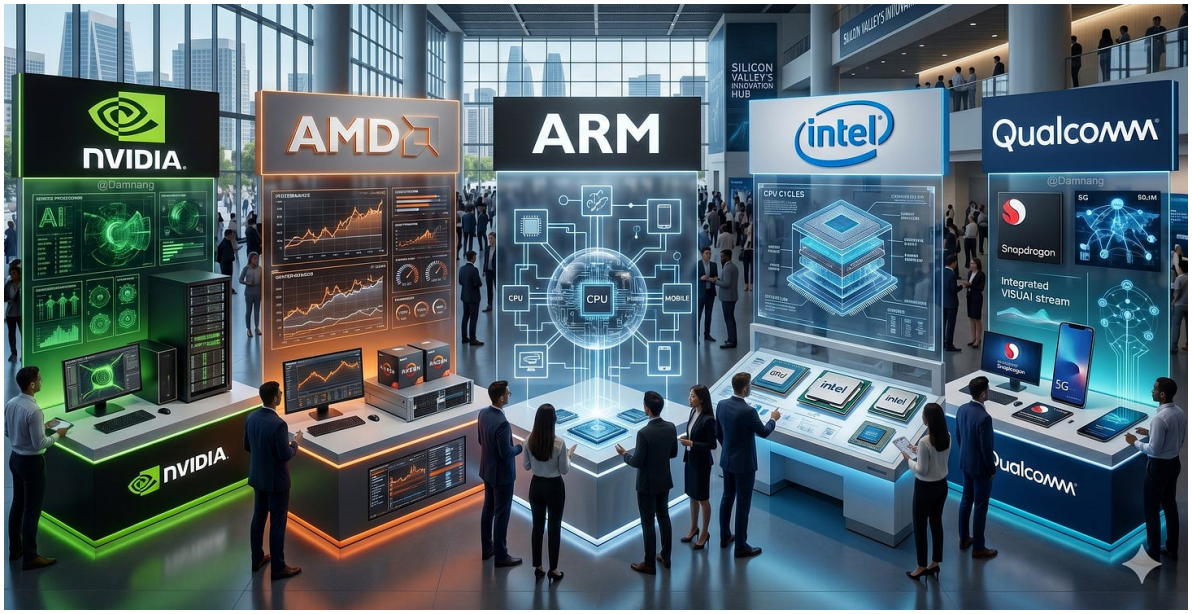
damnang2.substack.com

The CPU Bottleneck Trade: Who Actually Gets Paid in the Agentic AI Era?

Damnang

42–54 minutes



After the close on April 23, 2026, Intel CEO Lip-Bu Tan said something that mattered on the Q1 earnings call.

> "The CPU is reasserting itself as the indispensable foundation of the AI era. The CPU now serves as the orchestration layer and critical control plane for the entire AI stack."

The message: CPUs are moving back to the center of AI infrastructure.

The market reacted immediately.

The next day Intel jumped +23.6%, AMD +13.9%, ARM about +14%, Qualcomm +11%, and even NVIDIA was up +4.3%.

The Philadelphia Semiconductor Index crossed 10,000 for the first time.

What mattered more was CFO David Zinsner's framing. As AI workloads shift from training to inference and then to agentic AI, the GPU-to-CPU ratio could tighten dramatically. Per outside reporting, the training era ran on roughly 1 CPU per 7 to 8 GPUs. Inference brings that down to 1 per 3 to 4. Once agentic AI scales out, it could move to 1-to-1 or higher.

If you've been reading my articles, you already know I've been arguing for months that CPU becomes a clear bottleneck in the agentic AI era. That thesis is finally getting priced in.

Don't jump straight from this to "every company that sells a lot of CPUs is a winner." A higher CPU ratio doesn't automatically pull CPU ASP (Average Selling Price, the average revenue per chip) and margin up with it.

What hyperscalers want might be the highest-performing x86 CPU. It might also be lower power, lower TCO, enough host CPU to orchestrate GPUs, more memory bandwidth, easier software integration.

So the real question for this cycle isn't simply "who sells the most CPUs."

Who can convert rising CPU demand into their own profit pool most effectively.

This piece walks through the technical differences between ARM, x86, and RISC-V, the monetization structure of data center CPUs, what CPUs actually do in agentic AI, and how Intel, AMD, NVIDIA, ARM, and Qualcomm each get paid in this cycle.

Disclaimer

This article is for informational purposes only. It is not a recommendation to buy or sell any security. Every investment decision and its outcome belongs to the reader.

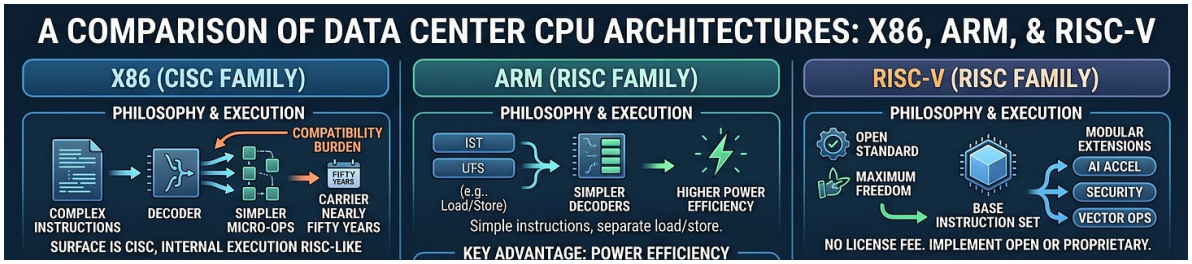
A CPU is a machine that decodes and executes the instructions software hands it. The agreement that defines what those instructions look like and how they get exchanged is the ISA, or Instruction Set Architecture. Think of it as the language between the CPU and the compiler.

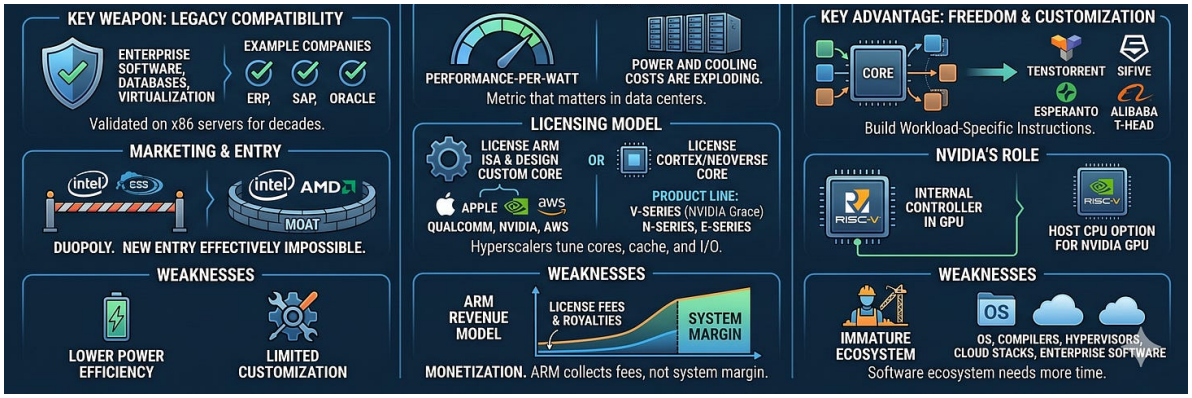
The same C code compiled for x86 produces an x86 binary. Compiled for ARM, it produces an ARM binary. That's why an x86 program doesn't just run on an ARM CPU.

ISA and CPU design aren't the same thing. The ISA is the visible instruction interface. The microarchitecture is the internal circuitry that decides how those instructions actually execute. That's why Intel and AMD both build x86 CPUs but ship completely different cores.

ISA isn't a technical taxonomy. It defines the software ecosystem, the licensing structure, the entry barrier, and the monetization model all at once. To see who can make money in CPUs, you start with ISA.

In modern data center CPUs, three ISAs matter: x86, ARM, and RISC-V. There are several ways to slice ISAs, but the most useful axis for investors today is x86's CISC family versus the RISC family that includes ARM and RISC-V.





x86 is in the CISC family. CISC stands for Complex Instruction Set Computing, an approach that includes a relatively large set of complex instructions. A single instruction can handle both memory access and computation, which improves code density but makes the decoder more complex.

Modern x86 CPUs internally crack complex x86 instructions into simpler micro-ops before executing them. So the surface is CISC but the internal execution looks closer to RISC. Carrying nearly fifty years of compatibility still leaves a real burden on the decoder and front-end.

x86's biggest weapon isn't raw performance. It's legacy compatibility.

Enterprise databases, virtualization, ERP, SAP, Oracle, all of these have been validated on x86 servers for decades. For a large company, switching CPUs isn't swapping a chip. It's redoing software certification, operational stability, and the entire maintenance stack.

The licensing structure is also a strong moat. The high-performance server x86 market is essentially a duopoly between Intel and AMD. VIA and Zhaoxin exist as exceptions but have no meaningful share in mainstream data center. New entry into x86 is effectively impossible.

The weaknesses are clear too. x86 trails ARM in power efficiency and customization freedom. Hyperscalers in particular want to design their own CPUs around their own workloads, and x86 offers almost none of that flexibility.

So x86 stays strong in general enterprise while ARM eats into hyperscaler internal workloads quickly.

ARM is in the RISC family. RISC stands for Reduced Instruction Set Computing, a design philosophy that keeps instructions simple and separates load and store. Instructions get simpler, code length can grow, but decoders get easier to build and power efficiency goes up.

This is exactly why ARM owned mobile. In smartphones, power efficiency mattered more than raw performance, and ARM fit that constraint best.

The same logic is now climbing into the data center. As power and cooling costs in AI data centers explode, performance-per-watt becomes the metric that matters.

ARM's other strength is its licensing model. A company can license the ARM ISA and design its own core, or take an ARM-designed Cortex or Neoverse core off the shelf. Apple, Qualcomm, NVIDIA, and AWS each

design or customize ARM-based CPUs for their own products.

In data center, the line that matters is Neoverse. The high-performance V-series goes into NVIDIA Grace, AWS Graviton4, and Google Axion. The N-series targets server designs that emphasize core count and power efficiency. The E-series leans toward low-power infrastructure and edge.

The advantages are obvious. Power efficiency is good, and hyperscalers can tune core count, cache, and I/O to fit their workloads. That's why AWS Graviton, Google Axion, Microsoft Cobalt, and NVIDIA Grace are all ARM-based.

The investor-side weakness is monetization. ARM doesn't sell chips. It collects license fees and royalties. Even if ARM-based CPU share rises, the system ASP that NVIDIA or AWS captures doesn't flow back to ARM.

ARM can be a share winner without being a system margin winner.

RISC-V is also in the RISC family. The decisive difference from ARM is that it's an open standard. The ISA itself is published, and anyone can use it without a license fee. Implementations can be open-source or proprietary.

The core point of RISC-V is freedom. On top of a base instruction set, custom extensions get added in a modular way: data types for AI acceleration, security features, vector ops, special instructions tuned to a specific workload. No license negotiation like ARM. No closed ecosystem like x86.

That matters to hyperscalers and AI chip startups. When AI workloads are moving fast, being able to extend and modify the ISA is an asset. Tenstorrent, SiFive, Esperanto, Andes Technology, Alibaba T-Head, and Xiangshan are all building on RISC-V for that reason.

NVIDIA isn't ignoring RISC-V either. NVIDIA has been using RISC-V inside its GPUs for internal controllers for years, and recent moves suggest it's opening up RISC-V CPUs as a host CPU option in the NVIDIA GPU ecosystem. This isn't about CUDA running on a RISC-V GPU. It's that a RISC-V CPU can play the host role for an NVIDIA GPU.

The weakness is just as clear. The data-center-grade software ecosystem isn't yet as mature as x86 or ARM. OS, compilers, hypervisors, cloud stacks, and enterprise software certification all need more time. An open ISA is one thing. Running at scale in production data centers is another.

So RISC-V today isn't a player taking large data center CPU share. It's a longer-term wildcard that could eventually shake the cost structure and customization model sitting between ARM and x86.

By analogy, x86 is closer to Windows. Old and heavy, but with overwhelming compatibility and installed base. In enterprise, that legacy is a serious moat.

ARM is closer to Android. Multiple companies build their own products under license, and the strengths are power efficiency and customization. ARM sits at the center of every hyperscaler's in-house CPU effort.

RISC-V is closer to Linux. Maximum freedom and the largest long-term potential, but the ecosystem is still catching up. For now it's mostly private

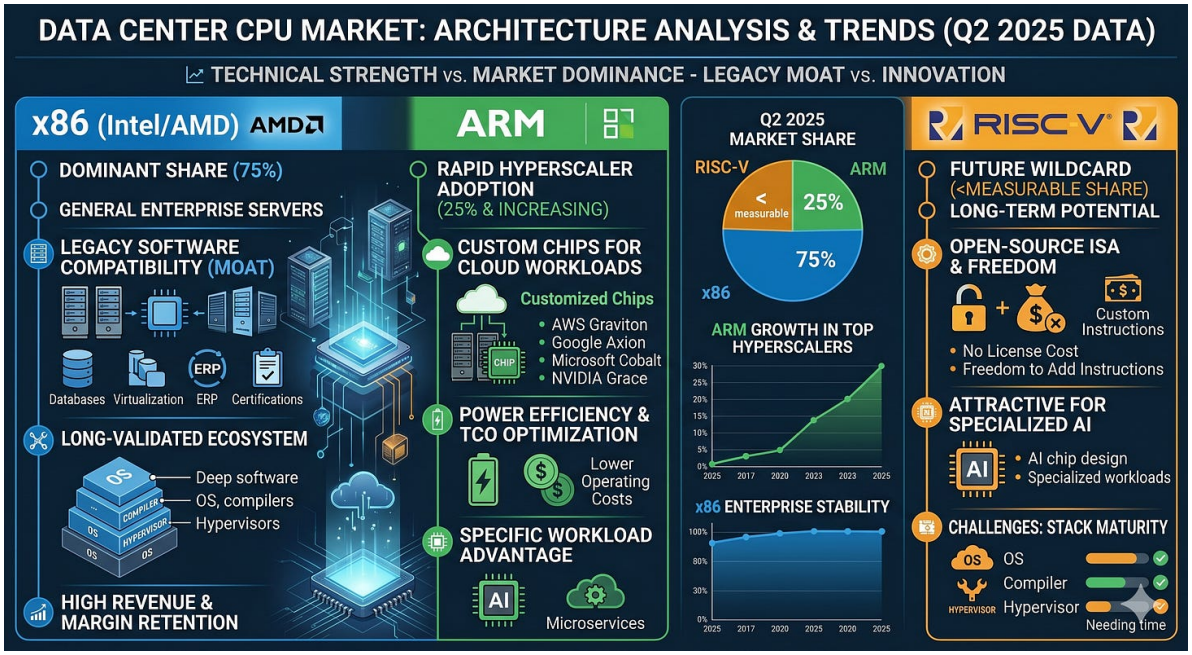
companies, internal controllers, and special-purpose accelerators. Over time it can reshape the data center CPU market structure.

ISA differences aren't just engineering preferences. x86 connects to Intel and AMD's socket margin. ARM connects to hyperscaler in-house chips and the royalty model. RISC-V connects to zero license cost and custom-silicon optionality.

So the right question for this CPU cycle isn't "which ISA is better."

The better question is this:

Whose revenue and margin does each ISA's technical advantage actually convert into.



x86 is still the center of the data center CPU market. In general enterprise servers especially, legacy software compatibility is a serious moat. Databases, virtualization, ERP, certification matrices have all been validated on x86 for years, so most enterprises can't move to ARM easily. That's why Intel and AMD still take the bulk of data center CPU revenue.

Hyperscalers are different. AWS, Google, and Microsoft know their workloads cold and control their own software stacks. Power efficiency, TCO, and customization can matter more to them than legacy compatibility. Which is why ARM-based CPUs like AWS Graviton, Google Axion, Microsoft Cobalt, and NVIDIA Grace are scaling fast.

RISC-V isn't yet mainstream in data center. The advantages are real. No license cost, freedom to add the instructions and extensions you need. As AI workloads get more specialized, that freedom looks attractive. But running at scale in actual data centers requires deeper OS, compiler, hypervisor, cloud, and certification stacks. For now it's a long-term wildcard.

The numbers still favor x86. As of Q2 2025, Intel and AMD combined hold roughly 75% of data center CPU share, ARM around 25%, with RISC-V below measurable levels. ARM, though, sees its share inside the top hyperscalers climbing faster.

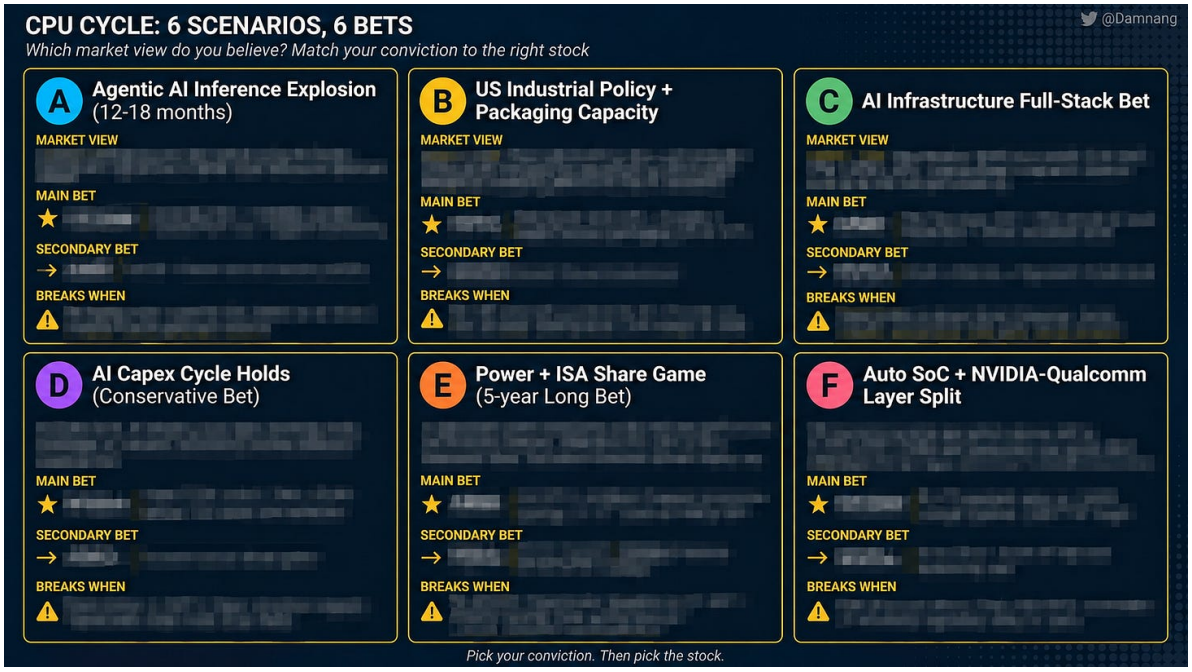
What matters more than the share figure itself is whose revenue and margin that share connects to.

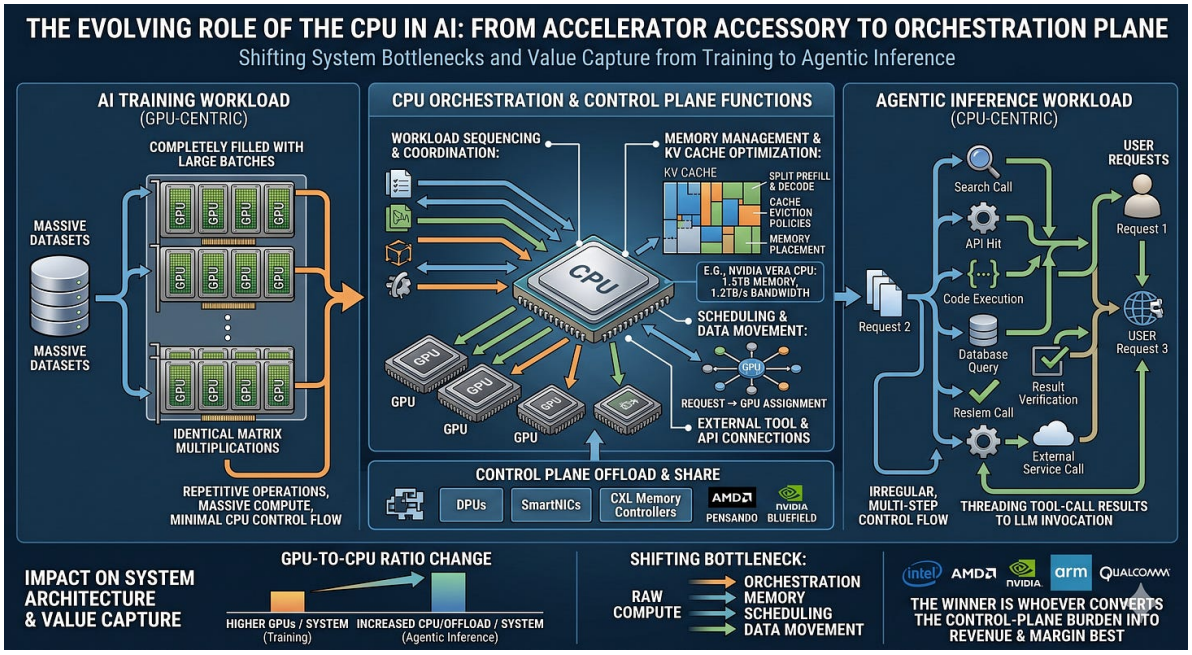
Anyone can say CPU becomes a bottleneck in the AI agent era.

Looking at how the stocks are moving, CPU demand is clearly turning. Figuring out which company actually benefits, and from what angle, is a separate problem.

You need the technology, the industry structure, the asset stack, the monetization model, and what's already in the price versus what isn't.

What follows is a deep dive on the five major CPU companies, plus six forward-looking scenarios and which name fits best inside each. If you're thinking about how to position for the CPU bottleneck, this is the part to read carefully.





When Lip-Bu Tan called the CPU the "orchestration layer and control

plane for the AI stack," it wasn't a marketing line. As AI workloads shift from training to inference and then to agentic AI, GPUs alone stop explaining total system performance.

In training, filling a GPU is easy. Massive datasets go in as large batches, and the same operations repeat. Agentic inference is far more irregular. Every user request is different, with search calls, API hits, code execution, database queries, and result verification cutting in along the way. From the GPU's seat, it isn't just running large matrix multiplications back-to-back.

That's where the CPU gets bigger. The CPU sequences and coordinates work, decides which request goes to which GPU, and threads tool-call results back into the next LLM invocation.

When an agent searches, computes, calls an external service, and verifies the answer to deliver one response, most of that is control flow. Control flow lives closer to the CPU than the GPU.

Memory management is another big piece. Long contexts, multi-agent workflows, repeated tool calls, all of these grow the KV cache. Where you keep prior token state, when you evict it, when you share it across requests, all of that shapes inference efficiency.

Splitting prefill and decode, deciding memory placement, designing cache eviction, each one moves system performance. NVIDIA's emphasis on 1.5TB of memory and 1.2TB/s of bandwidth on the Vera CPU lines up with exactly this trend.

So in agentic AI the CPU isn't a sidecar to the GPU. It's the control plane that splits the work, manages the memory, and connects the network, storage, and external tools so the GPU can compute well. If GPUs are the engines, the CPU is the operating system that keeps those engines from colliding inside an actual service.

That doesn't mean the CPU captures all the value. Some control plane work moves to DPUs, SmartNICs, and CXL memory controllers. AMD Pensando and NVIDIA BlueField become more important for the same reason.

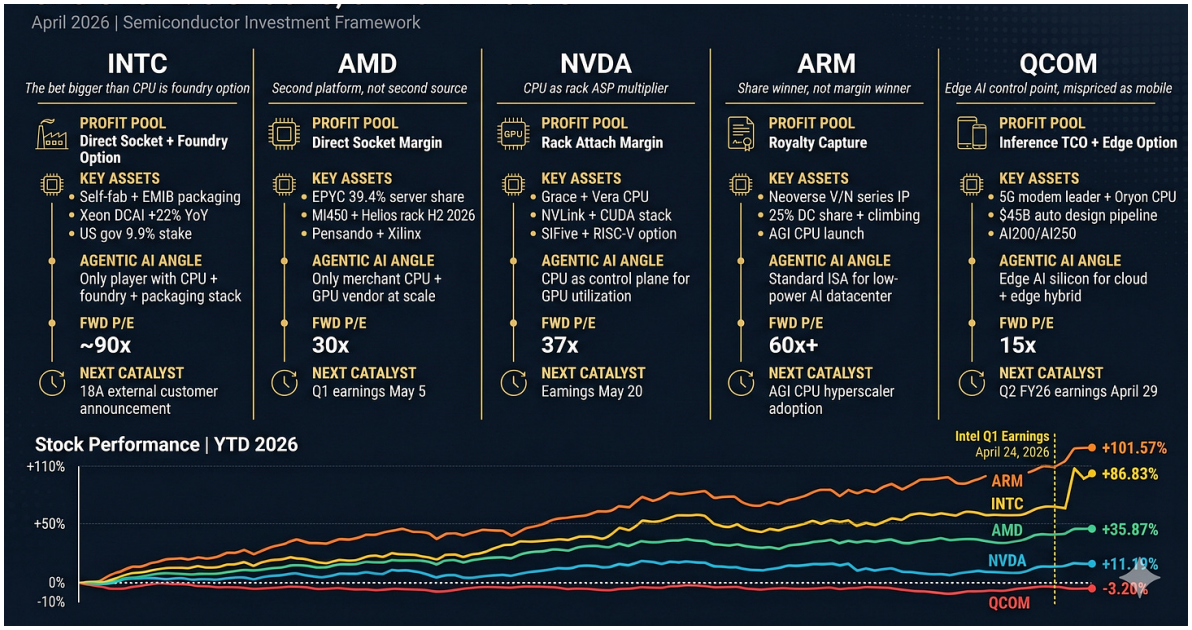
When you look at CPU demand, you have to look at the chips that share the control plane around the CPU as well.

So what Zinsner means by a changing GPU-to-CPU ratio isn't just "more CPUs per system." It's that as agentic AI spreads, the system bottleneck moves from raw compute to orchestration, memory, scheduling, and data movement.

That shift matters. The winner of this CPU cycle may not be the company that sells the most CPUs. The winner is whoever converts that control-plane burden into revenue and margin best.

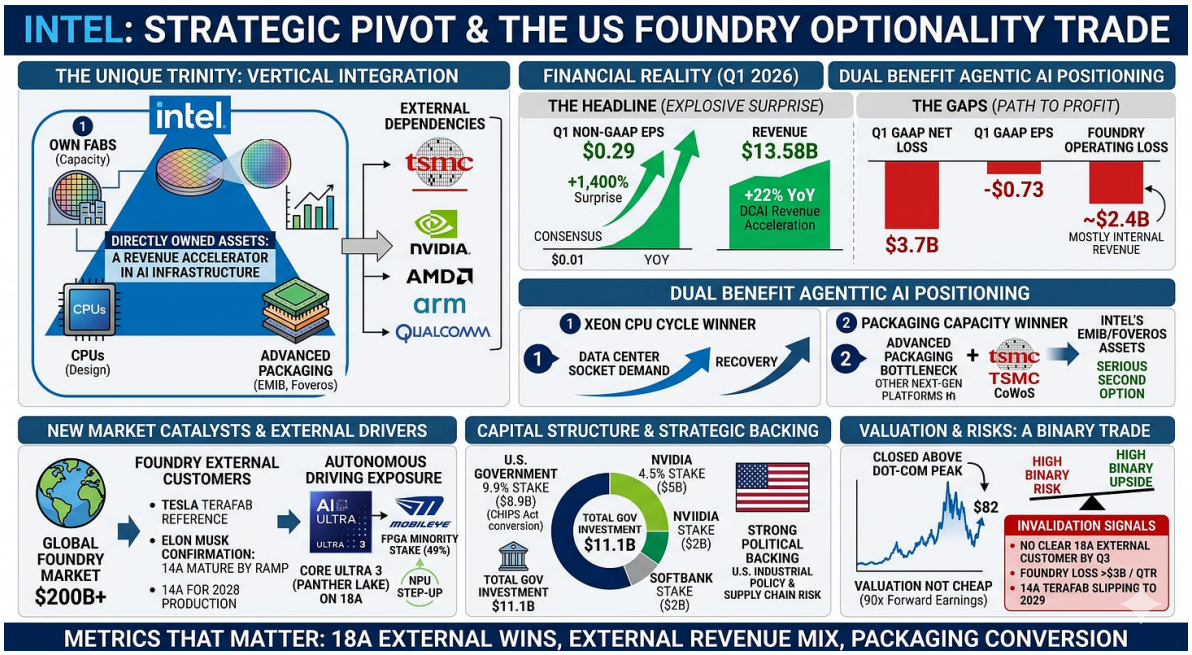
That's the lens for taking another look at Intel, AMD, NVIDIA, ARM, and Qualcomm.

For each company: current business, asset stack, positioning for the agentic AI era, and the invalidation signals that would break the thesis.



Intel's real asset isn't the CPU. **It's the only stock that owns its own fabs, its CPUs, and its advanced packaging in one company.**

AMD depends on TSMC. NVIDIA depends on TSMC. ARM only sells IP and Qualcomm has no fabs. In the AI infrastructure cycle, capacity and packaging are revenue, and Intel is the only company that owns those assets directly.



Current business: Q1 2026 non-GAAP EPS of $0.29 (versus $0.01 consensus, a 1,400% surprise) and revenue of $13.58B (versus $12.42B consensus). Headline numbers were explosive. Data center (DCAI) revenue accelerated to +22% YoY.

On a GAAP basis, Intel is still in the red. Q1 2026 net loss attributable to Intel was $3.7B, GAAP EPS of -$0.73. The profit narrative and the loss reality have to be read together. Q1 Foundry revenue came in at $5.4B, but most of it is intersegment internal revenue, with operating losses around $2.4B.

Agentic AI positioning: Intel can benefit two different ways at once.

First, Xeon CPU is a direct beneficiary of rising data center socket demand. Intel is in the recovery phase after losing share to AMD, but +22% YoY growth is the start of acceleration.

Second, advanced packaging is the new bottleneck in AI infrastructure. Next-gen platforms like NVIDIA Blackwell and Rubin lean heavily on TSMC's CoWoS capacity. As that bottleneck deepens, Intel's in-house EMIB and Foveros packaging assets get a chance to be repriced as a serious second option.

Intel is essentially the only company that can be both a CPU cycle winner and a packaging capacity winner.

New market options: The real new market is Foundry external customers. The global foundry market is roughly $200B+, with TSMC effectively running it as a monopoly. Intel is trying to build the second option.

Reports on Tesla Terafab and Intel 14A came out in early April, and Elon Musk confirmed on Tesla's Q1 earnings call that 14A would be mature enough by the time Terafab ramps (Reuters). That said, 14A is targeted for 2028 production, and the Tesla Terafab references to Intel 14A are an important signal but not yet a confirmed large-scale production contract.

Second, AI PC. Core Ultra Series 3 (Panther Lake) ships on 18A, and the NPU step-up makes it a real competitor against AMD Ryzen AI and Qualcomm Snapdragon X.

Third, autonomous driving exposure through Mobileye. Altera FPGA is no longer a controlled asset. Intel sold 51% to Silver Lake in September 2025, and the business deconsolidated. Intel still holds a 49% minority stake, so FPGA optionality hasn't disappeared completely, but it isn't a controlled asset anymore.

Cap table is another variable. The U.S. government bought a 9.9% stake for $8.9B in August 2025 (converting $5.7B of unpaid CHIPS Act funds and $3.2B of Secure Enclave commitments into equity). Total cumulative government investment now sits at $11.1B ($8.9B equity plus $2.2B in prior grants).

Separately, NVIDIA took a $5B / 4.5% stake and SoftBank added another $2B. The Trump administration is actively backing the company. Given U.S. industrial policy and supply chain risk, the strategic pressure on certain hyperscalers to seriously evaluate Intel Foundry could rise.

Valuation: After yesterday's +23.6% surge, valuation isn't cheap anymore. Per Reuters, the analyst median price target was raised to around $75, and after Q1 earnings, KeyCorp, HSBC, Jefferies, and other major analysts stepped up with PT raises in succession. At the current price of $82, the stock trades at roughly 90x forward earnings. +211% on a one-year basis, with the first close above the dot-com peak.

A 90x P/E only gets justified if the foundry turnaround executes precisely.

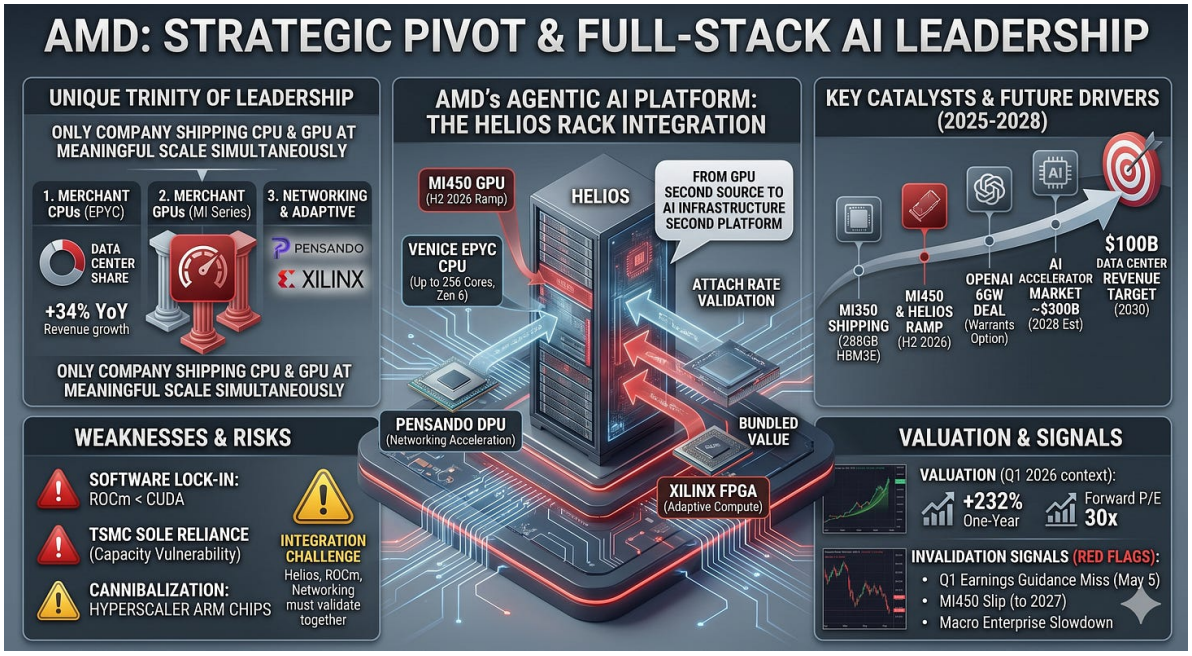
Invalidation signals: No clear 18A external customer announcement by Q3, Foundry operating loss expanding past $3B per quarter, or 14A

Terafab timing slipping to 2029. Any of these breaks the turnaround thesis.

Intel isn't a CPU cycle winner so much as a trade on the U.S. buying foundry optionality back. Binary risk is large, and so is the upside. The metrics that matter for Intel aren't Xeon revenue. They're 18A external customer wins, foundry external revenue mix, and how fast packaging capacity actually converts into customers.

AMD is **the only company shipping merchant CPUs and merchant GPUs at meaningful scale at the same time**.

NVIDIA has Grace and Vera, but those are rack-attach plays without a standalone CPU business. Intel has tried GPUs, but its data center GPU presence is thin. Only AMD has meaningful share in both EPYC (CPU) and the MI series (GPU).



Current business: EPYC's data center revenue share is 39.4% (Mercury Research, Q1 2025), nearly catching Intel in x86 server. Q4 2025 revenue of $10.3B (+34% YoY), data center segment of $5.4B, non-GAAP gross margin of 57%.

At its Analyst Day, AMD set a long-term target for major data center revenue growth over the next several years (with reporting referencing a $100B annual data center revenue target by 2030, per Reuters). The market is treating MI450 and EPYC ramp as the key drivers of growth from 2026 onward.

Agentic AI positioning: The real upside isn't EPYC share. It's whether **MI450 GPU ramp and EPYC CPU attach happen inside the same customer at the same time**.

Selling GPUs alone means going head-to-head with NVIDIA. Bundling CPU, GPU, DPU, and FPGA into one system turns AMD from a second source into a second platform. The Helios rack platform is exactly that full-stack integration play.

Since CPU memory capacity and GPU inference capability both matter

more in agentic AI, getting both from one supplier becomes attractive to hyperscalers.

The right question isn't MI450 unit volume. It's **whether MI450 customers also adopt EPYC, Pensando, and Xilinx**. If that attach rate gets confirmed, AMD gets repriced from a GPU second source into an AI infrastructure second platform. Q1 earnings on May 5 and the H2 2026 MI450 ramp window are the validation period for this thesis.

Per reporting and the roadmap, the next-gen Venice CPU is expected to scale up to 256 cores. While Intel is shifting SMT strategy on some next-gen designs, AMD is keeping SMT on Zen 6. MI350 (288GB of HBM3E memory) is shipping. MI450 and the Helios rack platform are scheduled for H2 2026. The Pensando (DPU) and Xilinx (FPGA) acquisitions add data-center networking and programmable silicon to the bundle.

New market options: Three angles.

First, AI inference GPU. Per industry estimates, the AI accelerator market grows to roughly $300B by 2028, and AMD is going for the second-option slot to NVIDIA with MI450 and the Helios rack platform. OpenAI signed a multi-year, 6GW AI chip supply agreement with AMD, and OpenAI can receive warrants for up to roughly 10% of AMD's stock. The first 1GW deployment, based on MI450, starts in H2 2026 (Reuters). Reports also suggest large-scale Meta deployment work is underway.

Second, AI PC. Ryzen AI 400 series is gaining laptop OEM traction and competes head-on with Intel Core Ultra.

Third, networking and adaptive compute. Pensando enters data center network acceleration and Xilinx layers on telecom infrastructure, automotive ADAS, and defense exposure.

Weaknesses: Software lock-in compared to NVIDIA is weak. AMD's GPU software stack (ROCm) hasn't caught CUDA. Sole reliance on TSMC is another vulnerability, and AMD trails NVIDIA in capacity competition. The trend of hyperscalers rolling their own ARM chips is a direct cannibalization risk for AMD.

The bigger weakness isn't the product portfolio. It's integrated experience. NVIDIA already sells rack-level reference architecture and a software stack as one product. For AMD to become a real second platform, MI450 performance isn't enough. Helios rack, ROCm, networking, and CPU attach all need to be validated together in real deployments.

Valuation: After dropping -20% post-January earnings, AMD pivoted on March 30 and rallied. April brought 11 straight up days (the longest streak since 2005), closing at $347.77 on April 24. A single-session +13.9% move set a new all-time high. +232% on a one-year basis, +41% in one month, forward P/E of 30x.

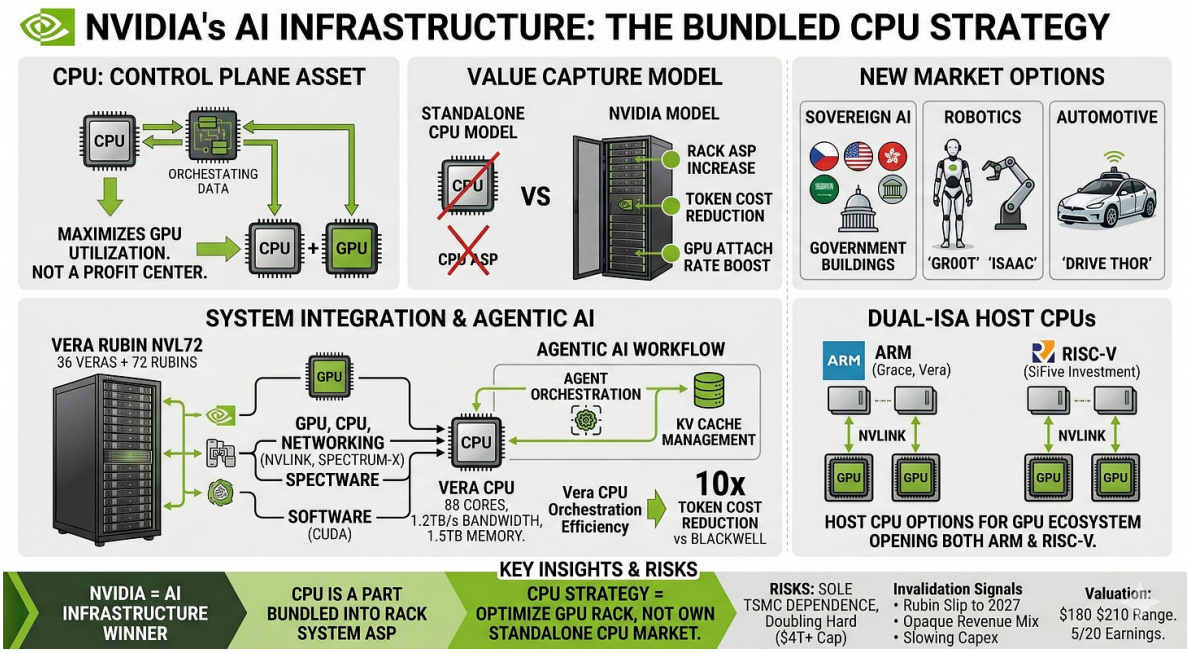
DA Davidson lifted its PT from $220 to $375, Stifel went to $320, and Barron's headlined AMD as the big winner of Intel earnings. Most of the thesis is already in the price.

Invalidation signals: Q1 earnings on May 5 missing data center growth guidance, MI450 slipping from H2 2026 to 2027, or macro slowdown delaying enterprise AI adoption by 12 to 18 months.

AMD has the cleanest fundamental story. At this price, though, “good company” and “good stock” aren’t necessarily the same trade.

NVIDIA is fundamentally a GPU company. It’s not a clean CPU cycle winner. But the role of CPU inside NVIDIA is clear. **It’s not a profit center. It’s a control plane asset that maximizes GPU utilization.**

So NVIDIA’s CPU upside doesn’t show up as CPU ASP. It shows up as rack ASP, token cost, and GPU attach rate.



Current business: Grace-class CPUs aren’t peripheral parts anymore. They’re meaningful system components inside NVIDIA’s rack architecture. A GB200 NVL72 rack carries 36 Graces and 72 Blackwells. The next-gen Vera Rubin NVL72 carries 36 Veras and 72 Rubins.

While the market keeps pricing NVIDIA as a GPU company, CPU value is getting absorbed into rack system ASP rather than showing up as a standalone product line.

Agentic AI positioning: The ability to bundle GPU, CPU, networking, and software into one system is the strongest of the five. NVLink (high-speed interconnect), Spectrum-X (networking), and CUDA (software) tie the whole system together.

As agentic AI workloads stop running purely on GPU and the CPU’s control-plane role becomes essential, NVIDIA builds that CPU itself with Vera and integrates it into the GPU rack. The Vera spec backs that integration. Per NVIDIA’s official page: 88 cores, 176 threads, memory bandwidth of 1.2TB/s, and support for 1.5TB of memory. The design intentionally scales memory up to handle the parts of agentic AI workloads that hit the CPU hardest, agent orchestration and KV cache management.

NVIDIA’s own claim is that the Rubin platform reduces inference token cost by 10x versus Blackwell, and NVIDIA says Vera CPU’s orchestration efficiency sits at the core of that improvement.

The strategy is dual-ISA. On the ARM side, Grace and Vera in-house. On the RISC-V side, an investment in SiFive (Series G, $400M, $3.65B

valuation, April 2026) and NVLink Fusion to plug RISC-V CPUs directly into NVIDIA GPUs. Rather than locking onto x86, NVIDIA is opening up both ARM and RISC-V as host CPU options inside its GPU ecosystem.

New market options: Data center AI infrastructure is the core, but expansion is happening on three fronts.

First, Sovereign AI. National governments are starting to build their own AI infrastructure (Saudi HUMAIN, UAE G42, the UK, Japan, India), and NVIDIA goes in at the system level.

Second, robotics. Project GR00T (humanoid robot foundation model), the Isaac platform, and Jetson Thor are an attempt to make NVIDIA the OS underneath industrial and service robotics.

Third, automotive. DRIVE Thor unifies autonomous driving and infotainment on a single chip.

NVIDIA's own estimate is that data center AI infrastructure capex reaches $1T+ by 2028.

Weaknesses: The real risk is sole TSMC dependence. As capacity competition intensifies, near-term revenue takes a direct hit. The CPU business is inside the GPU rack as an attach play, so NVIDIA isn't a direct beneficiary of the CPU cycle. And at $4T+ market cap, doubling is hard.

Valuation: Sideways in a $180 to $210 range for six months. April 24 close of $208.27, +101% on a one-year basis, -2% from the $212 ATH, forward P/E of 37x. May 20 earnings is the next inflection.

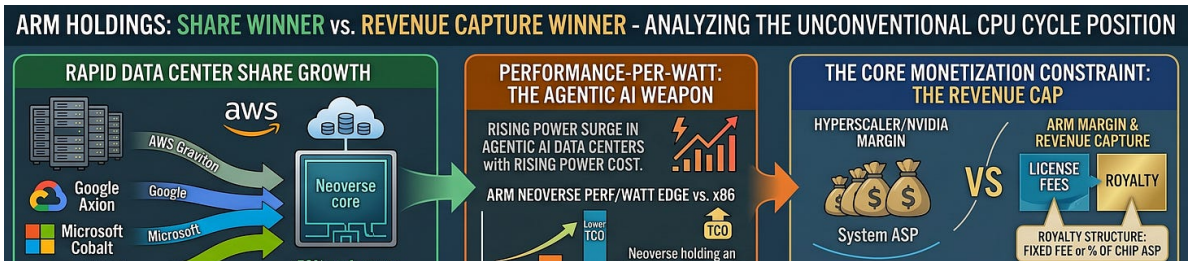
Invalidation signals: Rubin slipping from H2 2026 to 2027, continued opacity around Grace and Vera revenue mix, or meaningful signals of slowing AI infrastructure capex.

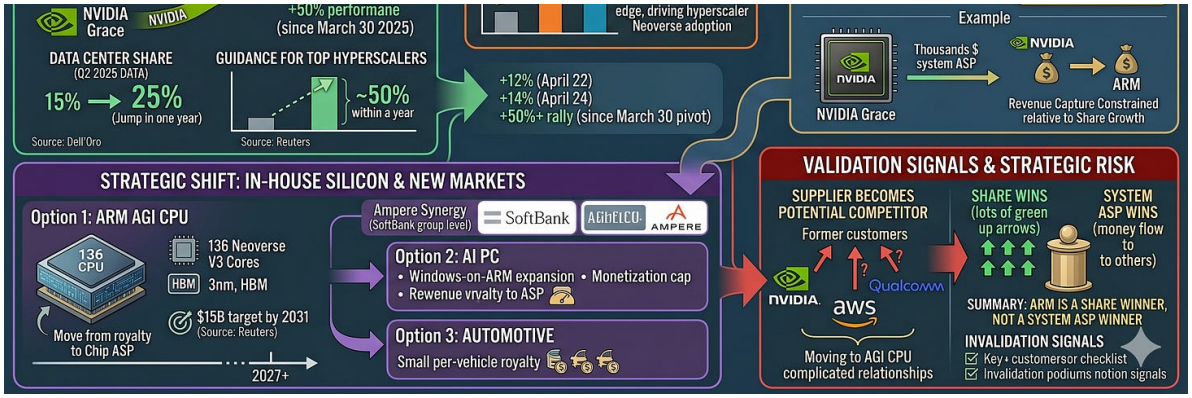
In NVIDIA, CPU isn't a separate product. It's a part bundled into rack system ASP. NVIDIA is an AI infrastructure winner, but it isn't a direct CPU cycle bet. NVIDIA's CPU strategy isn't about owning the standalone CPU market. It's about optimizing control plane and data movement inside the GPU rack. So it shouldn't be read through the same socket-thesis lens that fits Intel or AMD.

ARM sits in the strangest position in this CPU cycle. The share looks the best. The way it captures dollars is the most constrained.

AWS Graviton, Google Axion, Microsoft Cobalt, and NVIDIA Grace are all hyperscaler in-house chips built on ARM Neoverse. The power efficiency edge that comes with RISC design philosophy is a direct weapon as AI data center power costs surge.

That weapon, though, doesn't show up in ARM's revenue. It shows up in hyperscaler and NVIDIA system ASP.





Current business: Data center share jumped from 15% to 25% in one year (Dell'Oro Q2 2025). ARM has expressed an expectation that its share inside the top hyperscalers could reach roughly 50% in the near term (per Reuters reporting). Worth flagging that this is ARM's own expectation rather than measured share.

+12% on April 22, +14% on April 24 (Intel earnings spillover), +50%+ since the March 30 pivot. Strong momentum.

Agentic AI positioning: As AI data center power use surges, hyperscalers are putting performance-per-watt at the top of the stack. Once core counts grow in agentic AI, per-core power directly enters TCO, and ARM Neoverse holds a clear edge over x86 on performance-per-watt. That's why hyperscaler in-house ARM chips are scaling so fast.

The core constraint: Share gains don't convert proportionally into revenue. ARM only collects license fees and per-chip royalty. Royalty is structured as a fixed fee or a percentage of chip ASP, so even when NVIDIA sells Grace for thousands of dollars, ARM only collects the fixed per-core royalty.

The actual margin sits with NVIDIA. Data center ARM IP can carry a higher royalty intensity than mobile, so the right framing isn't "share doesn't matter." It's that revenue capture is constrained relative to how fast share is growing.

New market options: Three plays to break the revenue cap, and all of them take time.

First, in-house silicon. ARM AGI CPU (up to 136 Neoverse V3 cores, 3nm) is the first attempt to move past the royalty model and capture chip ASP directly. Per Reuters, ARM has guided that the in-house chip business could reach roughly $15B annual revenue by 2031. Initial revenue likely scales from 2027 onward.

Separately, SoftBank acquired Ampere for around $6.5B. It's not a direct ARM asset, but at the SoftBank group level, it opens an ARM ecosystem synergy story.

Second, AI PC. Windows-on-ARM expansion lifts client share, but the royalty cap on monetization stays.

Third, automotive. Almost every vehicle SoC is ARM-based, but per-vehicle royalty is small.

The moment ARM moves on AGI CPU for direct silicon revenue, the

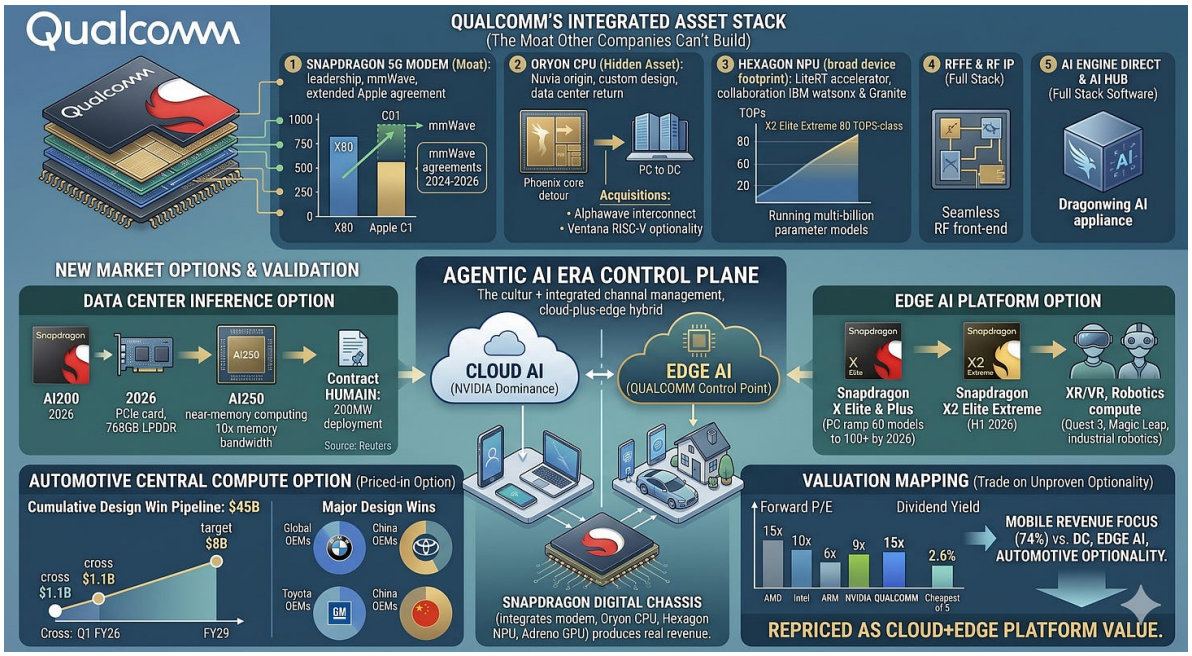
relationship with existing customers becomes complicated. NVIDIA, Qualcomm, AWS no longer see ARM as a pure IP supplier but as a potential competitor. That's both ARM's largest upside and its largest strategic risk.

Valuation: From $130s in January to roughly $234 on April 24. Forward P/E of 60x+, with the royalty model's monetization cap pressing on the multiple. Susquehanna's PT of $210 (raised April 16) has already been crossed. Most of the share narrative is already in the price.

Invalidation signals: No hyperscaler adoption announcement for ARM AGI CPU by H1 2027, SoftBank-level Ampere ecosystem synergy not converting into revenue, or in-house silicon driving licensing customers away.

ARM is most likely a share winner. It's not a system ASP winner.

Qualcomm isn't an obvious CPU rally beneficiary. AI200 and AI250 are inference accelerator cards and rack solutions, not CPUs in the strict sense. The reason for including it in the five is that the same agentic AI demand drives both CPUs and inference accelerators. And Qualcomm's real bet sits not in data center but in being a **control point for the edge AI era**.



Current business: A large share of revenue still comes from handset (smartphone SoC), and Apple is reducing dependence with its in-house modems (C1, C2). Between April 16 and 22, JPMorgan (Outperform to Neutral), BNP Paribas (Outperform to Neutral), and Barclays (Underweight reinitiation, PT $130) all downgraded the stock.

YTD -13% while the other four names rallied. April 24 +11% on Intel earnings spillover, closing around $148 to $149.

Asset stack, the part the market is missing:

First, modem leadership is a massive moat. Snapdragon X80 5G Modem holds the lead in premium 5G modems. mmWave 5G support is something Apple's C1 doesn't have. Apple extended its agreement to

keep using Qualcomm's 5G Modem-RF System on iPhones launched between 2024 and 2026 (Reuters). Despite throwing enormous R&D dollars at its in-house modem, Apple still can't fully replace it.

Modem isn't a single component. It's a full stack of RF front-end (RFFE), power management, and 5G standard IP wrapped together. As 6G arrives and communications and AI inference share the same chip, modem leadership gets repriced again.

Second, the **Oryon CPU produced by the Nuvia acquisition** is a hidden asset. The timeline tells an interesting story.

Nuvia was founded in 2019 by Gerard Williams III, the former Apple chip design lead, and **its original product line was a high-performance ARM core for data center (codename Phoenix)**. Qualcomm acquired the company for $1.4B in 2021 and went through a licensing dispute with ARM (ARM demanded license renegotiation, Qualcomm refused) before deploying the Phoenix core as Oryon for PC use first. Oryon is the core inside Snapdragon X Elite and Plus and Snapdragon 8 Elite. By Qualcomm's own statement, Oryon uses less than 1% of ARM's original IP, making it nearly a fully custom core (Tom's Hardware).

In 2025, Qualcomm pivoted back to data center. It hired the former lead architect of Intel Xeon, then closed the Alphawave Semi acquisition ($2.4B, completed December 2025, a quarter ahead of plan), bringing high-speed interconnect IP that's central to data center inference. Former Alphawave CEO Tony Pialis joined as head of Qualcomm's data center business. The Ventana Micro Systems acquisition (December 2025, terms not disclosed) added RISC-V optionality.

Oryon is a core that was originally designed for data center, took a detour through PC, and is now coming back to data center.

Third, **Qualcomm's broad device footprint in edge AI**. Cloud-plus-edge hybrid is the consistent vision Qualcomm CEO Cristiano Amon has been pushing at Snapdragon Summit. The reasons to do edge processing are clear. Latency (autonomous driving, voice response), privacy (personal data), context (sensor data), and cost (ballooning cloud inference cost).

Qualcomm already has a full edge AI stack: Hexagon NPU, Snapdragon X2 Elite Extreme (80 TOPS-class NPU), Snapdragon 8 Elite Gen 5, AI Engine Direct, AI Hub, and the Dragonwing AI On-Prem Appliance. The trajectory is toward running multi-billion parameter models on-device. Google launched a LiteRT Qualcomm AI Engine Direct Accelerator. IBM watsonx and Granite models run on Snapdragon as part of an ongoing collaboration.

Agentic AI positioning: In data center, an inference TCO option (AI200 / AI250). At the edge, one of the broadest device footprints among players positioned for agentic AI agents.

Two things the market hasn't priced.

First, if agentic AI moves to a cloud-plus-edge structure, the edge control point becomes revenue. Smartphones, AI PC, automotive, XR (such as Meta's smart glasses), and IoT, Qualcomm already has meaningful silicon footprint across all of them.

Second, integrating modem, Oryon CPU, Hexagon NPU, and RFFE on a single chip is something other companies can't easily replicate. That's why Apple's transition to its own modem hasn't eliminated Qualcomm modem dependence in the short term, and why most OEMs can't internalize CPU, modem, RF, NPU, and OS integration the way Apple can.

New market options: Three angles.

First, **data center inference option**. Per industry estimates, the inference chip market grows to roughly $400B by 2030. AI200 (2026) and AI250 (2027) target inference TCO directly. AI200 is a PCIe card with 768GB of LPDDR per card, designed for mid-scale inference. AI250 uses near-memory computing to target effective memory bandwidth above 10x. The first customer, HUMAIN (an entity under Saudi Arabia's sovereign wealth fund), signed a 200MW deployment contract.

Second, **edge AI platform option**. Snapdragon X Elite and Plus, and Snapdragon X2 Elite Extreme (H1 2026 launch) are driving Windows-on-ARM ramp. Roughly 60 OEM models are in production, with 100+ expected by 2026 (Microsoft Surface, Dell, HP, Lenovo, Asus, Acer). XR and robotics extend the same category. Snapdragon AR/VR platforms and robotics compute add another axis to edge AI expansion.

Combining AI PC, XR, and IoT, Qualcomm has one of the broadest edge footprints heading into the cloud-plus-edge era.

Third, **automotive central compute option**. This is the new market that's already producing real revenue. Automotive revenue went from $984M (+21% YoY) in Q3 FY25 to crossing $1.1B in Q1 FY26. The cumulative **design win pipeline is $45B**. The company guides $8B in automotive revenue by FY29.

Snapdragon Digital Chassis is already in 50+ vehicle models. Cumulative shipments of Snapdragon Ride-based ADAS SoCs are at the million-unit level. Major design wins include BMW Neue Klasse (a 50:50 JV, debuting in the iX3), Mercedes-Benz, Toyota RAV4, GM, Sony-Honda, Stellantis (Leapmotor D19), Volvo EX90, and Hyundai Mobis, covering most major global OEMs. In China, design wins include Li Auto, NIO, Zeekr, Great Wall, and Chery.

The implication is two-fold. The revenue multiple itself can rerate from a mobile company to an automotive plus AI company. And as automotive SoC evolves into a central computing platform that integrates ADAS and infotainment, Qualcomm's full stack of modem, Oryon CPU, Hexagon NPU, and Adreno GPU fits the requirement precisely.

Valuation: Forward P/E of 15x, the cheapest of the five (AMD 30x, Intel 128x, ARM 60x+, NVIDIA 37x). Dividend yield of 2.6% on top. The market is looking only at slowing mobile and pricing in nothing for data center, edge AI, or automotive optionality.

This isn't a trade about buying a great company expensive. It's a trade about buying unproven optionality cheap. Qualcomm's upside doesn't sit in AI200 alone. The bigger question is whether modem, Oryon CPU, Hexagon NPU, RF, and automotive footprint get repriced as one platform value once agentic AI runs on cloud-plus-edge rather than cloud-only.

SiFive and Tenstorrent are both private. Direct exposure is hard, but the thesis is worth tracking.

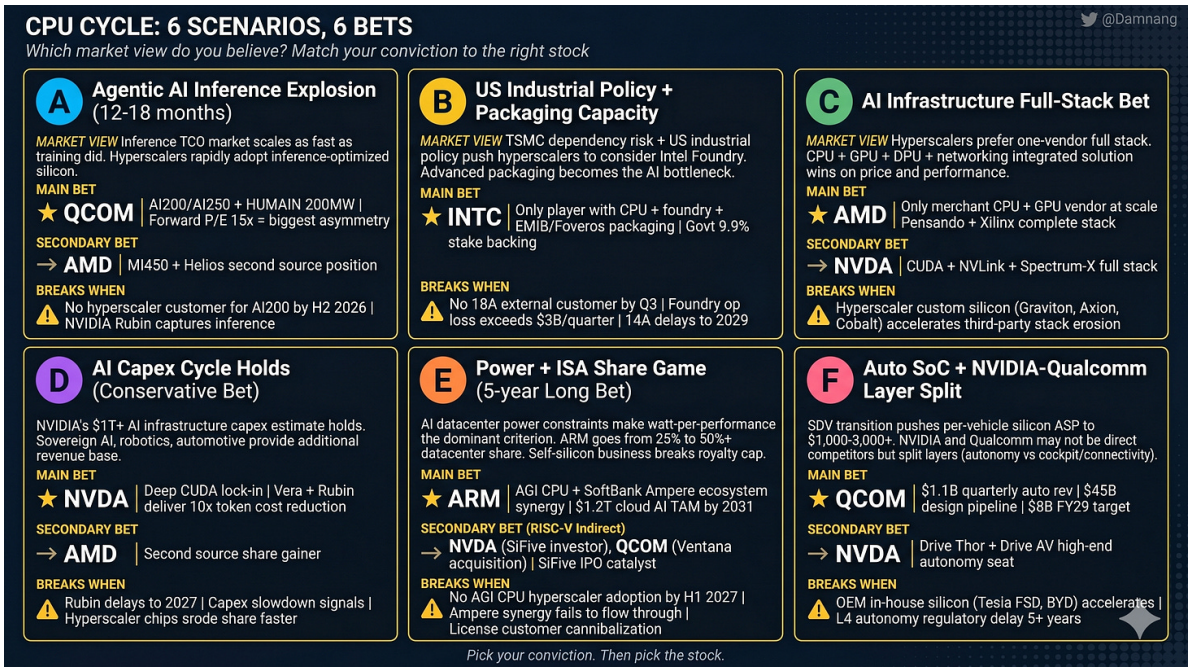
SiFive raised $400M in a Series G in April 2026. Valuation at $3.65B, with NVIDIA participating as an investor. The structure is that SiFive P870-D (data center RISC-V CPU, scaling to 256 cores) connects directly to NVIDIA GPUs over NVLink Fusion.

Tenstorrent is led by Jim Keller (former AMD, Apple, Tesla CPU architect). The roadmap includes Black Hole and Grendel, RISC-V-based AI chips. Tenstorrent's value is the ability to integrate a RISC-V CPU and an AI accelerator on one chip.

Direct exposure is hard because both companies are private. Indirect exposure runs two ways: NVIDIA (SiFive investor, NVLink Fusion, dual-ISA strategy) and Qualcomm (Ventana Micro acquisition). NVIDIA's position in SiFive can also be read as a signal that an IPO isn't far off (my own read).

Forcing a single ranking doesn't make sense. Investors have different time horizons, different views, different risk tolerances. Looking at the same data, one person sets up a 12-month catalyst trade, another holds for five years, another buys binary turnaround risk in option size.

What works better is splitting the market view into six scenarios and naming the stock that fits each one most cleanly. The way to use the framework is to first decide which scenario you have the most confidence in, then look at the stocks that fit it.



Market view: Zinsner's GPU-to-CPU ratio comments get confirmed quantitatively in 2026 to 2027 data, and the inference TCO market grows as large as the training market. AI infrastructure capex shifts from training to inference, and chips optimized for inference quickly accumulate hyperscaler customers.

Main bet: **Qualcomm**. AI200 (2026) and AI250 (2027) target inference TCO directly, and the HUMAIN 200MW deployment is the first customer

signal. Forward P/E of 15x with the market only pricing the mobile business sets up the largest asymmetry if the scenario plays out. Key catalysts: Q2 FY26 earnings on April 29 and hyperscaler customer announcements in H2 2026.

Secondary bet: **AMD**. MI450 and the Helios rack platform can take the second-option slot to NVIDIA in the inference market. Weaker ROCm software lock-in versus CUDA limits how much of NVIDIA it can replace.

What breaks the scenario: AI200 hyperscaler customer wins not landing by H2 2026, agentic AI adoption delayed by 12 to 18 months, or NVIDIA Rubin sweeping the inference market.

Market view: AI infrastructure capacity itself becomes the bottleneck, and U.S. industrial policy creates strategic pressure for hyperscalers to seriously evaluate Intel Foundry. As sole TSMC dependence becomes a more visible risk, the second foundry option gets repriced. Advanced packaging becomes the real bottleneck for next-gen AI chips, and Intel's own EMIB and Foveros assets get recognized as a separate profit pool.

Main bet: **Intel**. The only stock with fab, CPU, and advanced packaging under one roof. Political backing is in place: 9.9% U.S. government stake, 4.5% NVIDIA stake, $2B SoftBank investment, and cumulative government commitment of $11.1B. Tesla Terafab and Intel 14A reporting is already out, and 18A external customer announcements would be the trigger. Binary risk is large, and so is the upside.

What breaks the scenario: No clear 18A external customer announcement by Q3, Foundry operating loss expanding past $3B per quarter, or 14A Terafab timing slipping to 2029.

Market view: Buying every chip type for an AI infrastructure rack from one vendor is the optimal strategy for hyperscalers. A fully integrated solution wins on price and performance versus buying CPUs, GPUs, and DPUs separately.

Main bet: **AMD**. The only company shipping merchant CPU (EPYC at 39.4% share) and merchant GPU (MI series) at meaningful scale at the same time. Add Pensando (DPU) and Xilinx (FPGA), and AMD owns more of the core chip categories that go into an AI infrastructure rack than anyone else. Helios rack is the attempt to bundle the full stack into one system. The price already reflects most of this, so betting this scenario means waiting patiently for a pullback after Q1 earnings on May 5.

Secondary bet: **NVIDIA**. CUDA lock-in plus NVLink plus Spectrum-X builds a different version of the full stack. AMD takes the second-option slot, NVIDIA holds the first.

What breaks the scenario: Hyperscalers accelerating their move to in-house silicon (AWS Graviton, Google Axion, Microsoft Cobalt) and shrinking demand for third-party full-stack solutions overall.

Market view: NVIDIA's own estimate of $1T+ in AI infrastructure capex by 2028 holds up roughly, and even with macro slowdown or hyperscaler in-house chip transitions, NVIDIA holds its market position. New markets like Sovereign AI, robotics, and automotive add to the base.

Main bet: **NVIDIA**. AI infrastructure winner, with deep CUDA lock-in and

full integration of GPU, CPU, and networking. Vera plus Rubin sets up the next cycle by reducing inference token cost by 10x. The market cap is large enough that doubling is hard, but the base is solid. The six-month range itself creates some asymmetry.

Secondary bet: **AMD**. Secondary winner in the second-option slot.

What breaks the scenario: Rubin slipping from H2 2026 to 2027, meaningful deceleration signals in AI infrastructure capex, or hyperscaler in-house silicon eroding NVIDIA share faster than expected.

Market view: AI data center power costs hit a hard ceiling, hyperscalers anchor on performance-per-watt as the top criterion, and ARM's data center share moves from 25% to above 50%. At the same time, the in-house chip business (AGI CPU) and SoftBank-level Ampere ecosystem synergies break the royalty model cap, and the revenue multiple gets repriced. RISC-V's data center share also climbs to measurable levels.

Main bet: **ARM Holdings**. In a 12 to 18 month window, the royalty cap pushes on the multiple. From 2027 to 2028 onward, as in-house silicon revenue starts flowing, ARM becomes a different company. ARM's own estimate puts the cloud AI TAM at $1.2T by 2031.

Indirect RISC-V exposure: SiFive and Tenstorrent are private, so direct exposure is hard. Indirect exposure runs through NVIDIA (SiFive investor) and Qualcomm (Ventana Micro acquisition). A SiFive IPO would be the next catalyst.

What breaks the scenario: No hyperscaler adoption announcement for ARM AGI CPU by H1 2027, SoftBank-level Ampere synergies not flowing through to ARM revenue, or in-house silicon driving licensing customers away. In a 12-month window, the share narrative is mostly priced in, leaving limited room for further upside.

Market view: As cars transition to software-defined vehicles (SDV), per-vehicle silicon content grows dramatically. ADAS and infotainment stop being separate ECUs and merge into one central compute platform, creating a per-vehicle SoC market of $1,000 to $3,000+. As autonomous driving moves from L2+ to L4, compute requirements jump again. Industry estimates put this market above $50B by 2030.

The market is misreading one thing. Treating NVIDIA and Qualcomm as direct competitors in automotive misses what's actually happening. There's a stack-division dynamic running in parallel.

The Volvo EX90 already ships with NVIDIA Drive handling autonomous driving and Qualcomm Snapdragon Cockpit handling infotainment and cabin systems. NVIDIA is strong in ADAS and autonomous driving stacks (Drive AV, Drive Thor). Qualcomm is strong in cockpit, modem, vehicle connectivity, and the full vehicle connectivity stack including RFFE.

Both companies' silicon already shipping inside the same car is the current reality.

One step deeper, an interesting structure emerges. NVIDIA Drive Thor takes the autonomous compute headroom while Qualcomm Oryon CPU plus Hexagon NPU plus 5G modem takes vehicle connectivity and

cockpit. From the OEM's seat, that's a rational way to secure two second sources at once. Both run on ARM (Drive Thor uses ARM Neoverse, Oryon is built under an ARM Architecture License), so software integration is feasible.

As the automotive compute stack standardizes, it can move from head-on competition to layer-by-layer specialization. That's the core of this scenario.

Main bet: **Qualcomm**. Automotive revenue has already crossed $1.1B per quarter, with a $45B design win pipeline and FY29 guidance of $8B. Market valuation still treats it as a mobile company, so as automotive revenue mix climbs meaningfully, the multiple has room to rerate.

Secondary bet: **NVIDIA**. The high-end seat in autonomous driving and ADAS compute. Drive Thor and Drive AV launch on Mercedes-Benz CLA and enter U.S. roads. If the Qualcomm-NVIDIA layer division holds, both end up winners.

What breaks the scenario: Automotive SoC market getting eaten by OEM in-house silicon (Tesla FSD chip, BYD's own silicon), L2+ to L4 transition delayed five-plus years by regulation and safety concerns, or the layer division between NVIDIA and Qualcomm collapsing into one player taking the full stack. A deep automotive cycle pullback from macro slowdown also breaks it.

Share

CPU demand is coming back. The dollars don't follow share order. The winner of this cycle isn't the company that sells the most CPUs. It's the company that converts the CPU bottleneck into its own margin structure most effectively.

Recapping the scenarios. Betting on agentic AI inference exploding makes Qualcomm the largest asymmetry. Betting on U.S. industrial policy and foundry optionality makes Intel the only choice. Betting on AI infrastructure as a full-stack play makes AMD the company holding both ends. Betting conservatively on the AI capex cycle makes NVIDIA the base. Betting on a five-year ISA share game makes ARM the long hold. Betting on automotive SoC as the next big cycle makes Qualcomm and NVIDIA both winners through layer specialization rather than head-on competition.

Decide which scenario carries the most confidence first. Without that anchor, the question of where to buy doesn't have an answer.