α

Meta's Obsession With AI Is Great News For AMD

Jul. 09, 2025 1:49 AM ET | **Advanced Micro Devices, Inc. (AMD) Stock**, **AMD:CA Stock** | AMD, AMD:CA

LL Insights
501 Followers

Summary

Meta's embrace of AMD's MI300X, alongside new deployments at Microsoft, Oracle, and DigitalOcean, makes AMD the prime second source to Nvidia as AI compute demand accelerates.

The MI300X's 192 GB HBM3e, chiplet cost advantages, and maturing ROCm software stack uniquely enable massive LLM inference and strengthen AMD's long-term moat.

Robust Q1 2025 results - 36% revenue growth, expanding gross margins, and a net-cash balance sheet - underscore AMD's operational leverage and capacity to fund further AI expansion.

Challenges include Nvidia's imminent Blackwell launch, tight HBM supply, and Intel's lower-cost Gaudi 3, any of which could compress AMD's pricing power and slow share gains. Ask ChatGPT.

COM & O

Thesis

AMD's (NASDAQ:AMD) (TSX:AMD:CA) strategic position as the memory-dense second source for hyperscale AI accelerators, now endorsed by both Meta (META) and OpenAI, creates a multiyear, high-margin revenue runway that the market still underappreciates. Our core thesis is simple: by parlaying MI300-series wins at Meta, Microsoft (MSFT), Oracle (ORCL), and OpenAI into a steady double-digit share of the datacenter accelerator market, AMD can more than triple its datacenter revenue by 2027, lift consolidated gross margin above 55%, and compound free cash flow at a pace that forces a valuation rerating. Meta has ignited an arms race for artificial-general-intelligence talent, dangling bonuses rumored to approach $100 million to lure senior researchers away from OpenAI, DeepMind, and Anthropic into the newborn Meta Superintelligence Labs. That hiring spree signals an aggressive road map for ever-larger Llama models and multimodal agents that crave dramatically more on-board memory than mainstream GPUs can economically provide.

To meet that requirement, Meta has standardized on AMD's Instinct MI300X accelerator for its 405-billion-parameter Llama 3.1 model, ordering roughly 170,000 cards according to Omdia supply-chain checks. Each MI300X carries 192 GB of HBM3e and 5.3 TB/s of bandwidth, enough to hold an entire Llama 3.1 shard inside a single device, whereas Nvidia's (NVDA) H100 tops out at 80 GB. By anchoring the largest open-weight model to AMD silicon, Meta not only validates the CDNA 3 architecture but also gives ROCm the at-scale customer feedback loop it previously lacked. In June 2025, OpenAI publicly committed to deploying AMD's forthcoming MI350 accelerators and co-designing future MI450 silicon, adding a second marquee customer that magnifies software telemetry and volume leverage for ROCm.

Meta's commitment converts into billions of dollars of annual revenue potential for AMD because AI workloads scale almost linearly with context length and parameter count. Mark Zuckerberg has publicly promised to pursue a "personal super-intelligence" that will require successive Llama generations an order of magnitude larger than today's versions, bending the demand curve for memory-dense GPUs upward in AMD's favor. In parallel, Microsoft, Oracle, Samsung (OTCPK:SSNLF), and DigitalOcean (DOCN) have begun rolling out MI300X instances of their own, broadening the hyperscale footprint and diversifying the revenue base. The result is a virtuous adoption cycle: marquee wins attract software optimization, which in turn lowers friction for additional customers, undermining Nvidia's lock-in advantage.

The new dynamic comes at a moment when AI accelerator supply is structurally constrained. With HBM capacity tight and Nvidia's backlog stretching well into 2026, cloud providers want a credible second source. AMD's chiplet strategy lets it stitch together more memory channels at lower marginal silicon cost than monolithic rivals, giving it a bill-of-materials edge that hyperscalers can measure in the low tens of thousands of dollars per GPU at 192 GB densities. That economic spread compounds across tens of thousands of nodes and becomes impossible to ignore when power and floor-space budgets are capped.

Finally, software no longer looks like an insurmountable moat. ROCm 6.2 introduced native vLLM and Bits-and-Bytes support, FP8 kernels, and new profiling tools, narrowing the efficiency gap with CUDA for both training and inference. Meta's own disclosure that it is serving production Llama 3.1 traffic exclusively on MI300X clusters indicates that the stack is already good enough for the toughest real-time inference workloads. As more open-source repos accept HIP pull requests, switching costs will drip lower quarter after quarter.

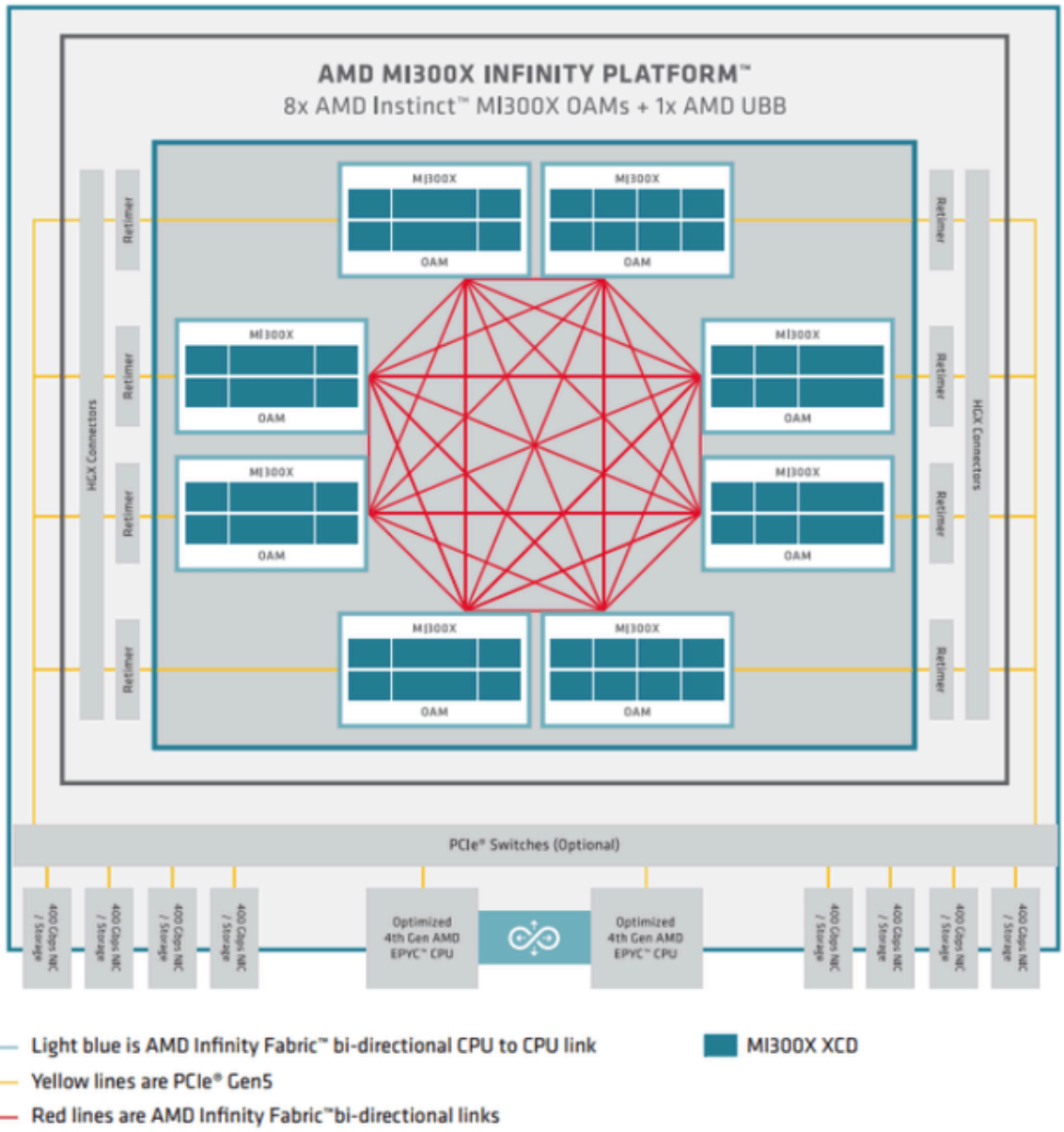
AMD CEO Dr. Lisa Su presenting the Instinct MI300X accelerator at a 2025 company event. (AMD)

Architectural Advantage

The Instinct MI300X is the first data-center GPU to pair 192 GB of HBM3e with 5.3 TB/s of bandwidth in a single package, more than doubling both the capacity and throughput of Nvidia's flagship H100 PCIe card. Large-language-model inference performance is often gated by memory rather than raw compute, so the extra VRAM allows an eight-way Grand Teton server to fit the entire 405-billion-parameter Llama 3.1 without partitioning, cutting latency while freeing network fabric for actual user tokens. Fewer hops mean higher throughput per watt, a metric hyperscalers use to size datacenter expansions.

CDNA 3's FP8 matrix engines also post impressive raw math. AMD claims 2.6 petaFLOPS of FP8 at 750 W TDP, and independent MLPerf submissions show parity or better versus H100 on GPT-J and BERT-Large inference at similar power envelopes. ServeTheHome's deep dive from Hot Chips corroborated the finding, noting 60% more bandwidth and double the HBM footprint as critical advantages when quantization is not an option. Because memory capacity scales poorly with lithography shrinks, AMD's decision to modularize HBM stacks delivers a forward cost curve that steepens slower than Nvidia's monolith, giving it room to underprice competitors once supply normalizes.

Price transparency is already emerging. While Nvidia's H100 still retails for roughly $30,000 per card and Intel's eight-way Gaudi 3 board lists for about $125,000, dealers quote MI300X units in the low- to mid-twenties, enabling hyperscalers to cut dollar-per-gigabyte costs by more than half at the rack level. That arithmetic becomes decisive in inference, where gross-margin sensitivity to opex outweighs time-to-train metrics. These architectural wins directly reinforce our thesis that AMD's memory leadership, not raw FLOPS, will dictate hyperscaler wallet share.



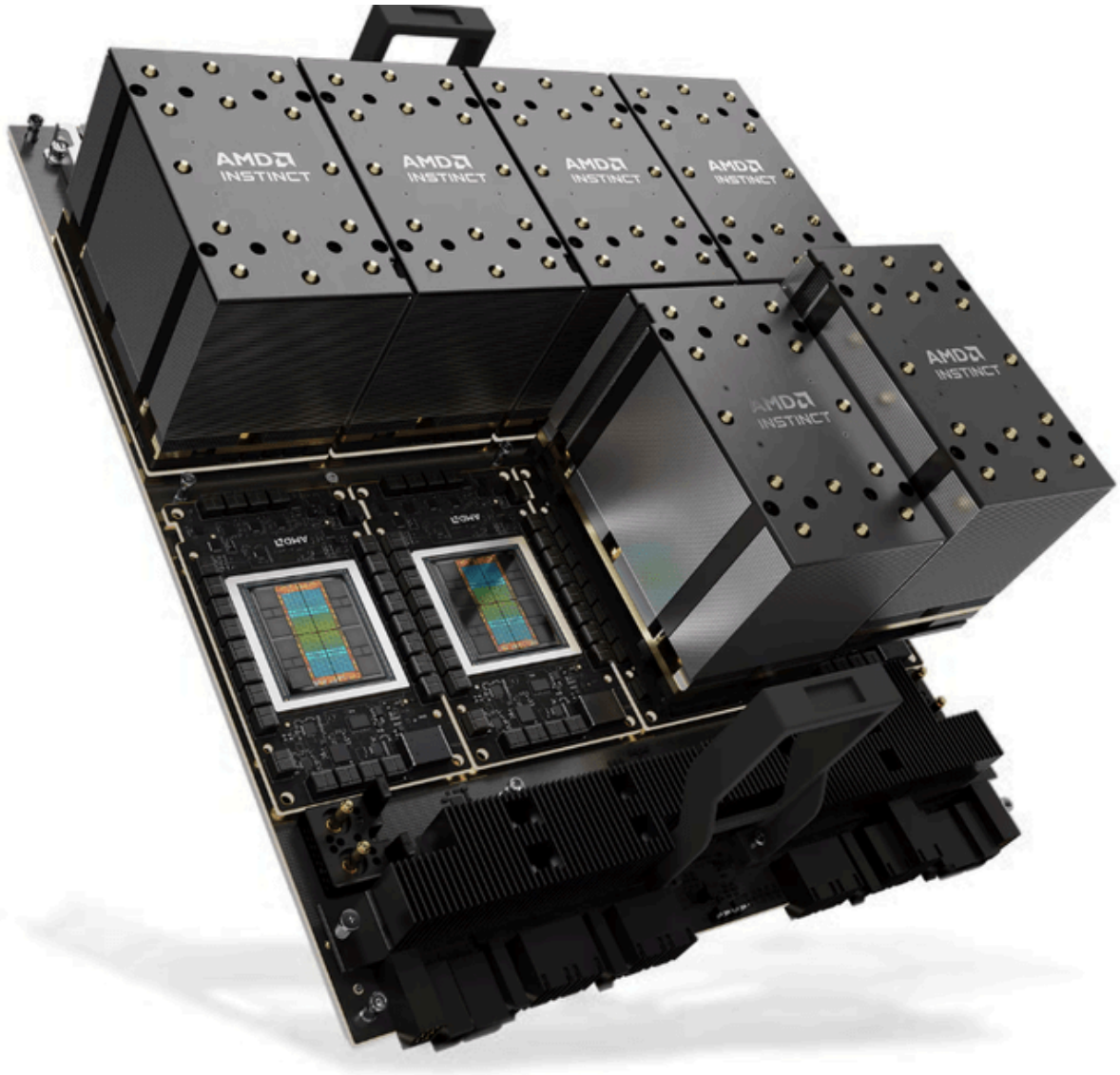
AMD platform diagram showing eight fully-connected MI300X OAM modules (AMD ROCm Docs)

Hyperscaler Momentum and the Software Flywheel

Seven of the top ten AI companies are now publicly deploying MI300-based systems, according to AMD's July "Advancing AI" event, and the firm reiterated a pathway for its data-center AI revenue to move from roughly $5 billion in 2024 to "tens of billions" by the decade's end. Oracle Cloud Infrastructure offers GPU.MI300X.8 shapes, DigitalOcean provides bare-metal MI300X for startup workloads, and Dell ships PowerEdge XE9680 nodes optimized for Llama 4 series models. Each of these deployments enlarges the ROCm telemetry pool, producing performance counters that AMD can feed back into kernel autotuning.

ROCm 6.x has landed crucial ecosystem integrations. vLLM now launches on HIP without code changes, FP8 kernels accelerate both training and cache-heavy inference, and Omnitrace-based profilers provide developers with end-to-end visibility across CPUs, GPUs, and NICs. The result is a cumulative-advantage loop: better tooling attracts more developers, whose optimizations in turn improve benchmark scores, further reducing perceived risk for procurement teams.

Interconnect openness compounds those benefits. AMD is a founding member of the UALink consortium, an industry effort to define an open alternative to Nvidia's proprietary NVLink for intra-server GPU fabrics. The MI300X already supports UALink 1.0 signalling and will inherit 2.0 bandwidth once UCI-Express retimers hit production servers. This ensures future backward compatibility without vendor lock-in. Hyperscalers that fear vendor concentration see standards-based fabrics as strategic insurance. The breadth of adopters, now explicitly spanning Meta to OpenAI, thus validates the central thesis and reduces perceived single-customer risk.



MI300X board showing CDNA 3 dies (AMD)

Market Opportunity

Industry analysts peg the data-center AI accelerator market at more than $500 billion by 2028, a 10-fold jump from 2023 levels. Even if Nvidia currently commands about 80% share, the residual pool still represents an unprecedented addressable market for alternative suppliers. Omdia estimates that MI300X shipments across Meta, Microsoft, Oracle, and TensorWave already surpassed 327,000 units in 2024. Meta alone accounts for roughly half of that total, and its publicly disclosed road map suggests annual refresh cycles that grow accelerator counts in step with Llama model parameters.

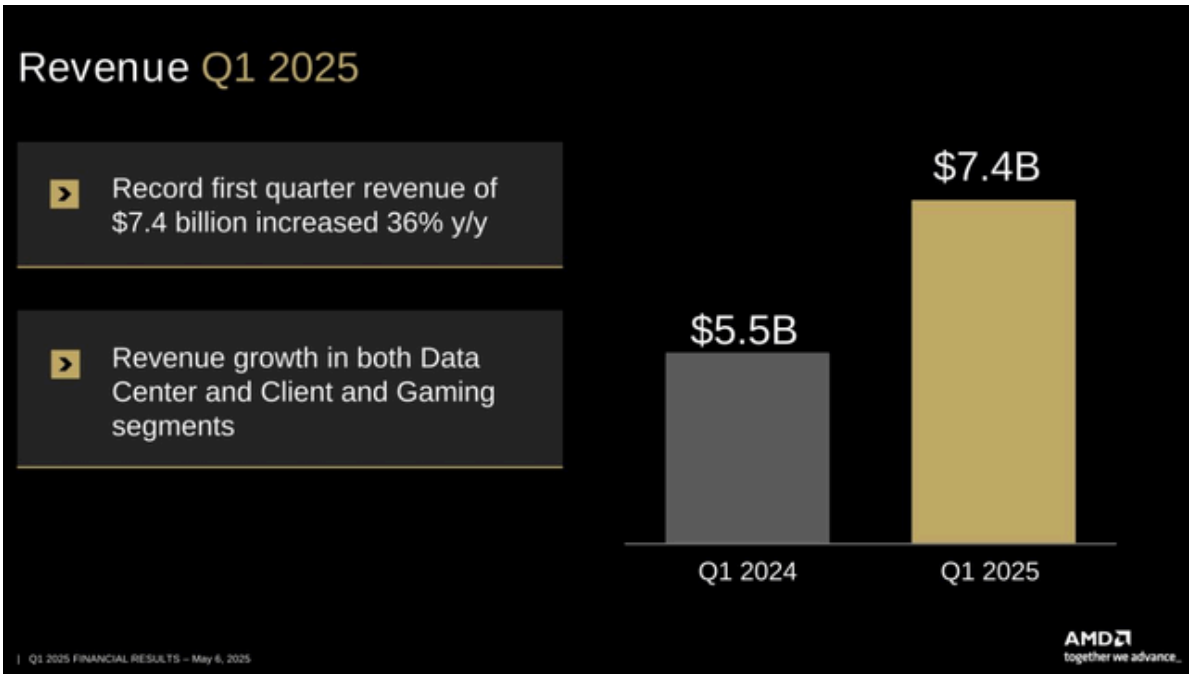
The modular nature of AMD's chiplets also opens adjacent verticals. The MI300A variant, blending Zen 5 CPU tiles with CDNA 3 GPUs, has been selected for the Department of Energy's exascale-class El Capitan supercomputer, a deployment that reinforces AMD's credibility in high-performance computing and cross-pollinates software optimizations back into commercial libraries. AMD recently acquired ZT Systems to offer full-rack AI infrastructure, capturing a richer share of system value and positioning itself to bundle CPUs, GPUs, NICs, and UALink switches in a single purchase order.

Crucially, hyperscalers want redundancy. Google's internal TPU road map is proprietary, Amazon leans on its own Trainium and Inferentia ASICs, and Meta is testing its "Artemis" AI accelerator, yet all three continue to buy merchant GPUs as timing safe-nets. By achieving functional equivalence to CUDA on core inference frameworks, AMD becomes the default second source in any multi-vendor stack.

Financials

AMD exited the March quarter with GAAP revenue of $7.44 billion, up 36% year over year, and a 50% gross margin that expanded 3% points despite mix shifts. Within that total, the Data Center segment contributed $3.7 billion (+57% YoY), Client and Gaming delivered $2.9 billion (+28% YoY), and Embedded added $0.8 billion (-3% YoY). Operating income for Data Center reached $932 million, a 25% margin, underscoring the segment's leverage. Non-GAAP gross margin reached 54%, underlining management's ability to price newer products above corporate average. GAAP operating income surged to $806 million from just $36 million a year earlier, proving that volume leverage in the Datacenter and Client segments is translating into real operating muscle.

Looking beyond the March quarter, consensus models already embed a step-function move: Wall Street expects roughly $37-38 billion of revenue and about $5.9 of non-GAAP EPS in FY 2026. More bullish sell-side scenarios that assume an MI350 ramp at Microsoft and Meta push FY 2027 revenue past $45 billion and EPS into the $8-$9 range. Management's May 2025 call guided that Q2 revenue would still grow 27% YoY even after a roughly $700 million export-control headwind and reiterated confidence that full-year datacenter sales will grow at a "strong double-digit" rate. Datacenter mix is the fulcrum (historical results show that every five-point shift toward accelerators has lifted corporate gross margin by roughly 100-150 basis points), providing operating leverage that can drive free cash flow well above $15 billion by 2027 even after aggressive HBM capacity reservations.



Q1 2025 revenue growth slide (AMD)

Data Center Outlook

If AMD secures even a 15% slice of the $500 billion accelerator market projected for 2028, datacenter revenue alone could exceed $50 billion (seven times the 2024 baseline) with incremental margins north of 60%. Such a ramp would push consolidated earnings power well beyond current consensus and support sustained multiple expansion. Management reinforced this trajectory on the May call, stating that datacenter-AI revenue could scale to "tens of billions of dollars" annually by the latter half of the decade. Adding the Q1 starting point of $3.7 billion and a double-digit YoY growth cadence implies a clear path toward that target, even after factoring in the $1.5 billion export-license headwind for MI308 shipments.

Valuation

At roughly 8x trailing price-to-sales and 47x trailing earnings, AMD trades at a steep discount to Nvidia's 25x sales and 51x earnings multiples despite a faster datacenter growth profile. Intel, by contrast, sits at about 1.8x sales and lacks a proven high-bandwidth GPU offering, underscoring the valuation spread that the market assigns to credible AI execution. Applying a discounted-cash-flow framework that builds off a 2027 free-cash-flow estimate of $15 billion (12% WACC, 3% terminal growth) produces an equity value of roughly $400 billion, or about $245 per share (40% above the current price) even before assigning option value to MI400-series rack-scale systems. A rerating toward half of Nvidia's price-to-sales multiple would yield a similar upside, corroborating the DCF output. A rerating toward even half of Nvidia's price-to-sales multiple would imply material upside for AMD as MI300 (and soon MI350) revenues compound through 2026. A rerating toward half of Nvidia's price-to-sales multiple would yield a similar upside, corroborating the DCF output.

Challenges

Nvidia is far from standing still. Hopper successor "Blackwell" promises dual-die GPUs with 192 GB HBM3e on each die and a proprietary NVLink 5 fabric that might close AMD's memory advantage before year-end. Should Nvidia ship in volume, the pricing umbrella supporting MI300X could narrow quickly. CUDA also remains a default setting for many AI researchers. While ROCm has closed much of the gap, performance irregularities still show up in edge cases such as irregular-sparsity kernels or bespoke transformer variants. Each friction point delays workloads from migrating.

Supply resiliency also bears watching. HBM3e relies on a handful of packaging nodes at TSMC and Samsung; any yield hiccup could restrict AMD's ability to fulfill Meta's and Microsoft's aggressive forecasts. Export-control shocks illustrate another vulnerability: the $800 million inventory write-down tied to new U.S. licensing requirements demonstrates how geopolitical risk can swing margins and working capital in a single quarter.

You can’t control the market. But you can

Enter email address

Continue with Google

control how you invest.

Breaking stock news is now free. Create your account to stay informed—and explore the insights behind every move.

Create Free Account or Continue with Apple

By creating an account using any of the options above, you agree to the Terms of Use & Privacy Policy

Already a member? Sign in

will likely hold the crown in absolute unit share, AMD's differentiated memory footprint, cost structure, and growing software acceptability position it as the indispensable second supplier in a market that punishes single-source risk. Patient investors prepared for volatility stand to benefit as hyperscaler capex redoubles and the Meta deal rolls into full revenue contribution over the next twelve months.

This article was written by

LL Insights

501 Followers

I am a retired quant with a PhD in mechanical engineering. I started off my professional career as an engineer and eventually transitioned into a hybrid developer/quantative analyst role at the investment arm c

Show more

Analyst's Disclosure: I/we have a beneficial long position in the shares of AMD either through stock ownership, options, or other derivatives. I wrote this article myself, and it expresses my own opinions. I am not receiving compensation for it (other than from Seeking Alpha). I have no business relationship with any company whose stock is mentioned in this article.

Seeking Alpha's Disclosure: Past performance is no guarantee of future results. No recommendation or advice is being given as to whether any investment is suitable for a particular investor. Any views or opinions expressed above may not reflect those of Seeking Alpha as a whole. Seeking Alpha is not a licensed securities dealer, broker or US investment adviser or investment bank. Our analysts are third party authors that include both professional investors and individual investors who may not be licensed or certified by any institute or regulatory body.

Comments (16)

Sort by Newest