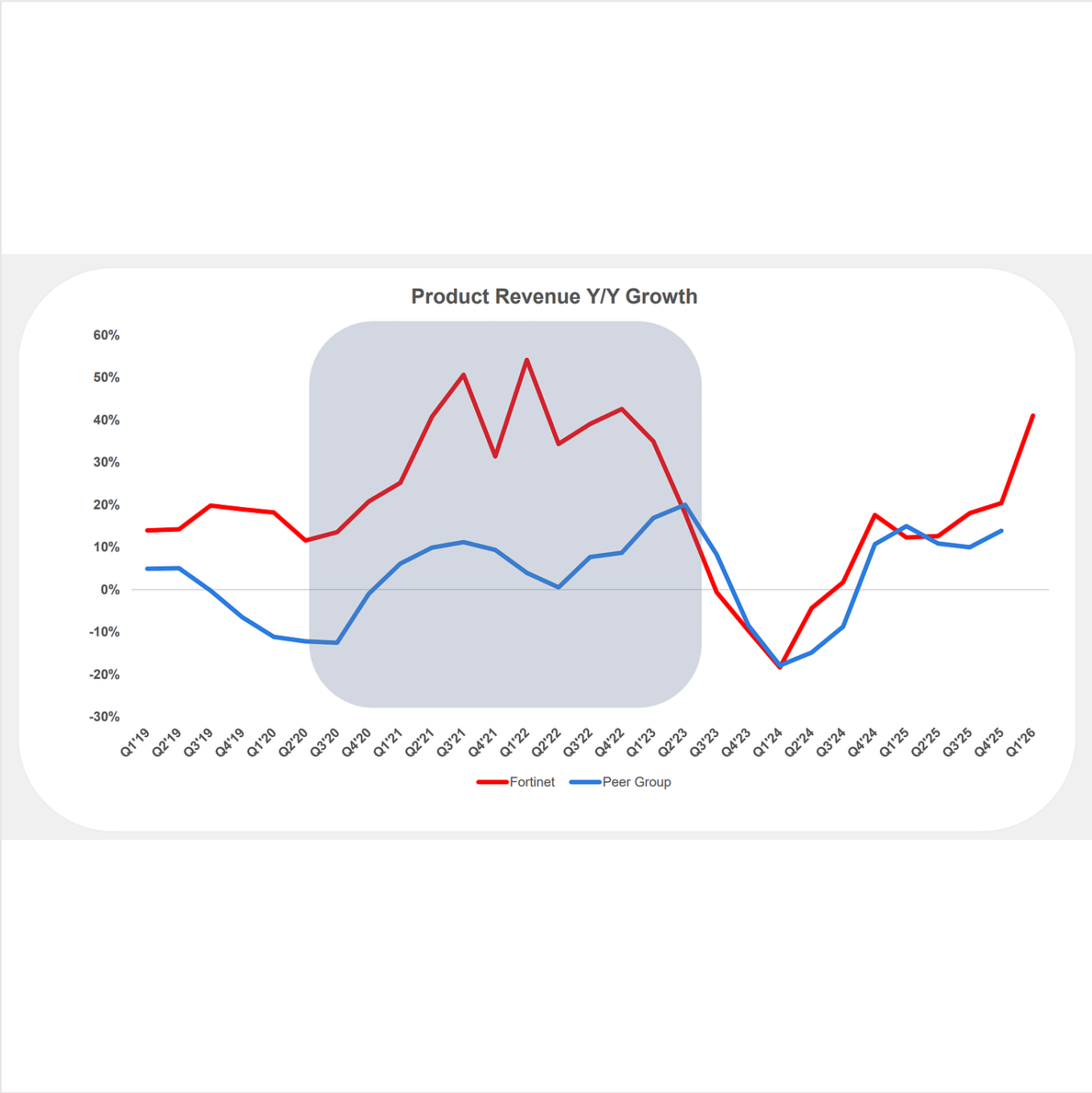
We've argued for a while that the market leading cybersecurity companies were mispriced. In January's Muscle Memory, we argued that selling cyber on agentic coding advancements was – to put it bluntly – *stupid.* But with leading cyber-software names already well discussed (in our notes and many others) and meaningfully re-rated in the past couple months, we turn our attention to some other types of beneficiaries.

Physical Cybersecurity

The prevalence of software in Project Glasswing is notable, but the *physical* side of the ledger can be shaped to reach outcomes that a stack assembled from off-the-shelf parts simply cannot replicate.

The first to discuss is firewall leader **Fortinet (FTNT US).** This is a leading hardware incumbent with software-like economics.

The company's value proposition lives within its custom silicon. Every FortiGate runs its own FortiASIC processors – the NP7 network processor, the CP9 content processor, and the newer SP5, which consolidates both functions onto a single die. The end result is a chip architecture that can do what a standard x86 CPU cannot. And as shown below, product revenue growth has sharply accelerated.



The appliance is the moat, but it is also a vehicle through which Fortinet can upsell its over 900,000 customers onto vertically integrated services like FortiGuard, SASE, and Security Operations.

Over the last decade, Fortinet's product sales have shifted from over 40% of sales to the low-to-mid 30s today. The product side is not disappearing – Fortinet reported record product sales in Q4 2025, and consensus estimates are pointing to another record of $763 million in product sales for Q4 2026. ***This is the model unfolding in practice:*** Sell the appliance, lock in your position within the hardware stack, and differentiate with vertically integrated software.

As a kicker, we think they'll play an important role in the emergent market for OT security, as management called out 70% billings growth in this segment for Q1. With their converged IT/OT architecture, we think Fortinet has a strong position to take share of the OT security market as it scales.

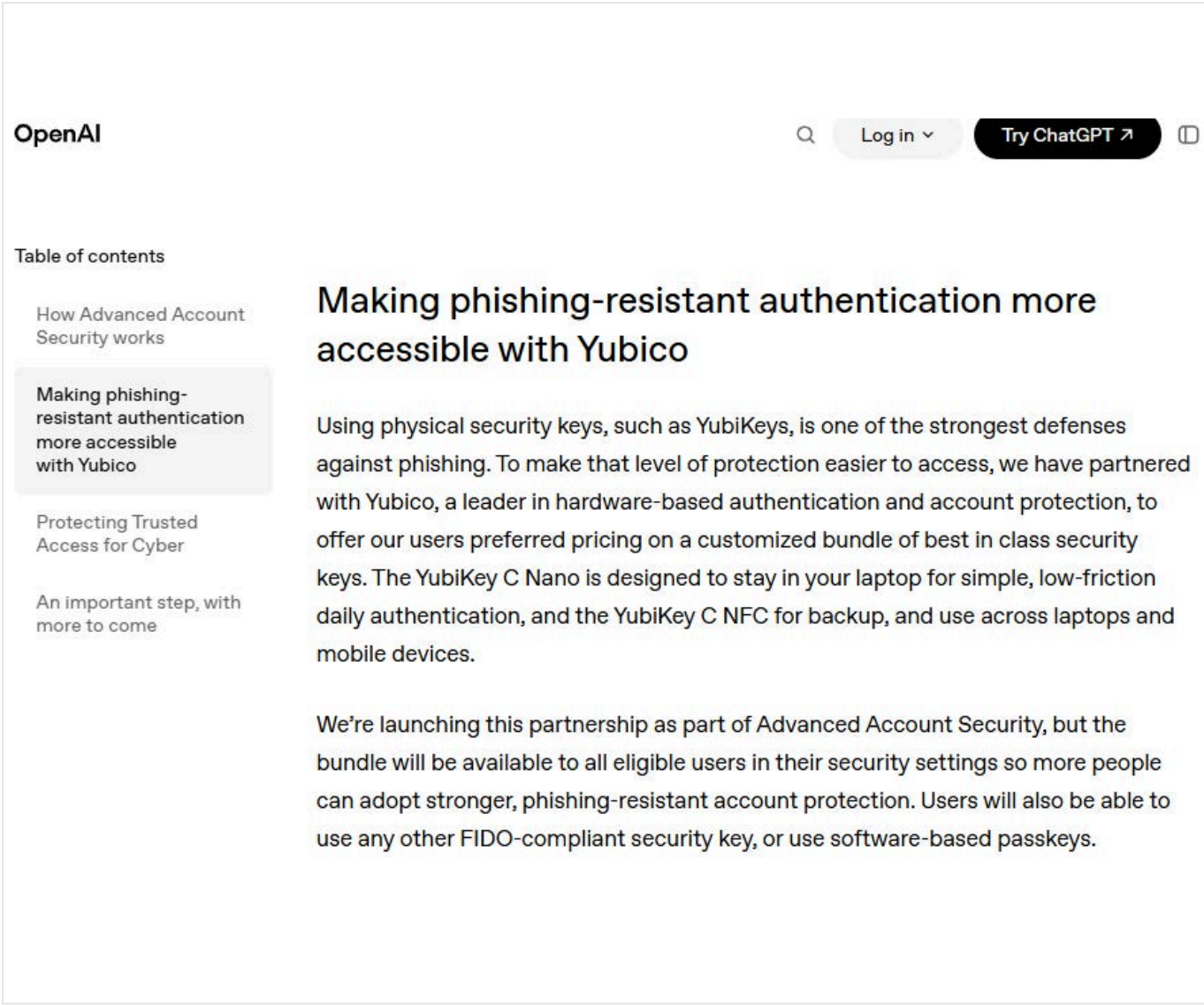
Another off-the-radar name that's caught our attention in the physical security space is **Yubico (YUBICO SS)**, a Swedish provider of hardware-based passkeys and authentication devices. Its flagship product – the **YubiKey 5** – is exactly what it sounds like: a physical "key" used for multi-factor authentication (MFA).

If it sounds a little analog, *that's because it is*. The explosion of AI-enabled threats, especially those from agentic identities, has forced the industry to actually meet the zero-trust standard. SMS codes or authenticator apps can be intercepted or phished, but there is no substitute for physical human verification.

In the agentic era, the "tap" mechanism ensures that there is a human being behind a high-consequence identity. In the company's words: *"the only thing more powerful than the AI itself is the identity of the person controlling it."*

Yubico's financial and stock performance haven't shown a good trend over the past year or so, but there is a major catalyst (other than Mythos) that could radically reshape the trajectory – the partnership with OpenAI announced in late April.

OpenAI's **Trusted Access for Cyber (TAC)** framework is a pilot program to ensure that models are deployed safely and responsibly. Those who enroll in the program are required to enable Advanced Account Security (AAS), which can be satisfied through hardware-backed keys like YubiKeys. OpenAI offers a YubiKey bundle through OpenAI subscriptions, reinforcing the partnership between the two (and perhaps trialing OpenAI's ability to compete in AI-native commerce).



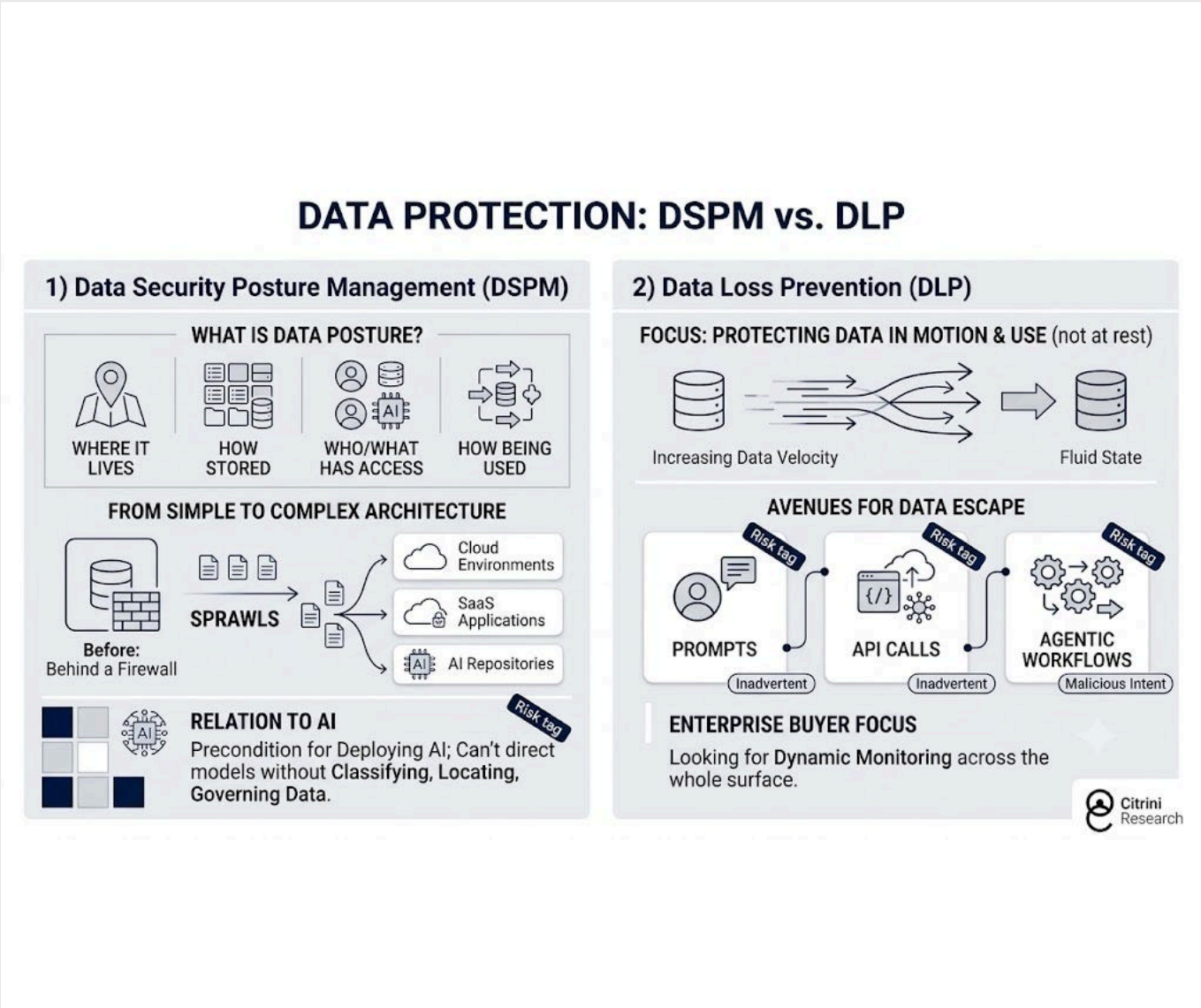
Like Fortinet, Yubico has grafted a recurring revenue business upon their physical hardware. They've coined this business **YaaS,** or YubiKey as a Service, which embodies SaaS-like economics with the added lock-in of a physical device.

Data Security

The data protection layer is one area we haven't spent a ton of time on. In cybersecurity parlance, this service is broken down into two complementary functions: **data loss prevention (DLP)** and **data security posture management (DSPM)**.

DSPM dictates the “posture” of data – that is, where it lives, how it’s stored, who (or, increasingly *what*) has access, and how it’s being used. That job used to be fairly simple when data lived in a database behind a firewall, but data now sprawls out across cloud environments, SaaS applications, and entirely new repositories created by AI.

DLP, in a similar vein, is focused on protecting data in motion or in use, rather than data at rest. The velocity of data continues to rise, and the state of data is increasingly fluid. Every prompt, API call, and agentic workflow creates another avenue in which data can escape – either inadvertently or by malicious intent.



The rising cadence and sophistication of cyberattacks creates tangible ramifications for the data layer: DLP backstops the digital infrastructure, while DSPM serves as the watchtower. However, as we’ve discussed in depth before, cybersecurity businesses are rewarded for *scale.* Customers are exhausted by vendor sprawl while vendors are looking to grow the top line. The endgame being that cybersecurity companies are increasingly becoming platforms, acquiring adjacent capabilities rather than trying to build in-house.

The playbook being run at Palo Alto Networks is the same formula applied to the data security market. **Rubrik** – who was added to Project Glasswing in June – acquired Laminar, a DSPM business, in 2023. The collective

business now infuses DSPM capabilities within the Rubrik Security Cloud, alongside a full suite of services available to customers.

Commvault (CVLT US), a Rubrik rival, acquired Satori Cyber in July of last year. In March, the company extended its DSPM reach into structured databases — including the vector stores underneath AI applications — with real-time access governance.

The same dynamic is unfolding in the private markets: Veeam (backed by Insight Partners, TPG, and Neuberger Berman) paid $1.7 billion for Securiti to bolster their DSPM offering. Cohesity (backed by Nvidia, IBM, Sequoia, and others) hasn't made an explicit bet on this service, electing to partner with Cyera for DSPM capabilities. However, the company's acquisition of Veritas Data Protection in late 2024 reinforces the case that ***scale*** is the North Star in the data security roadmap.

Curiously, however, with six DSPM vendors off the board, one publicly listed pure-play remains up for grabs: **Varonis (VRNS US)** – it's no surprise that they've entered the M&A rumor mill.

AI Sovereignty: Servers and Storage

Within the realm of AI, there has always been a push to "repatriate" the supply chain in the name of national sovereignty.

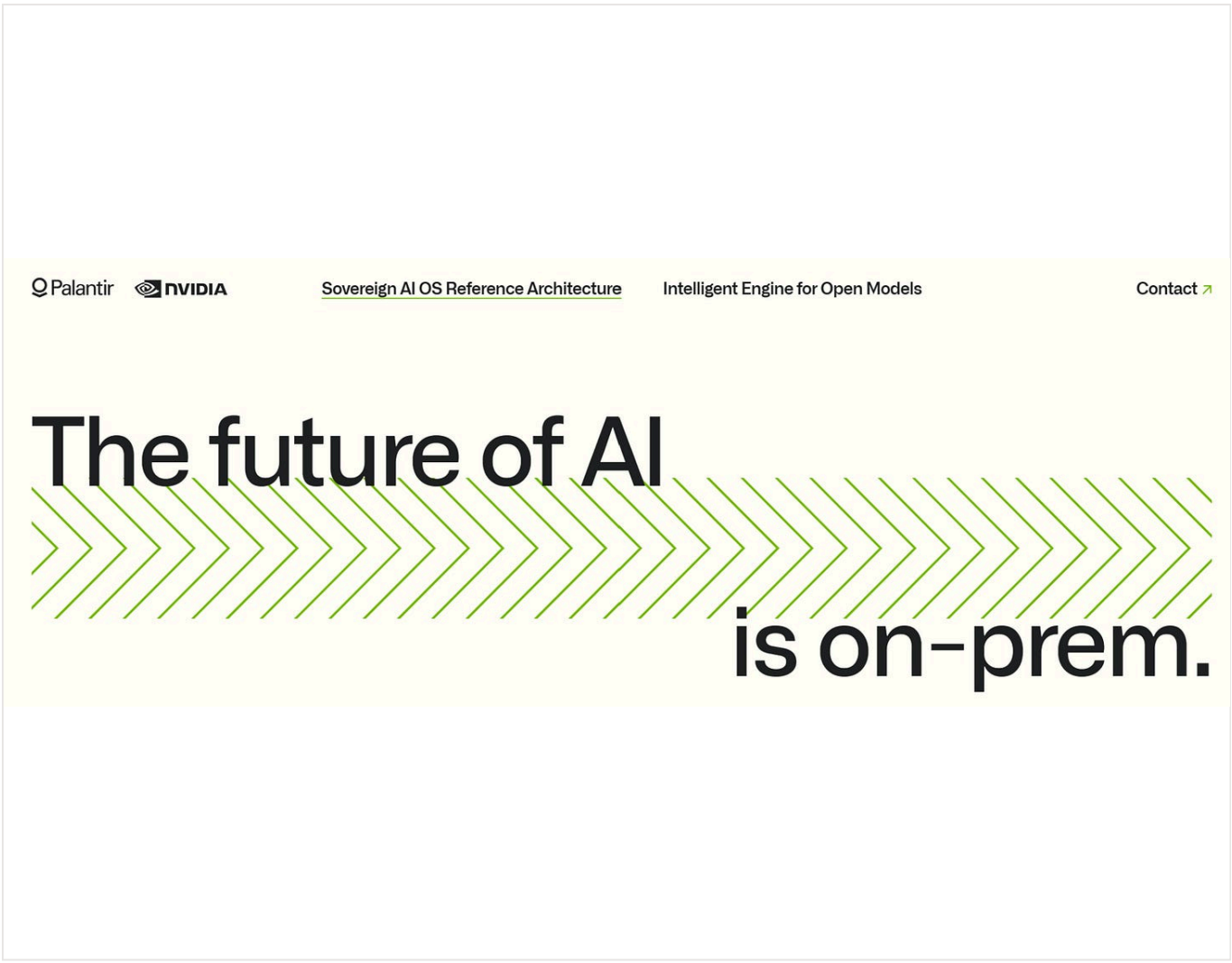
But there is a second interpretation of **sovereign AI** that creates a more investable universe: the proliferation of open-sourced models and the imperative for organizations to own their entire AI stack – from compute (on-prem, private clouds, or colocated), to data, to the custom tailored and continuously updated models themselves. This kind of sovereignty is what can preserve competitive advantage, or "business alpha" in the words of **Palantir (PLTR US)**.

Palantir's argument is that the foundational model complex has begun to exact taxes on the businesses built atop it – running inference through the frontier labs is expensive, and enterprises are increasingly unwilling to build their most differentiated workflows on rented intelligence that they don't control. Not to mention, the privacy anxiety is real – there's skepticism that these businesses won't train their next model on the very data that you're using to power your company.

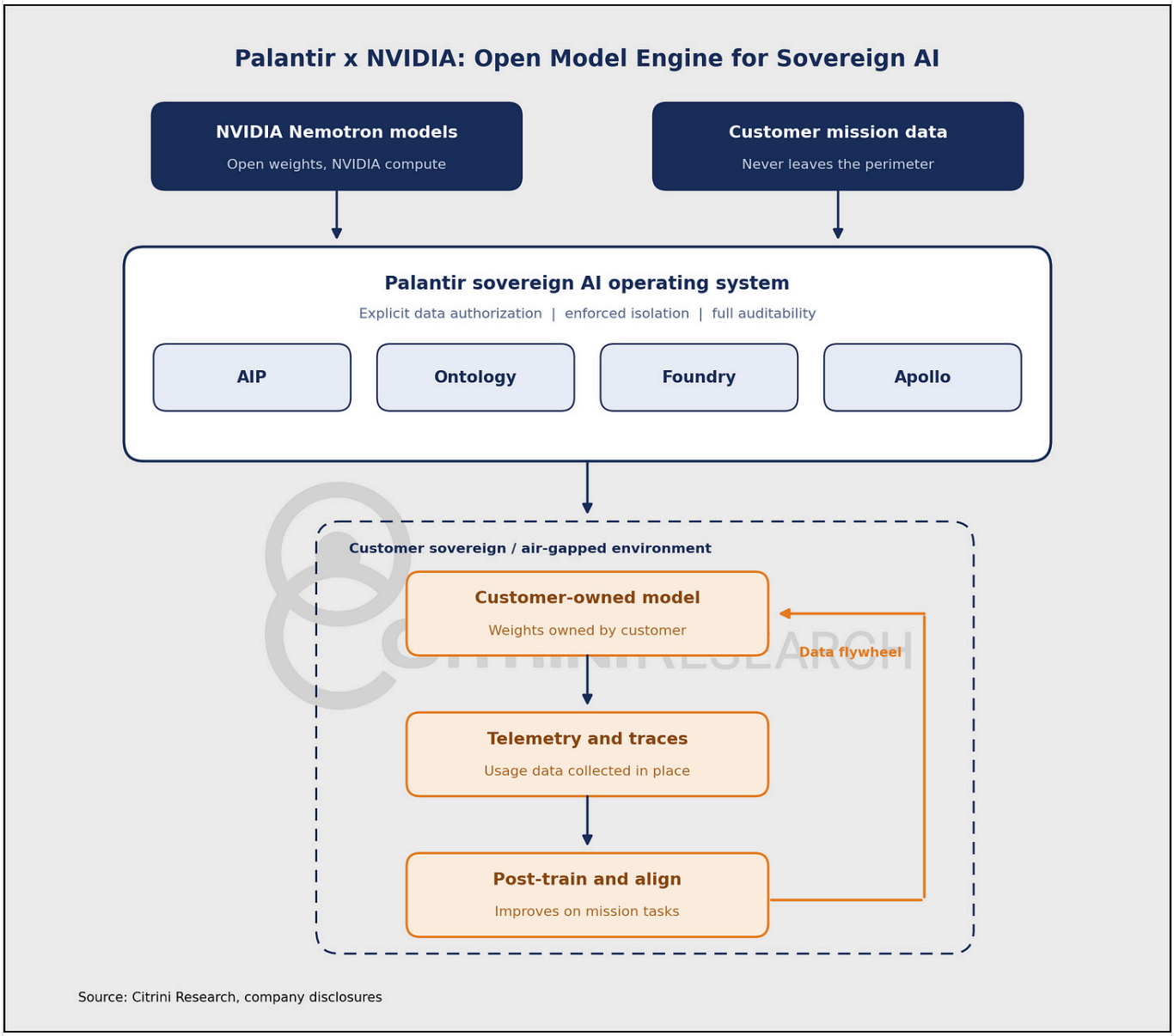
Microsoft CEO Satya Nadella emphasized the point quite directly this past week, in his latest barb against the frontier labs:

> *"In the AI age... You essentially pay for intelligence twice, once with money, and again with something even more valuable: the proprietary knowledge you must reveal to make that intelligence useful."*

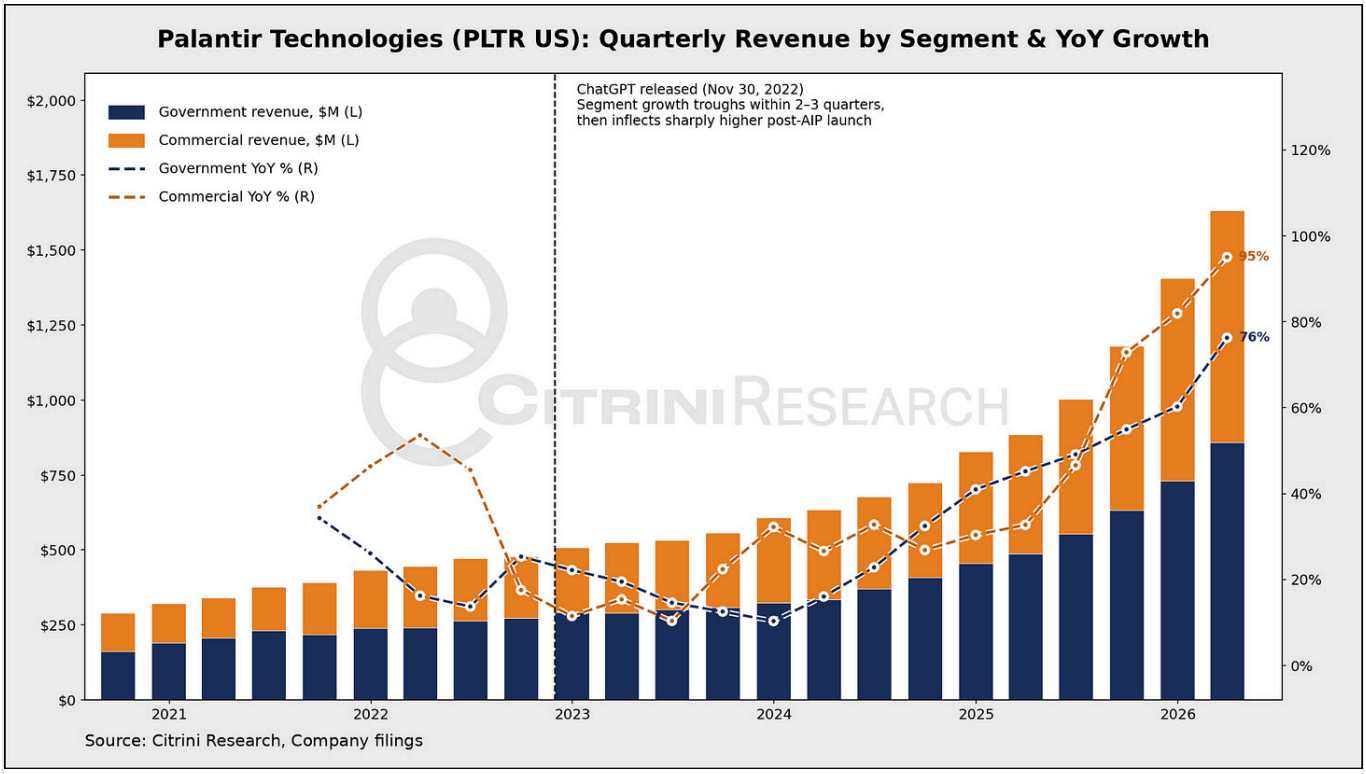
Taken in the context of Microsoft's recent decision to shut down internal use of Claude Code due to "costs", maybe the concern was less around the headline token bill and more about the idea of routing a large share of its software development through a competitor's data-hungry systems.



On June 29, Palantir and **Nvidia (NVDA US)** declared “The Future of AI is on Prem”. The duo released a “Sovereign AI OS Reference Architecture” which combines NVDA’s open weight Nemotron models and GPUs, Dell’s PowerEdge R670 server, and Palantir’s software suite to create a truly self-contained AI platform – one in which data is safe and your intelligence is both tailored and compounding.



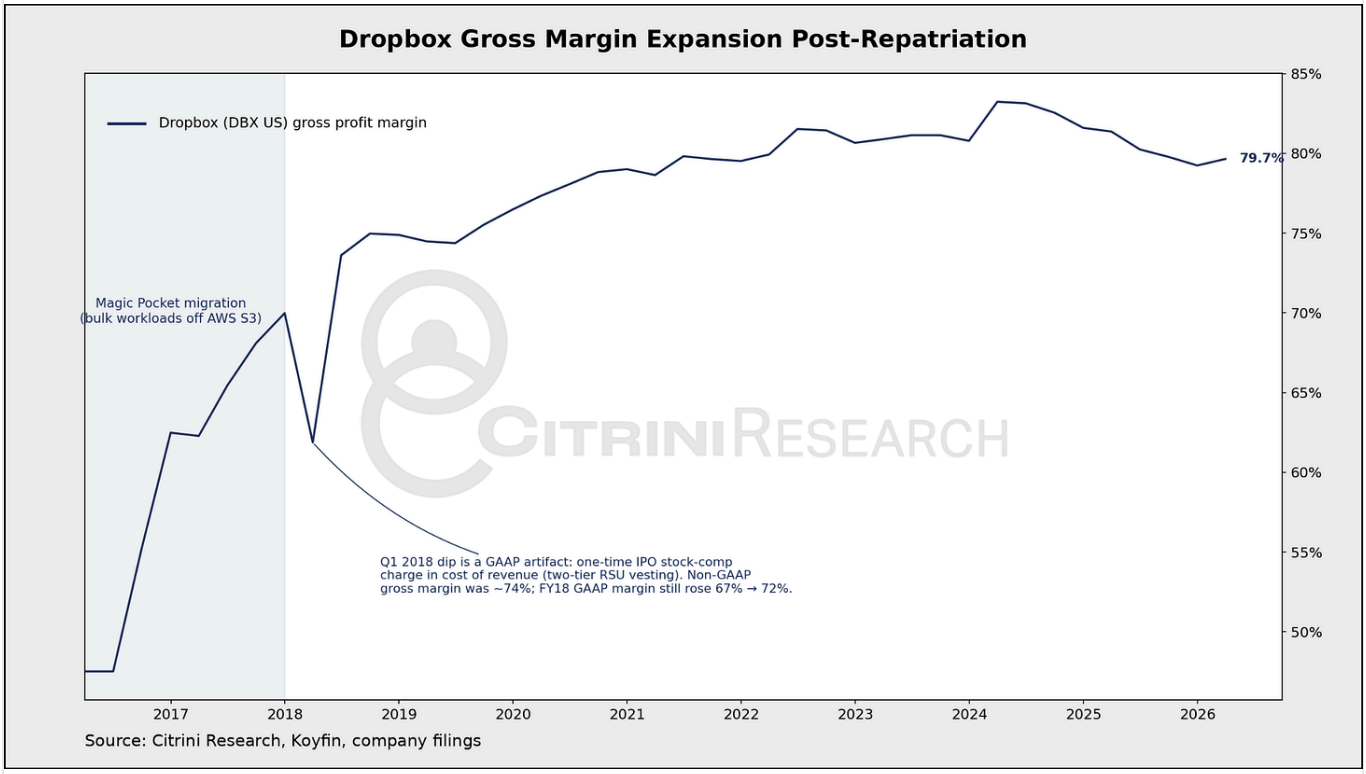
To be clear, implementing one of these PLTR/NVDA solutions with the hardware specs it outlines will run from $5-$6 million for the smallest configuration (32 GPUs across 4 nodes) to $50-$70 million at the largest. That's before the ongoing cost of PLTR software licenses, hardware refreshes and electricity/colocation bills.



But it's not a coincidence that the question of sovereign AI is running in parallel to a market dynamic that has been unfolding, albeit quietly, across the software industry. That is, the repatriation of computing architecture from public clouds back to on-prem, colocated, private clouds, or hybrid workloads.

For the better part of the 2010s, the common wisdom was to spin up an application and host it on a public cloud – AWS, Azure, GCP. *Elasticity* was the key selling point. A software-native business could scale from 100 users to 100,000 users without the time and investment required for an on-prem buildout. Of course, **those economics change once you've hit critical mass.**

Dropbox (DBX US) is the most salient example of this in practice. Prior to their 2018 IPO, the company made the decision to shift 90% of user data off of AWS and onto custom-built infrastructure in colocated facilities. The project was dubbed *Magic Pocket* and substantially improved DBX gross margins.



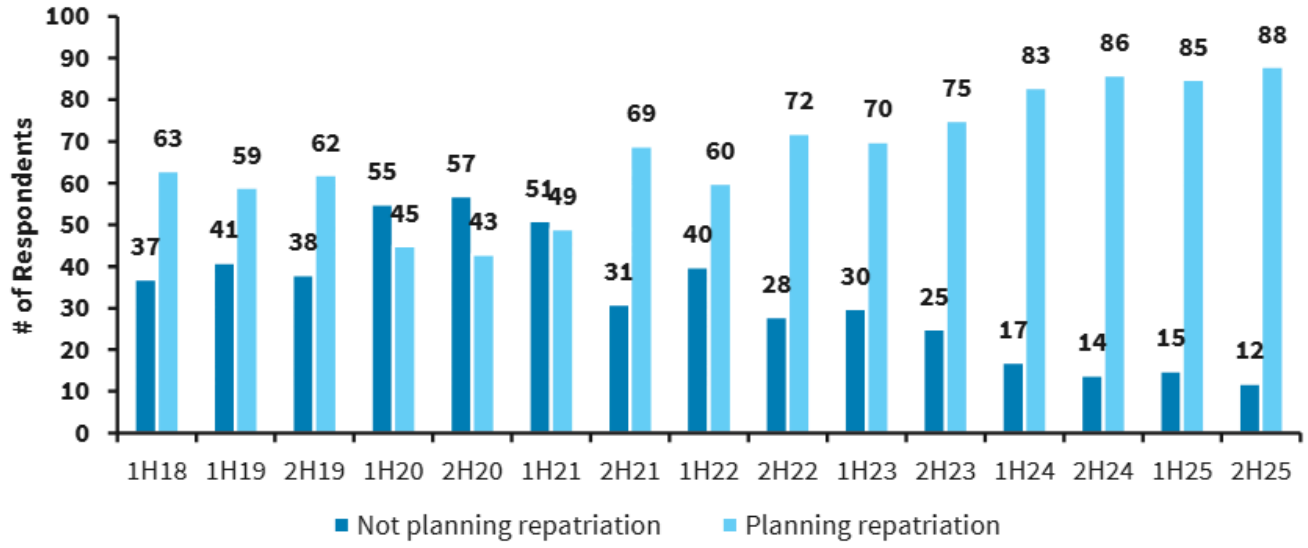
Other companies followed suit – including Berkshire-owned **GEICO** and privately-held **37signals**, the parent company of Basecamp and Hey. GEICO specifically spent a decade shifting hundreds of applications into the cloud, which resulted in a $300 million annual cloud bill.

They are now in the process of repatriating workloads onto an OpenStack private cloud built on Open Compute Project hardware, targeting a 50% reduction in compute spend and a 60% decrease in storage costs per unit. 37signals, in a similar vein, transformed a $3 million-plus annual AWS bill into roughly $700,000 worth of Dell servers and a further $1.5 million of Everpure arrays. The company officially deleted their AWS account in 2025 and anticipates over $10 million in savings over the next five years.

While these are only a few examples, the trend is real. According to a survey conducted by Barclays, 88% of enterprise CIOs intend to move at least some public cloud workloads back to private cloud or on-prem

infrastructure. A study by Broadcom echoed this sentiment, with a significant shift from 2025 to 2026.

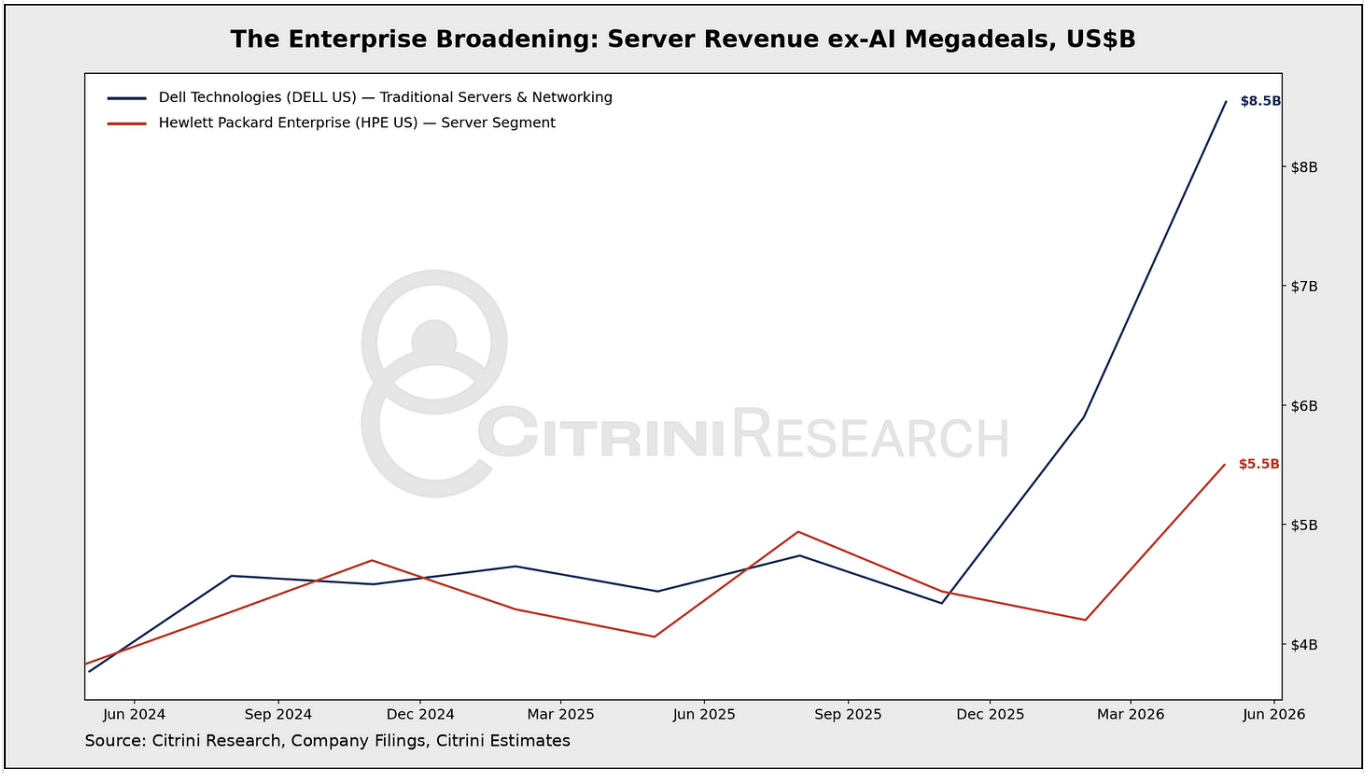
FIGURE 20. **Barclays CIO Survey – Percentage of Respondents Planning to Move Workloads Back on to Private Cloud / On-Premise from Public Cloud**



Source: Barclays CIO Survey program.

To be clear, this doesn't mean that the public cloud is dead – cloud spend continues to grow at a +20% rate and cloud computing is still optimal for up-and-coming businesses with a need for elastic compute. That is – bursty, dynamic, and unpredictable workloads.

Still, the combination of AI sovereignty and cloud-cost trends set up a dual tailwind for enterprise hardware servers and storage – one that is already quite visible in the traditional numbers at **Dell (DELL US)** and **Hewlett Packard Enterprise (HPE US)**.



Likewise, **Lenovo (992 HK)** and **Fujitsu (6702 JP)** are also beneficiaries and even after the re-rating at Lenovo, the valuations are very reasonable relative to other AI hardware beneficiaries if this tailwind holds.

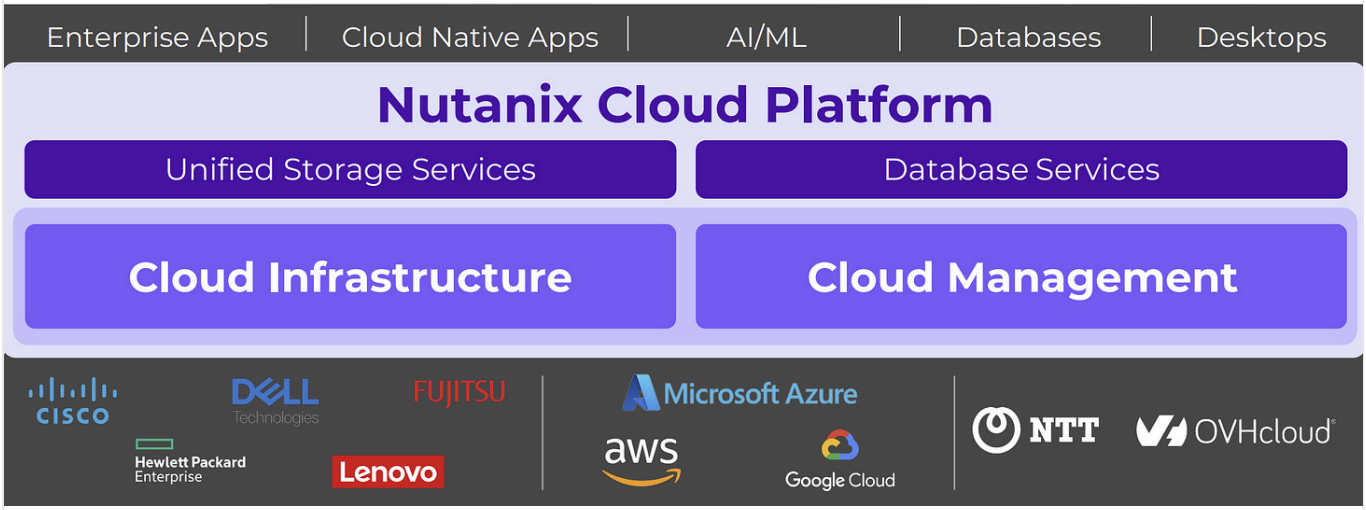
How expansive is the market? Maybe not "every American home will have a GPU attached to it" but certainly *larger*, especially as the hyperscalers migrate towards their own silicon (Trainium, TPUs, Maia). The marginal buyer of merchant accelerators is increasingly the enterprise, the sovereign, the neocloud – alongside the hyperscaler. The market has widened to capture the long tail of buyers: Sovereigns like the EU's €200 billion InvestAI program, Saudi Arabia's Humain, Stargate UAE, and Japan's METI-subsidized compute buildout as well as private-sector buyers like Citadel Securities, Pfizer, and SLB.

Storage players like **NetApp (NTAP US)** and **Everpure (P US)** are first-order beneficiaries of the explosion of data across multi-cloud and hybrid environments. NetApp builds on commodified x86 compute, while

Everpure designs its own proprietary DirectFlash hardware. Data gravity was always the strongest argument against repatriation – moving petabytes out of the cloud is strenuous and expensive. In the sovereign AI framing, it becomes the strongest argument for never letting the data leave in the first place. Management at **F5 (FFIV US)** – whose application delivery and security gear straddles exactly these hybrid environments – put it plainly at their May investor meeting:

> *"And then thirdly, data gravity and governance in AI. We're seeing a large number of companies wanting to keep their proprietary data on-premise, in many cases, repatriating data that was in public cloud on-premise because they want this data that has now become enormously valuable for their AI models, they want their data pipelines to be close to the infrastructure they trust and that's causing, again, a wave of investments in private infrastructure."*
>
> \- *F5 Analyst & Investor Meeting, May 2026*

Of course, there is friction inherent in overhauling your compute footprint. Smoothing those transitions is where **Nutanix (NTNX US)** makes its living – the company helps organizations run and move workloads across every genus of computing environment.



Ironically, Nutanix has also seen tailwinds following Broadcom's acquisition of VMware. Post-acquisition, VMware overhauled its business model – eliminating perpetual licenses, implementing bundle pricing, and jacking up rates on renewals.

To capture the *winners* of this shift in enterprise spending, we've created the Enterprise Security and Sovereignty Basket below.

4\. PEMDAS (Please Excuse Meta's Determined AI Spending)

-

In August of 2025, in our **State of the Themes** update, the AI narrative felt like it was in a very similar position with a clearly defined winner (OpenAI), an upstart that was making up ground (Anthropic) and a gaggle of hyperscalers that were resigned to being infrastructure providers to the two winners.

Despite the prevailing negative narrative on **Google (GOOGL US)** web search and Gemini disappointment, we highlighted the company as an unappreciated winner and it's up nearly 80% since. Now, a year later, does **Meta (META US)** provide a similar setup?

Source: Citrini Research

Like Google, Meta is becoming an integrated and diversified AI provider with compute, data, chip design, and most importantly, captive users. Unlike Amazon, Microsoft and Google, Meta entered this buildout without a cloud backlog or external customer base to absorb the capacity.

However, Meta has the right order of operations in AI to turn GPUs into revenue more effectively than anyone else. PEMDAS exists because arithmetic punishes the wrong sequence. The META AI bears are subtracting capex first and ignoring the exponents.

While the market is mostly hand-wringing over its capex figures (which are probably about to be revised much higher), there has been less focus on why exactly Meta is throwing so much money behind this plan, and the possibility that it may actually work out. Consider…

- Meta is the fastest growing hyperscaler *without* external cloud revenue and its compute translates directly into P&L
- Foundational AI models have yet to be monetized
- Capital expenditures are a choice

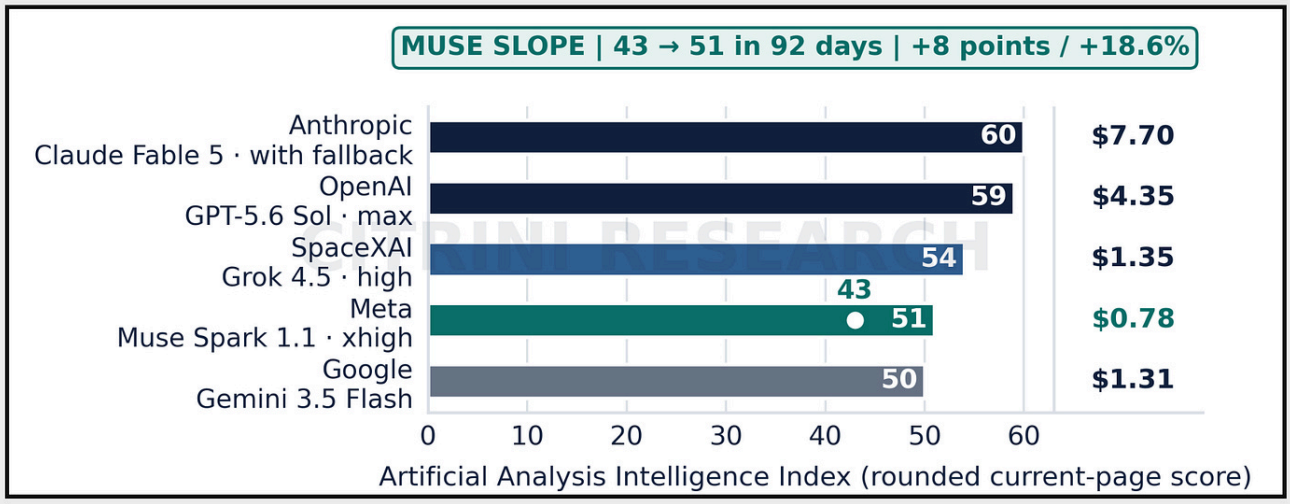
Meta Superintelligence Labs (MSL) is younger and lacks the pedigree of some of the DeepMind greats. But what MSL lacks in research talent, it makes up for in the sheer amount of data talent that Alexandr Wang was able to wave in from Scale AI. MSL has turned into the premier RL factory which offers it a strong captive advantage over the labs.

Given that Meta has the homogeneous compute on its roadmap to train a large 10 trillion+ parameter model, the question is whether it can reliably post-train its models to the reasoning efficiency of frontier class models.

The Muse Spark 1.1 is certainly encouraging.

Muse Entered the Near-Frontier Cluster

Selected provider maxima; Meta ranks 4th here. Exact configs and blended USD per million.



SOURCE: Artificial Analysis model leaderboard; Citrini Research

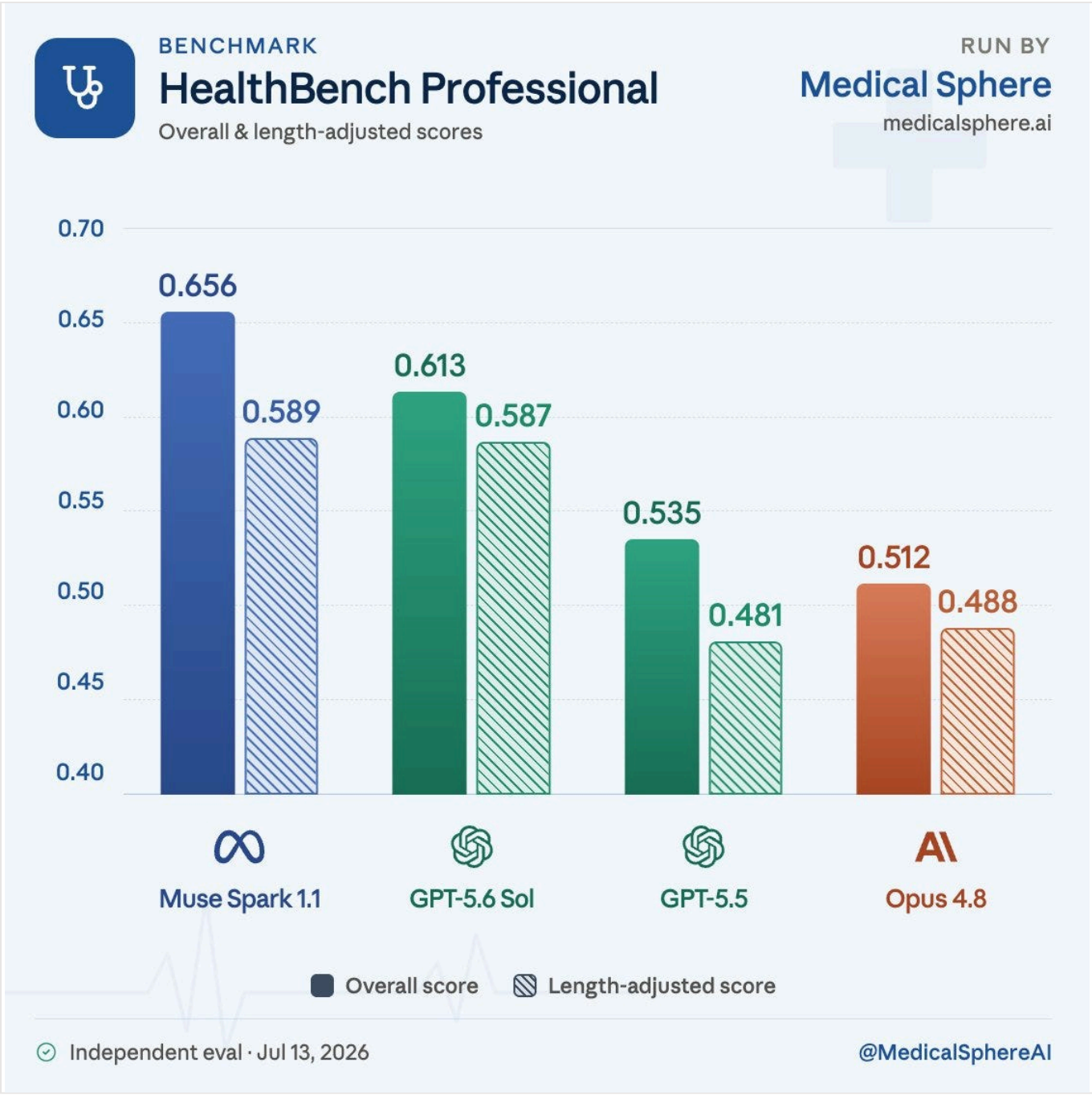
AS OF JULY 13, 2026

Scores are rounded; configurations and fallback routing remain attached. Rolling speed and latency are omitted.

But like we did with Google DeepMind last year, let's just assume that MSL is simply an option. In the best case, MSL wins the race due to its concentration of RL data and homogeneous compute.

Most likely, Meta emerges among the winners of low-cost closed source model providers. Meta's Muse is by far the cheapest quality US model on the market, beating out even Grok in pricing. As Zuck himself even says: "The pricing from some of the other labs is very extreme and has very high margins. We think that there's a real ability to be able to offer frontier or very high-level intelligence at a much more affordable cost."

And he has a point about frontier intelligence. Muse has been performing especially well on benchmarks that actually matter like HealthBench.

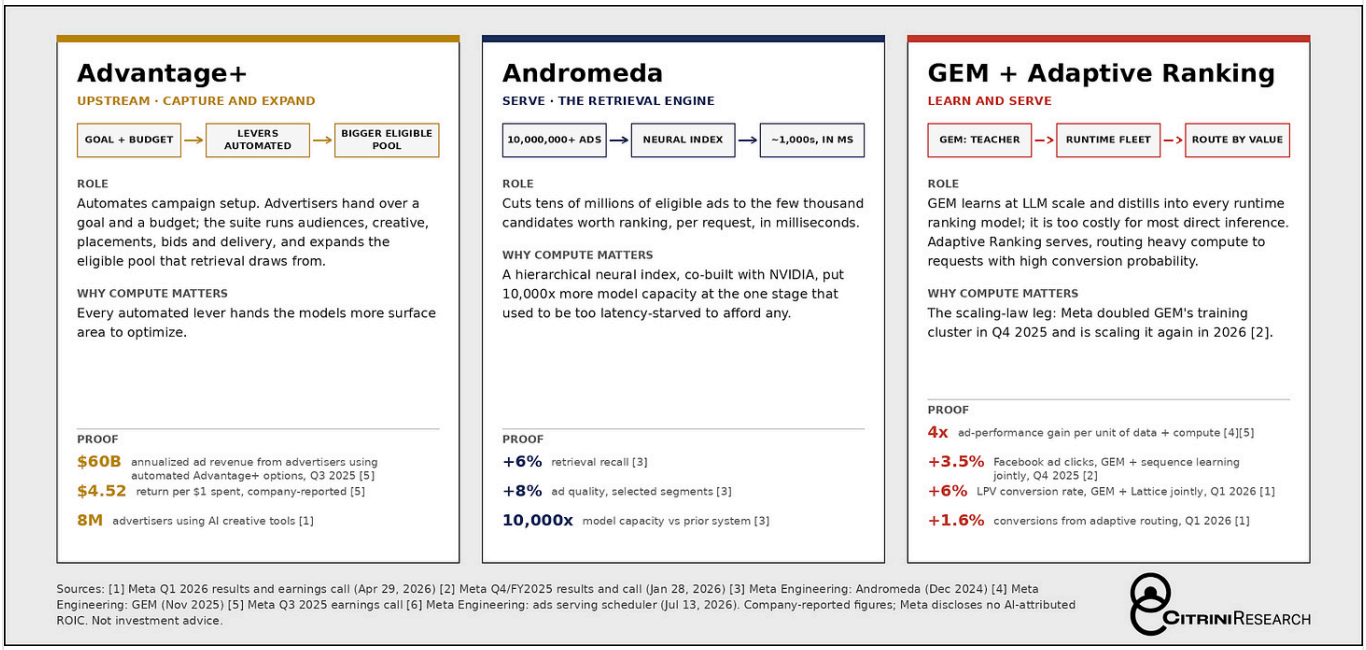


Even at blended token pricing of $1, the economics for an integrated provider are still attractive as Meta has shown the ability to optimize their hardware for parallelization and latency. And for enterprise users that are sensitive about cost, sacrificing a few percentage points of performance for 90% cost savings versus Anthropic is going to merit real discussions as we enter renewal season in Q4.

But even more attractive are Meta's existing generative AI capabilities and their cash generation potential from the enterprise. Meta, more so than Google, has already demonstrated that it can use generative AI to improve recommendations, creative, and advertising performance in a way that is generating extremely high margin revenue ***today***.

Exponents: Monetizing the Models

The same principle feeds directly into Meta's advertising business. Meta has three core GPU-optimized AI products in its advertising business:



Although Andromeda and GEM run on separate inference infrastructure rather than the Titan training clusters themselves, Andromeda was co-designed around the Nvidia Grace Hopper memory hierarchy and interconnect. We can assume the next generation of generative models for advertising will follow the same principles.

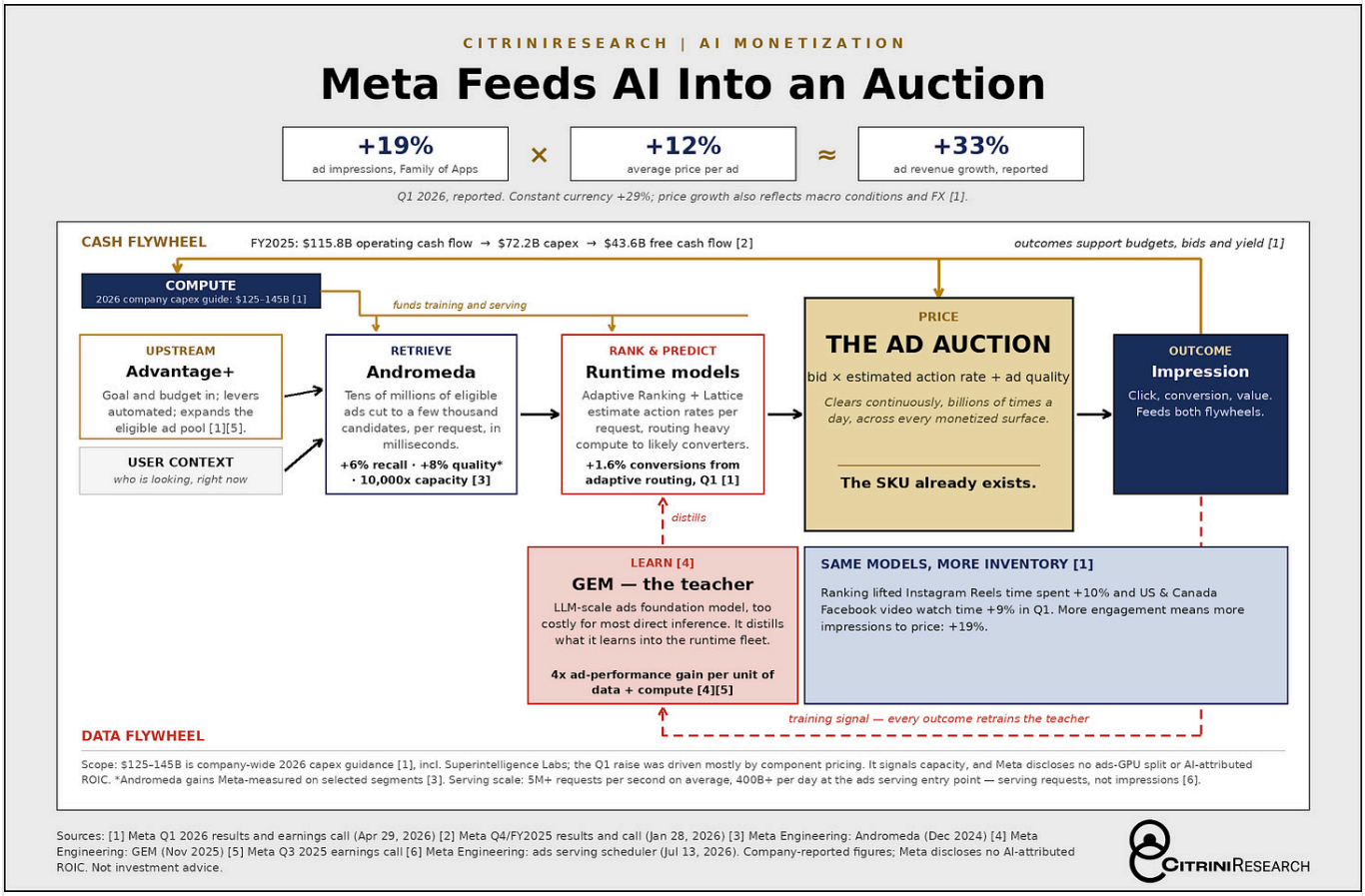
Meta can therefore tune the model architecture, feature pipeline, memory placement and fused GPU kernels against a defined hardware target, extending its lead over Google in monetizing its AI compute for advertising

gains. Precomputed ad embeddings remain in local memory, avoiding repeated CPU-to-GPU transfers and allowing the system to exploit massive parallelism for low latency workloads.

Meta reports that this co-design increased Andromeda's model capacity by 10,000 times, improved inference efficiency by 10 times, and more than tripled end-to-end query throughput. Deployment produced a 6% retrieval-recall improvement and an 8% increase in ads quality across selected segments.

Advantage+ automation generates more audience combinations, placements and creative variants, increasing the number of eligible ads Andromeda must search. **Meta reports that advertisers using Advantage+ shopping campaigns saw a 22% increase in return on ad spend, while those using Advantage+ creative's image generation saw an estimated 7% conversion lift – fueling a large portion of relative market share gains versus Google.** We think that these performance improvements are part of the value Meta is extracting from its GPU-based advertising optimization system.

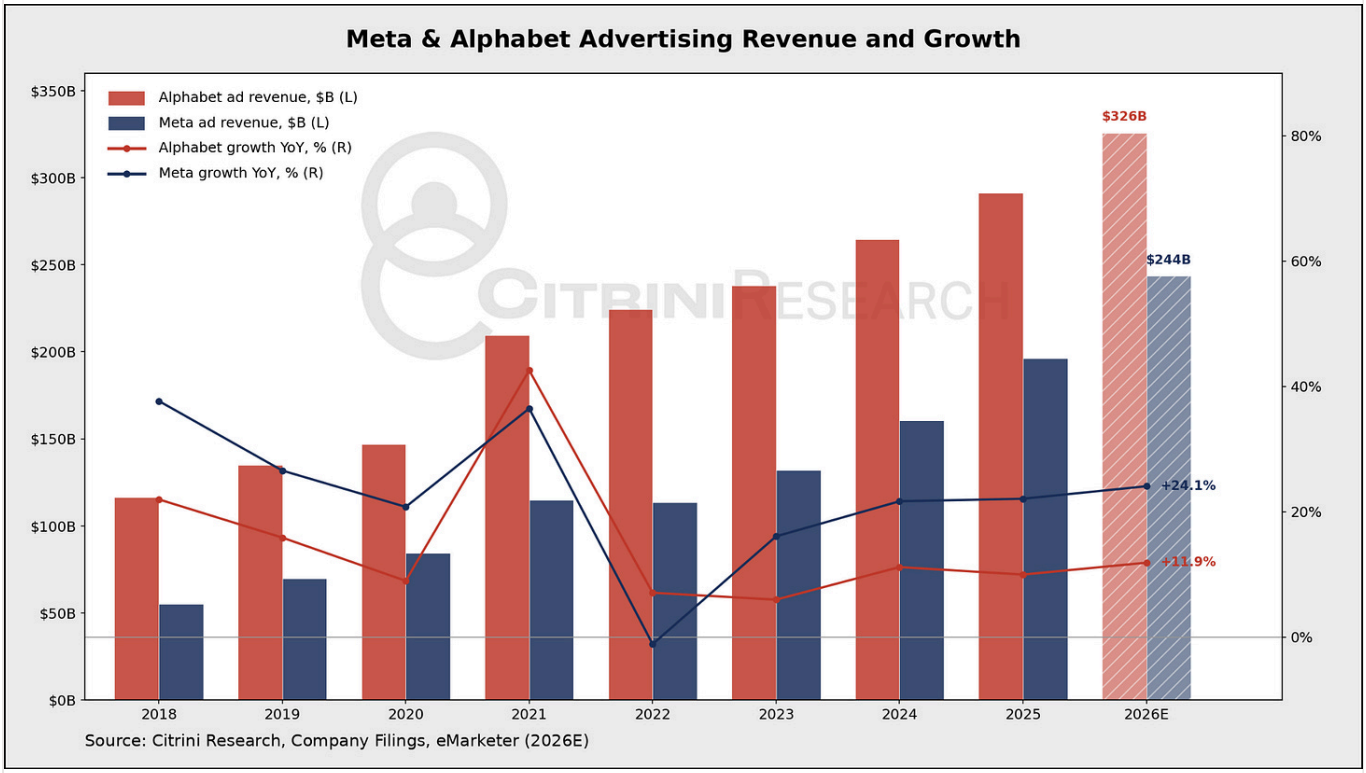
The cash-to-compute-to-ad-revenue-to-compute flywheel is elegant.



Meta also has another trick up its sleeve that we believe the market hasn't been paying attention to. Currently advertisers on Google can access CTV and programmatic inventory through DV360, but there is no equivalent on the Meta Ads platform. Meta is now laying the groundwork to take its advertising machine beyond its own apps. Digiday reported discussions with Magnite, FreeWheel, TV manufacturers, and publishers to secure third-party CTV inventory. "The goal would not be to own the media but to own the demand."

Advantage+ supplies advertisers and creative, Andromeda retrieves the right inventory, and Meta's optimization and measurement stack closes the performance loop in a way that no one is offering right now. **The U.S. opportunity is $40-50 billion of annual advertiser spend — IAB sized CTV and online video at $45.2 billion in 2025, while EMARKETER expects open-web programmatic display to reach $42.5**

billion by 2027. Given that Meta has the best closed loop system currently, we think it's not unrealistic to expect Meta to take a huge chunk of the $40 billion+ in spend quickly.



This makes Meta's compute position structurally different from that of a standalone model lab. Frontier training consumes the GPUs, but advertising can monetize the same infrastructure expertise immediately. Advantage+ creates more candidates, Andromeda retrieves them more efficiently, better retrieval improves advertiser returns, and the resulting revenue funds the next generation of compute. Meta is building both the factory and the cash-generating application that keeps it loaded and is doing it much more efficiently than Google despite not having their own training silicon yet.

The idea that Meta would have to become a cloud platform to soak up its near 14GW planned over the next three years is a complete smokescreen to quell investors in case it takes time for any of GEM, Andromeda, or

MSL to ramp up to full capacity.