Can AMD break the CUDA Moat? AMD Advancing AI 2026

Up to 105% Equity Rebate Discounts for OpenAI, Agentic Kernel Generation, Improvement in Software Quality, Unstable Internal Development Clusters, Helios MI455X Production Ramp Hell

BRYAN SHAN, DANIEL NISHBALL, MYRON XIE, AND 9 OTHERS

JUL 25, 2026 • PAID

When we published our first AMD software article, we gave AMD a 0% chance of closing the gap with Nvidia in AI accelerators. Software was broken, progress was unexciting, and we were the top bug submitter for many months with dozens of AMD engineers triaging our bug reports.

Six months later, in our AMD 2.0 article, we took the non-consensus position of upgrading from 0% chance to a much more meaningful chance at success. We published this opinion when sentiment towards AMD was sitting at rock bottom, and when most of the market thought we were were being far too optimistic towards AMD.

That view was based on our observations that AMD has the leadership that can create change rather than suffer from committee style leadership. Lisa quickly hopped on a calls with us and has since then implemented many of our suggestions. We saw an AMD that had finally recognized the importance of software and had a sense of urgency, moving in the right direction even if the destination was still far off. Since then, the signal has only gotten stronger.

Based on our experience on AMD software stack this year, we update our view again from non-zero percentage chance to now a great chance of success as long as AMD solves the two major risks we outline below. It is important to highlight that just because AMD gains market share, that doesn't mean that Nvidia will do poorly. The pie is growing rapidly for everyone, and Nvidia will continue to massively grow revenue. AMD poses potential competition to Nvidia on the software front, and Jensen will need to cut bureaucracy and flatten the layers of different required internal stakeholder approvals

Anthropic has publicly announced that they will deploy 2GW of AMD's chips, Lisa Su and team have leaned into an agentic oriented engineering culture. Anthropic Head of Compute Tom Brown explained how **he used Claude over the weekend with "/goal" to bring-up internal Claude inference stack on AMD hardware** as a case in point. We believe that since AMD's compiler and most of their kernels are open sourced, that they are better positioned for the agentic age **besides one major risk that we will discuss later**. Three months ago, our Accelerator model noted that Anthropic will be an AMD customer. We also noted this on our socials a week ago.

In 2023, Microsoft dropped AMD after the MI300X due to unreliable Samsung 2023 HBM memory and poor software quality, subsequently skipping both the MI325X and the MI355X. Microsoft has since made an about face and has announced that it will deploy MI455X Helios. We believe that OpenAI will be the main end customer for Azure's MI455X racks. In a strategy somewhat similar to the Nvidia-Groq deal, AMD is announcing a deal with Cerebras to do PD disagg for ultra-fast interactivity inferencing.

There are two major risks that AMD needs to guard against:

1. Supply chain checks and engineering first principles analysis and understanding show that AMD's first AI rack scale system, Helios, is going currently going through a slow rack production ramp given that it is not using a cableless tray design, which the Rubin Oberon rack is now adopting. Furthermore, since AMD has a weak SerDes design, up to 85% of its backplane needs to be retimed, requiring over over 550 Broadcom ethernet retimers per rack. Furthermore, it is running into backplane reliability challenges during rack production ramp hell.

2. **The chief complaint from most AMD engineers internally is that there is a persistent lack of stable GPU clusters for internal software development teams and a lack of stable GPU clusters for automated testing CI**. This is blocking AMD's rate of progress and **it is holding AMD back from harnessing the potential upside of AI coding Agents** because each AI Agent requires GPUs as well and also requires a testing tool use loop. We will more explain below on our recommendation section.

If AMD can overcome these challenges, we strongly believe that AMD will be well positioned to do well and take market share **especially considering AMD is giving Meta & OpenAI close to an 105% equity rebate discount** via their finance engineering & stock

is practically negative cost. AMD is practically giving away an Helios and 5% extra on top of that to SF based nonprofit OpenAI.

In the first part of our article, we do a deep dive into the Helios architecture and MI455X (gfx1250) instruction set which is an clone of Hopper SM90's ISA. Then we will go into details of the software stack. And then for part 3, we will break down the economies of OpenAI/Anthropic/Meta's equity based rebate discounts & how that effects TCO.

We are a daily user of ROCm software on MI300X, MI325X, MI355X & also the #1 bug report consistently every quarter, there is a couple 10x engineers we want to shoutout work 997 improve AMD software & quickly triaging our bugs reports including Hongxia, Chun Fang, HaiShaw, Thomas Wang, Andy Luo, Seungrok, Bill He, Teresa Shan, Parth, Duyi Wang, Gilbert, and many more. Most of AMD best 10x engineers are in Shanghai. AMD's MoRI collective & UMBP KVCache offloading team, AMD's disaggregated application forward deployed engineering team, and other AMD teams that understand how to do first principles-based inference engineering are all mostly based in Shanghai. **A lot of the ROCm software stack's most important pieces are built in China.**

Recommendations to Lisa Su, Mark Papermaster, Sharon Zhou, Anush Elangovan & Vamsi Boppana

There has been good progress over the past year on automated testing to improve software quality, but the pace of progress has not as aggressive as needed and there continues to lack of sense of urgency on providing enough GPU clusters for automated testing CI & for internal software development teams. It is always too little too late. Every time we have meeting with y'all (Vamsi, Anush) every couple of months & occasionally meet with Lisa, CI get better for specific examples we point out and things are heading in correct direction, but overall strategy need to be aggressively ramped up. For example, Kubernetes Inferencing Pollara NIC CI still sits at 0% parity with NVIDIA's ConnectX Nightly CI. Kubernetes is the layer that most inference deployments across the world use. It isn't due to AMD engineering not *wanting* to add it, rather they are blocked by lack of investment in internal CI capacity. The planned ETA of advancing AI 2026 reaching parity on this was missed due to cluster issues.

were making good progress on vLLM gating over the past couple of weeks to reach the ETA of reach 90% parity with CUDA on gating by Advancing AI 2026, until AMD leadership started pulling clusters away from AMD's internal vLLM team to deploy elsewhere due to AMD's internal capacity overrelying on backstopped temporary clusters.

Gating/blocking tests mean the tests are of the highest quality, since PRs cannot merge unless the test is passing. This prevents bugs from being merged. While AMD leadership may distract non-technical folks by showing their non-gating pass rate, gating parity & gating pass rate are what truly matter.

We hope that AMD leadership can re-prioritize giving their internal vLLM team stable clusters & update their philosophy on capacity planning to prevent issues like this in the future, so that AMD's internal hardcore vLLM engineers can do the work of reaching 90%+ parity with CUDA vLLM on gating and have the tools needed to operate at the same velocity as the AMD internal SGLang team.

This is the chief complain from most AMD engineers is that leadership needs to update their viewpoint on providing stable CI clusters that don't just randomly need to be migrated and shifted from 1 CSP to another CSP.

Furthermore, there is a constant lack of GPUs internally for development. For single node aggregated inferencing, there is enough GPUs to go around internally. But in the age of Distributed Multi-node Inferencing optimizations (wideEP & disaggregated PD), there is nowhere near enough GPUs. Even with the additional 2000 MI355X coming online this month and the 6000 MI325X/MI355X coming online later this year, it will not be enough and still more than an order of magnitude less than the stable long term clusters NVIDIA has for internal development. This lack of GPU nodes is even getting worse due to the rise of agentic coding. Previously, each human engineer required a couple of nodes to do DI inference software development. But now with agentic coding, each agent requires GPUs to test their code against and each human can have dozens of agents running at the same time and each agent can only dozens of sub agents running at the same time too. Thus, without Distributed Inferencing simulation tools like DynoSim, there will be even greater shortage of internal GPU capacity internally.

Furthermore, unlike Rubin (SM107) which is an very similar ISA to Blackwell (SM100), MI455 (gfx1250) is an complete different ISA from MI355 (gfx950) and has completely different codepaths and kernels which will require testing on both gfx950 & gfx1250. This is

was missed. We hear that the new timeline has been delayed till October 2026 but they are trying to leftshift it to August/September timeline.

We view this slow ramp of internal software development GPU clusters & automated testing CI clusters is one of the major risks slowing them improvement of software quality & performance and hope AMD leadership revisits their capacity planning strategy going forward.

Part 1: AMD MI455 Silicon & Helios Rack

Source: AMD

AMD has done it again on silicon engineering. The MI455X is the most advanced chip that comes out of a fab on the silicon front. AMD is the first company to ship 2nm datacenter silicon with both the MI455's compute tiles as well as the Venice CPU being early adopters of N2. All of the competing accelerator platforms are on N3.

AMD continues to lead on package integration as well. The MI455 package is the largest CoWoS-L module shipping at 5.5x reticle size. AMD is still the only company adopting TSMC's SoIC-X hybrid bonding, which allows AMD to scale silicon footprint in the z dimension as well as the x and y dimensions. All this comes together to bring the MI455 to

The package layout borrows heavily from the MI355X. 8 N2 ‘XCDs’ are hybrid bonded atop of 2 reticle-sized base dies which contain SRAM, HBM controllers, as well as the compute fabric that allows for the XCDs to communicate with each other.

One of the major areas where the MI455X departs from the MI300 family is the addition of 2 separate I/O dies that house all the PHYs for off-package communications. While the floorplan representation below shows one IOD for the UALoE scale up fabric, and the other IOD for the link to the host CPU and NICs for the scale out fabric, the I/O dies are in fact identical. This is a result of the clever implementation of flexible I/O, with each I/O die supporting 72 lanes that are compatible with different protocols at different line speeds including: 212G UALoE, 64G Infinity Fabric, PCIe Gen 6, 128G UALink, xGMI4. With many of these protocols contributed to the open UALink consortium in one form or another, the I/O is designed from the ground up to integrate well with UALink based interconnect solutions.

Source: AMD

This silicon spam is what allows AMD to deliver industry-leading peak theoretical dense FLOPs with the MI455X delivering 20PF of FP8 vs Rubin at 17.5 FP8. However, absolute numbers aside, these performance advantages are relatively mild compared to Rubin given

LUT tensor cores that Rubin SM107 has. **3 bit LUT tensor cores has potential advantages for reducing relative HBM bandwidth needs.** If anything, AMD is forced to be aggressive on silicon to compensate for this deficiency, not to mention its relative weaknesses in software and system design, though it's software is getting better. Alternatively, AMD's product marketing team needs to find a few more tricks to juice marketed performance numbers.

Source: AMD, Nvidia

The large package size also allows AMD to fit 12 stacks of HBM4 for a total of 432GB per package. This is again industry leading compared to Nvidia and Google shipping only 8 stacks which is 288GB per package. Each base die is now closer to reticle sized with the HBM-facing edge measuring 32mm to squeeze in 3 cubes per edge, compared to the MI300X AID which was 29mm per edge which isn't enough to fit 3 cubes.

Along with Rubin, AMD is the only company shipping HBM4 this year. Having 12 stacks also allows the MI455 to be the winner on memory bandwidth at 23.3 TB/s per chip, implying HBM4 pin speeds of 7.6Gamdbps. This is a slight upgrade from the 19.6TB/s advertised at last year's Advancing AI. Despite having a 50% wider bus than Rubin, the total bandwidth is barely above Rubin at 22TB/s.

need to make up for the memory bandwidth deficit to AMD MI455X. At Nvidia's marketed 22TB/s for Rubin, this is a pin speed of 10.7Gbps: 40% faster than MI455's HBM4 which means NVIDIA will be forced to use an much higher quality bin than AMD.

As we have mentioned before and covered extensively in our newsletter and Accelerator Model, this change was challenging for memory suppliers to keep up with. Memory suppliers had to rework their HBM4 to deliver the spec Nvidia was targeting. This also pushed out upstream output for Rubin but as of today these issues have been finally resolved. This move proved to be a worthwhile gambit for Nvidia as they could close one of the big shortfalls of Rubin and despite the delays, Vera Rubin will still be the first deliver tokens at scale compared to MI455 Helios. More on that later.

Active LSI

Hidden beneath the surface, we believe that the MI455 is also the known first chip to ship with active Local Silicon Interconnects ("LSI"), which are the bridges that connect the various chiplets within the CoWoS-L assembly. In the brief history of CoWoS-L, the bridges have been passive, containing only wiring as well as capacitors. Active LSIs come with actual circuitry. During TSMC's aLSI presentation at ISSCC 2026, TSMC demonstrated active bridges with a low-power repeater circuit that regenerates signals mid-channel. The benefit is that because the bridge now shares the burden of maintaining signal integrity, the PHYs on the top dies can shrink meaningfully, at a nearly negligible energy cost, reclaiming leading-edge silicon and shoreline for compute and memory. The give away that this is being shipped in the MI455 is the "test vehicle" TSMC showed is the MI455 interposer: with two base dies, 12 HBM4 stacks, and two IO dies.

Source: TSMC Active LSI Die Shot and Power Breakdown, ISSCC 2026

Source: SemiAnalysis

Meta Recsys Infra Strategy Odd Choice Hurting AMD's chances of success At Meta

not take advantage of AMD's innovation here. Most of Meta's orders for the MI455 is a customized variant that is a cut-down version of the full MI455X: the compute silicon is halved from 8 dies to 4, and the HBM drops from 12 stacks to 6 per package. The HBM4 itself is also a step down, with 8-Hi stacks in place of the standard SKU's 12-Hi. Halving the compute and memory silicon raises the CPU-to-GPU compute ratio, mirroring Nvidia's Meta-specific "Ariel" variant of the GB200 NVL72. This chip configuration is intended for Recsys workloads and was something that Recsys infrastructure teams decided. However, the decision was made before TBD Lab was formed or could have its say. Given the significant compute and HBM deficiency for the scale up domain versus Rubin, TBD has no interest in this system, nor is it attractive for external customers.

As we specifically stated on our last piece on Meta infra strategy, this decision on doing an half size custom MI455, is going to nuke AMD's volume at Meta because TBD will vastly prefer Rubin if the half MI455 design is chosen. AMD needs to step in, put on their big boy pants, and work directly with teams at TBD to make sure they get the normal MI455 instead of the gimped Meta custom version which is terrible for LLM training & inferencing. The normal MI455 will be competitive with Nvidia's Vera Rubin.

We do think there is hope though as after our article, Mark Zuckerberg has already began rapidly exploring changes to their infrastructure strategy org culture.

Source: SemiAnalysis

On the network side, MI455X marks the first AMD GPU deployment with switched scale up and in a rack-scale scale up domain. This a huge upgrade over the 8-GPU point-to-point mesh used from MI300X through MI355X.

AMD's Helios rack connects 72 MI455X GPUs through 12 102.4T Tomahawk6 switches in a single tier all-to-all network. Each GPU has 72 lanes of 200G UALoE for the scale up fabric, resulting in total scale up bandwidth of 1.8 TB/s uni-di per GPU. Each switch uses 432 of its 512 200G lanes, with the over-provisioning a result of AMD using merchant TH6 switches from Broadcom, rather than a proprietary co-designed switch like Nvidia. Scale-out moves in parallel from 400G Pollara to 800G Vulcano, with two NICs delivering 1.6 Tbit/s per GPU with the option of adding a third NIC for each GPU. MI500 is expected to extend the scale-up domain to 256 GPUs across three racks, where copper will likely give way to co-packaged copper or co-packaged optics.

Source: SemiAnalysis

Helios Rack Explainer Revisit

It has been a year since AMD's Helios architecture was announced at Advancing AI 2025, and a lot has happened since then. Let's revisit the Helios architecture and review some of the changes and nuance that happened over the last year.

Rack Elevation Diagram

Source: SemiAnalysis

The Helios rack features 18 compute trays and 6 scale up switch trays, sitting in the middle of the compute trays. Each compute tray will house 4 MI455X GPUs and 1 Venice CPU, adding up to 72 GPUs and 18 CPUs in aggregate. Each scale up switch tray will house 2 102.4T Tomahawk 6 switch from Broadcom adding up to 12 switch ASICs in aggregate. For Meta's version of MI455x, there will be six 51.2T Minipack Ethernet switch within the rack, and the power supply shelf will be in a separate IT rack.

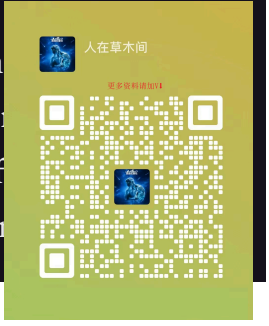
Compute Tray Layout

Source: SemiAnalysis

The compute tray design has not largely changed since Advancing AI 2025. Each MI455X GPU sits in an EAM module connecting to the scale up links via backplane connectors and the Venice CPU and Vulcano NIC via flyover cables. Notable, the LPDDR5x directly attached on the EAM is not to be seen, we will explain this below. The Venice CPU sits in its own module board with 16 slots of RDIMM and 5 NVMe SSD slots attached. The design is modular however the use of flyover cables to connect between modules could lead to challenges in production.

Partial Codesign

AMD has presented a wide portfolio of products for their rack scale solution
MI455x GPU, Venice CPU, Pensando Vulcano 800G NIC, and Pensando Sali
However, the scale up switch, which enables the true rack scale scale up perf
missing. Unlike Nvidia, AMD has to rely on a partner for an off the shelf me



partners. Because Broadcom is the only merchant provider shipping a 100T switch with 200G SerDes, Broadcom is AMD's only choice for the switch. This also complicates logistics and responsibility between vendors, which requires back and forth communication and problem solving together. Due to the lack of a scale up switch product, AMD cannot reach extreme codesign for the complete rack solution.

Source: AMD

Memory Despec

One notable omission is there is no longer any direct attached LPDDR running off of each accelerator. Previous roadmaps had up to 1TB of LPDDR as second tier memory for each MI455X attached on the EAM module, however this has now disappeared. We believe this is another consequence of tight memory supply.

Backplane retimed

Another feature of Helios is the addition of Ethernet re-timers on the scale up switch. This is unlike Nvidia's Oberon backplane which is completely passive. AMD's 200G SerDes struggles from the loss on the copper path between MI455X and the scale up Tomahawk 6. It could be a result of their flexible I/O ambition or simply inferior SerDes quality leading to inadequate signal performance in the backplane. For Meta's MI455X deployment, ~85% of

extra cost and power budget to the entire system. Also, it complicates server assembly as it will be a very tedious task tuning all the retimes when bringing up a rack.

Manufacturability and Efficiency Challenge

The MI455X Helios design was AMD's response to Nvidia's GB200/GB300 Oberon architecture. AMD referenced a lot of Nvidia's architecture for Helios. While the design enables high density single rack scale up, AMD has also borrowed the design that made Nvidia suffered, namely the flyover cables. We have discussed the manufacture challenges regarding the flyover cables in our Vera Rubin architecture deep dive article. The difficulties with flyover cables of GB200/GB300 made Nvidia pivot toward the cableless design for Vera Rubin NVL72. By the time this was announced, it was too late for AMD to implement this design philosophy for MI455x Helios.

In the compute tray, the Genesis cable from Molex handles the 128G UALink between the MI455X and Pensando Vulcano NIC. There will be up to 12 Genesis cables per compute tray cramming into the 1U space alongside liquid cooling tubes and power cables. On the scale up switch side, flyover cables are between the backplane connector and the TH6 switch. 1,728 cables will be routed from the backplane connectors to the 16 ports (32 ports per tray) around each TH6 ASIC. All the cables become potential points of failure during assembly and will make manufacturing inefficient.

Compute Tray Topology

Each of the 18 compute tray in the Helios rack holds four MI455X EAMs, each of which support 36 UALoE links, which in turn are built up using two lanes of 200G Ethernet. 72 lanes of 200G ethernet sums up to 14.4Tbit/s uni-di of total scale-up bandwidth per GPU. The MI455X features two N3P based I/O blocks on the North and South of the package – one I/O block hosts these UALoE links for the scale-up network, while the other implements Infinity Fabric for connecting to CPUs as well as UALink128/PCIe Gen7 for connecting to NICs.

Source: AMD

Scale Up Topology

Source: SemiAnalysis, AMD

On the other end of these UALoE links are 12 Broadcom Tomahawk 6 Ethernet switches across 6 switch trays with two switch ASICs per tray. UALink over Ethernet (UALoE) is the equivalent layer of the stack of NVIDIA's NVLink. Each Tomahawk 6 switch ASIC is built

of bandwidth going from each switch ASIC to the 72 GPUs in the rack. A switch aggregate bandwidth that divides evenly among 72 GPUs would be ideal – 115.2Tbit/s for instance, but AMD is using the only switch that is available today. UALink switches with 115.2T and 57.6T aggregate bandwidth are on the horizon, but not close enough for them to anchor this architecture. In contrast, Nvidia designed the 28.8T NVSwitch to match the needs of a 72 GPU rack, and this divides bandwidth evenly among 72 GPUs, 400Gbit/s uni-di each, with no wastage of bandwidth at all. Each switch tray also includes a small host x86 CPU.

Source: AMD

AMD uses a scale-up topology that should by now be familiar to our readers. In this flat one-layer network, each GPU connects to each switch using 6 lanes of 200Gbit/s uni-di per lane, for a total of 1.2Tbit/s uni-di to each switch of the 12 switch ASICs. Each switch ASIC in turn connects to all GPUs in the rack.

Source: SemiAnalysis

The scale-up switch to GPU links is implemented using a copper cable backplane. For each GPU, each of the 72x 200G lanes is carried over two differential pairs (DPs) of copper channels, one for transmit and one for receive. This totals to 144 DPs of copper cables per GPU, or 10,368 differential pairs of copper cables for the entire rack. Each GPU will also need a 144DP male and a 144DP female connector to interface between the backplane cables and the compute tray itself. On the switch side, four banks of connectors will be used per switch tray, with each bank hosting four 108 DP connectors supporting each – a total of 432 DP per bank.

Flyover cables are used to connect the switch ASICs to the connectors on the back of the compute tray. Flyover cables offer better signal integrity than PCB traces, but the disadvantage is that they tend to limit serviceability and thermal efficiency, and add manufacturing challenges on top. Nvidia has flip flopped between the use of flyover cables and PCB – initially aiming to use flyover cables before changing to use PCB traces for better serviceability and airflow.

Source: AMD

The total backplane and compute tray content per rack will add up to $68,928, with $44,352 coming from the backplane and the remaining $24,576 attributable to the flyover cables.

Source: SemiAnalysis AI Networking Model

Helios rack with 9 compute trays on the upper portion and 9 in the lower portion – this allows signals to travel equal distances in both directions.

Source: SemiAnalysis

Scale-out Networking

For scale-out, each MI455X can connect to as many as three AMD Pensando Vulcano 800 AI NICs delivering 2.4 Tbit/s of scale-out bandwidth, with each NIC attached through an x8 UALink128 interface delivering 256 GB/s of bidirectional GPU-to-NIC bandwidth. A separate Pensando Salina 400 DPU handles the front-end network.

Source: SemiAnalysis AI Networking Model

We expect the dominant deployment configuration to carry two Vulcano NICs per GPU for 1.6 Tbit/s of scale-out bandwidth.

Outside the rack, Vulcano supports a multi-plane leaf-spine topology that is non-blocking within each plane which is now a common implementation. We have written extensively on how multi-plane architectures are crucial in AI clusters today as it can fit more GPUs on a two-layer network. We elaborated more on the intuition and advantages behind multi-plane fabrics in our Vera Rubin article here.

Source: SemiAnalysis AI Networking Model

Source: SemiAnalysis AI Networking Model

In a 131k MI455X 8-plane cluster, each 800G Vulcano NIC breaks into four 200G links, giving every GPU one connection into each of the eight independent planes given each GPU is attached to two NICs. Each plane has 512 leaf and 256 spine switches, for 6,144 TH6 switches across the cluster. The configuration requires two 800G DR4 modules at the NICs and three 1.6T DR8 modules per GPU – two for the leaf-side uplink and downlink and one spine-side. We expect the scale-out networking content per MI455X GPU on an eight-plane two-layer configuration to be ~$8,000 per GPU.

Source: SemiAnalysis AI Networking Model

newsletter, we highlighted the issues faced on the data-plane and control-plane level in AI scale-out networks.

In large AI training clusters, there are three main issues: elephant flow, low entropy and suboptimal fabric utilization.

- Elephant Flow: AI workloads tend to have long-duration, heavy-traffic flows that can congest the network and reduce the overall performance of the training batch.
- Low Entropy: Depending on the training job, the number of IP flows can be limited and congest only a few links while the overall fabric still has plenty of capacity.
- Suboptimal Fabric Utilization: Finally, as an overall effect of both elephant flows and low entropy, there is a large skew in the bandwidth utilization of fabric links, which drives how much the fabric needs to be overprovisioned to run smoothly.

RDMA traffic is highly sensitive to packet loss and congestion, and RoCEv2 is built on Ethernet, a lossy protocol, and UDP, which does not allow for recovery mechanism. Meta's DSF addressed this inside the network with virtual output queues, cell spraying, credit scheduling and deep-buffer Jericho-AI switches but at the cost of a much more complicated data plane, control plane and hardware stack.

Vulcano tries to solve the same problems but through the NIC. Here, the NIC comes with 3 elements, Intelligent Packet Spray, Path Aware Congestion Control, and Out-of-order Packet Handling and In-order Message Delivery, that allows to offload the network complexity from the fabric to the NIC.

Intelligent Packet Spray attacks the low entropy problem at the core. In traditional ECMP fabrics, a flow is hashed onto a single path and stays there. Since AI traffic is composed of a handful of massive, synchronized elephant flows, hash collisions are frequent and some paths get congested while others sit idle. Instead of pinning a flow onto a single path, AMD's AI NIC sprays the packets of the same flow onto the available paths of the fabric, raising the entropy level to a per-packet level. The result is a 1:1 fabric utilization, and a failure recovery in milliseconds because no link failure can take down the whole flow.

However, spraying packets creates a new set of problems, which is where the 2 other elements come into play. First, packets coming from different paths can arrive in the wrong order. Traditional RoCEv2 networks force the packets to go back N retransmissions, which

enforces ordering at the message level, removing the need for buffering. Packets are written directly onto the GPU memory as they land, and only lost packets are retransmitted.

Second, spraying packets blindly across path cause congestion around instead of avoiding it. Path Aware Congestion Control closes the gap as the NIC tracks the real time status of each path and shifts traffic away from congested ones before queues builds up. This is aimed specifically at incast, as collective operations like all-reduce tend to have many senders for one receiver, causing head-of-line blocking cascades through the fabric. Because decisions are made at the NIC level, with per-path visibility, ramp up is fast enough to handle the bursty nature of AI workloads. Also, tail latency is minimized here since packets are sprayed intelligently across the fabric.

Taken together, those 3 mechanisms solve the same problem Nvidia solves with its ConnectX NIC coupled with Spectrum-X adaptive routing. The main difference here is that AMD is adopting the open UEC standard and a multi-vendor fabric rather than a vertically integrated one.

AMD's AI NIC does not impose the transport model and allows for flexible scale-out networking. It supports switch-based packet-spray, for operators whose fabric already handles load balancing, NIC-based packet-spray, for dumb commodity fabric where the intelligence is offloaded to the NIC, and NIC-based source routing, for operators who want path control from the endpoint. The whole point is that the network behavior adapts to the fabric you already own and the workload you intend to run, not the other way around.

CDNA5 Microarchitecture

Designs Converging with NVIDIA

In the same way that AMD has drawn inspiration from Nvidia's boring keynote format, AMD's CDNA 5 in many ways draws huge inspiration from Nivida's Hopper (SM 90) architecture. First, CDNA 5 reduces the number of threads per wave to 32, matching it with NVIDIA's 32 threads per warp. CDNA 5 also replaces CDNA3/CDNA4 Infinity Cache and small L2 cache with a single large 96 MB L2 cache per FCD (Fabric and Cache Die). This design converges closer to NVIDIA's "global memory -> L2 cache -> shared memory" memory hierarchy. Managing memory latencies across different hierarchies has always been a pain point of AMD kernel writers; we expect simplifying the hierarchy should mitigate this issue.

Source: AMD

Increased Staging Memory and MMA Shapes

CDNA 5 has larger memory staging buffers than its NVIDIA counterparts. It has 320 KB of LDS (roughly SMEM-equivalent) and 32 KB of VGPR (roughly thread register-equivalent), so each thread has access to 1024 registers. In addition, we have only seen CDNA 5 supporting mostly 16x16xK shapes like its predecessor. Combining these facts, we theorize that the larger staging buffer is for adapting to the increase in wave count: Under the same number of parallel threads, smaller wave size would lead to a larger number of waves. Since the MMA shape didn't increase, and each thread has access to 4x number of registers, AMD doesn't have to resort to increasing MMA scope to warp groups like NVIDIA does.

Tensor Data Mover

Tensor Data Mover (TDM) is almost identical to NVIDIA's Tensor Memory Accelerator (TMA). TDM moves data from HBM to LDS without register staging. It supports 5-dimensional tiling, out-of-bound checking, and even multi-cast into different work group clusters, the equivalent of NVIDIA's thread block cluster. One difference is that TDM descriptors are loaded from SGPR, unlike NVIDIA loading from host to shared memory. NVIDIA's Rubin has just showcased inline TMA descriptor updates to improve user ergonomics, so we look forward to seeing whether AMD will get the design right.

Source: AMD

The two FP4 formats now competing for inference share the same E2M1 element: one sign bit, two exponent bits, one mantissa bit. But they differ in how they carry the block scale that makes so narrow a type usable. MXFP4, the OCP standard AMD has built its four-bit path around, pairs each 32-element block with a power-of-two E8M0 scale. NVFP4, the format NVIDIA introduced with Blackwell, uses a finer 16-element block, an FP8 E4M3 scale per block, and an FP32 per-tensor global scale on top, resulting in a costlier but more accurate arrangement that is fast becoming the default for FP4-quantized checkpoints.

The notable thing about AMD's gfx1250, the architecture target behind MI455X, is that its matrix engine speaks NVFP4 natively. MLIR-level scale-format enum in ROCm WMMAMatrixScaleFormat, includes e8, e5m3, and e4m3 members.

More concretely, AMD has already shipped a gfx1250-compiled NVFP4 GEMM code object inside AITER and wired it into the runtime dispatch table, where a format discriminator distinguishes NVFP4 from MXFP4 and the dispatch logic selects the NVFP4 assembly kernel on gfx1250. This is a gfx1250-specific capability, not a CDNA4 one. MI355X / gfx950 does four-bit only as MXFP4, through scaled-MFMA with an E8M0 scale over 32-element blocks — no scale-format field, no 16-element blocks, no E4M3 — which is all AMD's public matrix-core documentation describes.

We also found that CDNA 5 additionally supports unsigned E5M3 (UE5M3) scaling factor format, in contrast to NVIDIA Rubin supporting E4M3 and E5M2. UE5M3 repurposes the

additional per-tensor FP32 scaling, and future adoption of those formats will show their efficacy.

Source: https://arxiv.org/pdf/2601.19026

Conservative Microarchitectural Changes

The microarchitecture evolution from CDNA 4 to 5 shows AMD believes less in scaling than NVIDIA does. NVIDIA has bet on models requiring larger multiplications and scaled MMA shapes aggressively every generation, scaling to MMA shapes that need 2 SMs to execute in Blackwell. Since CDNA 5 barely scales MMA shapes, we doubt we will see an equivalent feature in CDNA 5. **We also have yet to see innovations on data compression techniques from AMD like Rubin's 3-bit lookup table weight compression MMA mode.**

Part 2: AMD Software Defeat the Cuda Moat?

AMD Software Is Improving Quickly, But the Moat Has Moved

AMD's software story is not "ROCm is broken" anymore. It is that ROCm is finally moving with real urgency, but the competitive frontier has moved faster. We said in April 2025 that

https://newsletter.semianalysis.com/p/amd-2-0-new-sense-of-urgency-mi450x-chance-to-beat-nvidia-nvidias-new-moat

What Has Actually Improved

CI Is Maturing, But not fast enough

We praised AMD for moving stable ROCm support into upstream vLLM releases in January 2026 and adding nightlies shortly afterward. CI has since made concrete but incomplete progress. A June change added AMD mirrors and gates for eight major test groups: V1 attention, engine, OpenAI API correctness, small-model evaluation, multimodal pooling, and three speculative-decoding paths. A separate July patch cleaned up flaky jobs and prepared existing mirrors to become merge-blocking again. But CUDA parity has not yet been demonstrated: public regression dashboards, AITER accuracy gates, end-to-end disaggregation CI, and automatic performance gating remain roadmap items.

Encouragingly, AMD's distributed-inference work has begun moving from one-off recipes into upstream CI. SGLang merged a two-node MI355X 1P1D disaggregation nightly in June for DeepSeek-V4 Flash and Pro, in both FP8 and FP4, then added DP-attention, EP8, MTP, and Kimi K2.6 coverage in early July. This is the important shift: disaggregated inference is starting to become a continuously tested upstream feature rather than a demo.

Kubernetes powers most inference deployments across the world, and AMD is a founding partner of llm-d, an open-source distributed inference Kubernetes orchestration engine. But it has yet to have sufficient CI automated testing for the first party Pollara NIC. It isn't due to engineering not wanting to add it, but they still to this day under-invest in internal CI capacity planning. This results in 0% parity to NVIDIA's ConnectX-7 NIC on llm-d Kubernetes inference nightly testing. The planned ETA of advancing AI 2026 reaching parity on this was missed.

On the vLLM side, gating automated test progress has massively regressed due to AMD cluster infra stability issues this week. AMD's hardcore engineers were making good progress on vLLM gating over the past couple of weeks to reach the ETA of advancing AI to reach 90% parity with CUDA on gating, until AMD leadership started pulling clusters away from AMD's internal vLLM team.

non-gating pass rate, gating parity & gating pass rate are what truly matter.

We hope that AMD leadership (Anush, Vamsi, Mark Papermaster) can re-prioritize giving their internal vLLM team stable clusters, so that AMD's internal hardcore vLLM engineers can do the work of reaching 90%+ parity with CUDA vLLM on gating and have the tools needed to operate at the same velocity as the AMD internal SGLang team.

Single-Node Performance and Reproducibility Are Real Wins

In March, we highlighted an up to 18× improvement in Kimi K2.5 1T MXFP4 interactivity in under 30 days from AITER/vLLM fixes that were already upstreamed into vLLM 0.18. AMD's own February 2026 technical write-up, "Speed is the Moat: Inference Performance on AMD GPUs", argues that AITER-driven single-node optimizations deliver a roughly 1.08x–1.2x throughput uplift over baseline framework configurations. That is the correct direction: the baseline open-source frameworks must get faster on AMD, not just AMD-only demos.

AMD's MiniMax M3 performance has also caught up to B200, with optimizations via their ATOM stack.

Source: https://x.com/RyanLeeMiniMax/status/2080142342288445553

Where the story has improved most is in recipes, documentation, and reproducibility. ROCm now has a much deeper public recipe layer than it did a year ago. The ROCm inference docs expose a full AI-inference section spanning vLLM, SGLang, distributed

to multi-node scaling. The ROCm/MAD repo now publishes blueprints spanning vLLM, SGLang, training stacks, large-EP microbenchmarks, and disaggregated prefill/decode recipes.

The Posture Is Right

This is why we are mostly aligned with Anush on the software diagnosis. The developer-first posture, upstream alignment, day-0 model enablement, and faster release cadence are the right playbook. We praised Anush's developer-relations push in 2025 and, more recently, credited his team for moving ROCm from a second vLLM fork toward something closer to a first-class upstream experience. The official vLLM blog now says the era of "just porting" AMD support is over, documents seven ROCm attention backends, and shows 1.2x–4.4x throughput gains from the latest AMD/vLLM orchestration work. That is what a credible catch-up path looks like.

George Hotz put it well back in 2025:

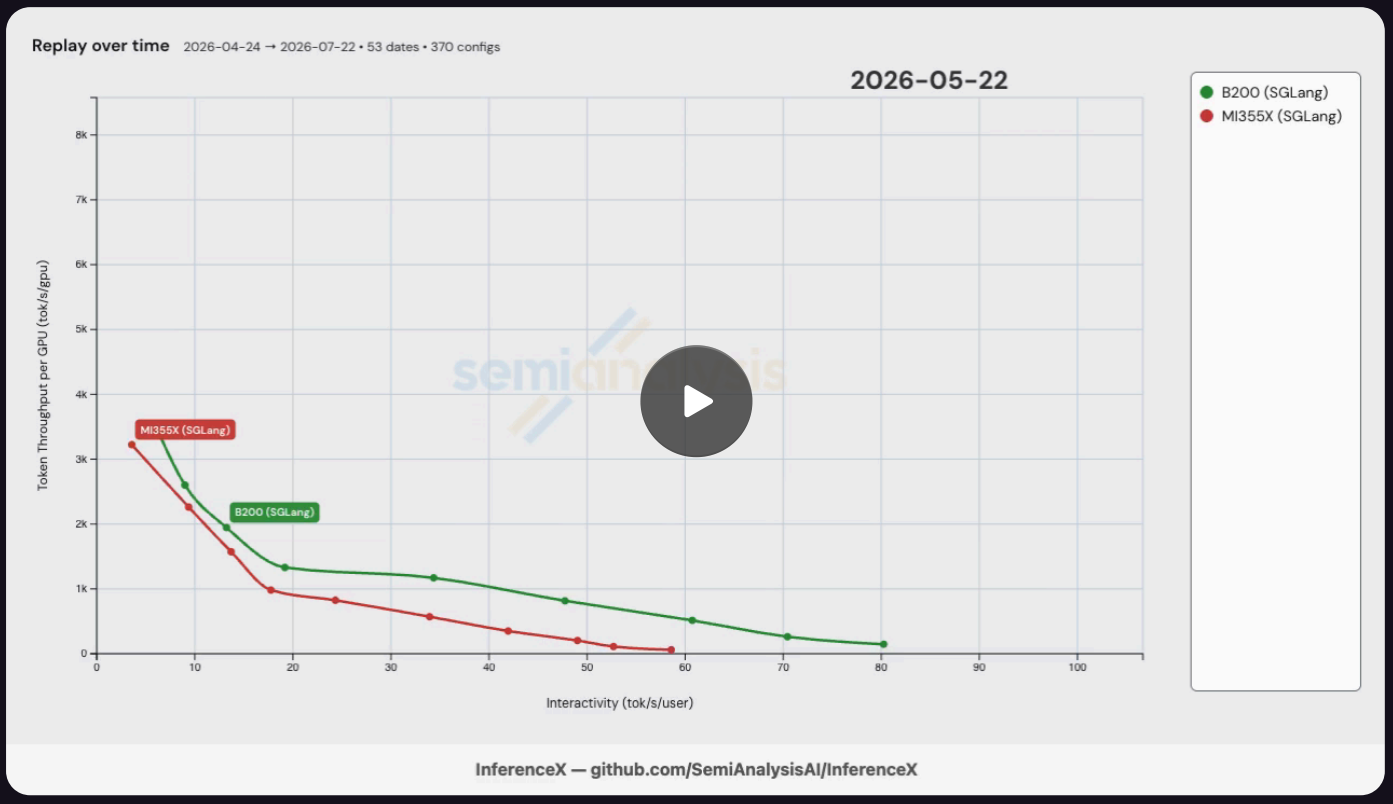
> AMD's dysfunction is different. From the beginning they had leadership that can do things (Lisa Su replied to my first e-mail), they just didn't see the value in investing in software until recently. They sort of had a point if they were only targeting hyperscalers. but it seems like SemiAnalysis got through to them that hyperscalers aren't going to deal with bad software either. It remains to be seen if they can shift culture to actually deliver good software, but there's movement in that direction, and if they succeed AMD is so undervalued. Their hardware is good.
>
> -George Hotz

InferenceX: AMD Dev Velocity is Getting Better

In the past, we have seen AMD struggle in efficiently bringing up new models on their hardware in both the aggregated and disaggregated scenarios. We track all iterative performance on InferenceX which allows us to estimate this "velocity" to some extent.

As mentioned in our previous article, the AMD inference team did a great job at rapidly improving its DeepSeek v4 performance over the first month or so, especially in the single node case. Below is a chart labelling all improvements made in the first 47 days of the DeepSeek v4 release to MI355X SGLang single-node config.

Source: SemiAnalysis InferenceX

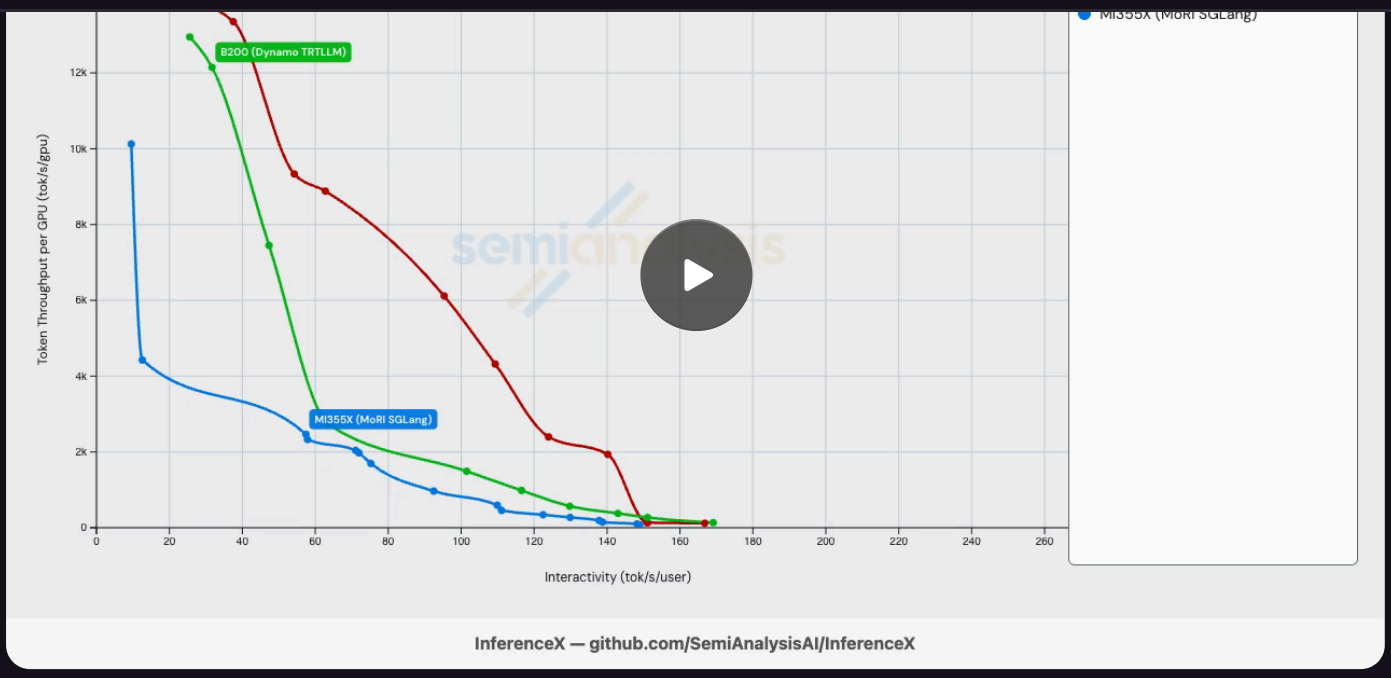


More recently, we saw a similar story for MiniMax M3, where the AMD inference team rapidly iterated to achieve competitive performance.

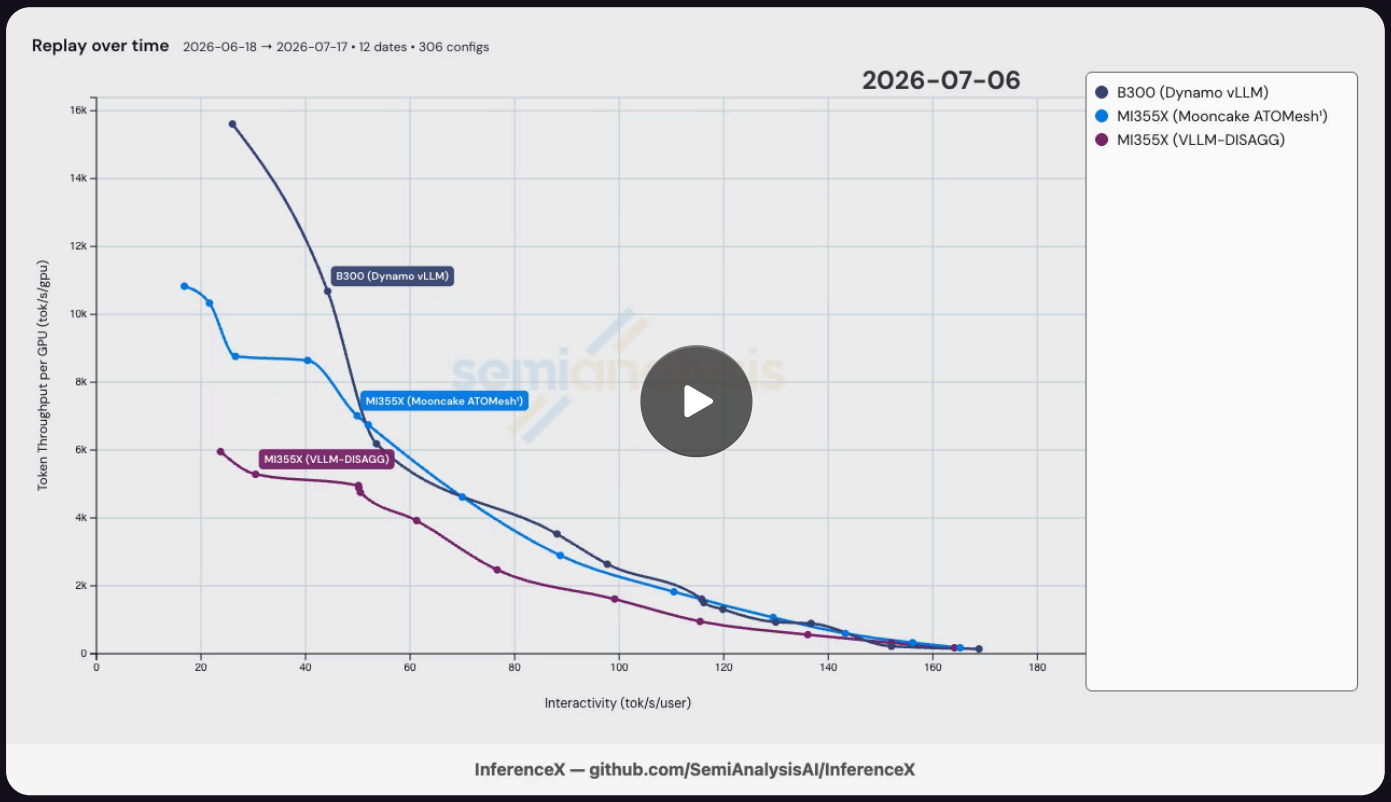
Source: SemiAnalysis InferenceX

For the disaggregated case, AMD is also doing quite well. This is a huge improvement from ~6 months ago, where the team struggled to achieve parity with Nvidia on DeepSeek R1 for many months. This is a result of a couple of things. First, AMD has made significant concrete improvements to their software stack with MoRI backend for distributed inference, as well as various improvements to Mooncake. More importantly, the AMD distributed inference team, led by Hai Xiao, has been working with a greater sense of urgency to provide better support for distributed inference.

The following video shows that MI355X FP4 disaggregated solution lagged many months behind Nvidia's.



Note that this was AMD's first public disagg recipe, posted on InferenceX in January. Take this in stark contrast with the Day 0 progress on MiniMax M3 FP4 disagg shown below, and you can see AMD is in a much better position now to be competitive in terms of distributed inference solutions.



Source: SemiAnalysis InferenceX

However, going fast isn't easy. Especially when the competition has a head start. The missing one piece is now real: AMD's distributed stack was missing above the engine came with ATOMesh . It is a ROCm-native distributed-inference gateway featuring a Rust routing and orchestration layer that handles the cluster-level work disaggregated serving needs, such as prefill/decode-disaggregated routing, cache-aware load balancing, RDMA KV-cache transfer (via MoRI-IO or Mooncake).

It is not a from-scratch build: ATOMesh is derived from SGLang's sgl-model-gateway, and substantially reworked around ATOM and AMD hardware. However, it works, as shown in the MiniMax M3 InferenceX results above. Routing policy stays separate from model execution, and it plugs into vLLM and SGLang as naturally as AMD's own ATOM engine.

Source: SemiAnalysis

Chipping Away at the CUDA Moat: Using Agents to Enable Day 0 Support

perf is an important baseline for showcasing performance *over time*, which is the north star goal of InferenceX.

Obviously, the 100x cracked engineers at AMD, Nvidia, Inferact, RadixArk, etc are responsible for the hardcore inference engineering work to truly enable day 0 support for these models. However, often times there are bugs to fix or low hanging fruit that the SemiAnalysis team adds in order to enable day 0 sweeps (currently, these are more prevalent with AMD configs). This has become increasingly easy with the rise of capable agents. The flow looks something like this: for a given config (I.e., MiniMax M3 vLLM MI355X FP8), start a new Claude Code/Codex agent to pull the day 0 recipe from the internet, create the necessary plumbing in InferenceX, and then kick off the sweeps. The agent has access to both the GitHub Action as well as direct access to the physical runner to continuously monitor the status. In the case of an engine error, the agent can automatically decipher what the root cause is, and either iterate automatically and re-run or consult the human for intervention. We can execute this pipeline in parallel for multiple configs on various SKUs.

When an error is identified, the current models are quite good at identifying the cause in the upstream engine code (vLLM/SGLang/TRT) and few-shot implementing a fix, with some guidance from the team. Additionally, we use agents to identify low-hanging fruit performance improvements. Some examples of upstream contributions by the SemiAnalysis team + agents:

- TRT: [fix] Fix fused MHC for DeepSeek-V4-Pro hidden size#13710: a Day 0 DeepSeek v4 fused MHC kernel fix
- [Bugfix] Fix NixlConnector handshake block_len validation for GQA-replicated KV heads#45879: MiniMax M3 disagg Day 0 enablement
- [Bug Fix] [MiniMax-M3] Implement EAGLE3 support on the AMD MiniMax M3#45546: enable speculative decoding for AMD MiniMax M3 Day 0
- [Bugfix][ROCm] Fix MiniMax-M3 FP8 KV cache dtype: enable FP8 KV cache support for MI300X and MI325X for Day 0 MiniMax M3
- More...

Shoutout to vLLM community Roger Wang, Hongxia, Michael Goin and other Inferact engineers for the kind mentorship and help on getting these fixes merged.

Source: GitHub

This kind of rapid iteration would simply not have been possible 3-6 months ago on a team of 2.5 engineers. Current frontier models in a capable harness are legit. They can make meaningful contributions to OSS serving engines and kernels. This may not *necessarily* be attributed to the "intelligence" of the models, rather the pure "grit" they have when told to achieve a goal. This combined with the ability to execute many of these tasks in parallel make it clear that testing and writing software is not the moat it was last year. This is broadly positive for AMD. By enabling AI agents to take on work previously performed by human engineers, it diminishes the relevance of the so-called "CUDA moat" and helps AMD offset Nvidia's advantage in engineering headcount.

AMD seems to be leaning into this thesis heavily, as at Advancing AI 2026, they announced ROCm.ai, a suite of skills, harnesses, and framework integrations that aims to further enable developers to iterate quickly on kernels and performance tuning using agents.

Source: AMD

GEAK, Hyperloom, and Optimizing Against InferenceX

ROCm.ai is more than just a keynote slide. Most pieces are already sitting in the open on AMD's AMD-AGI org, and they map almost one-to-one onto the day-0 loop we described above. At the center is GEAK ("Generating Efficient AI-Centric Kernels"), a mini-SWE-agent-based agent that writes and tunes Triton/HIP and FlyDSL kernels, wrapped by Hyperloom, the orchestrator that profiles a serving workload, finds the bottleneck kernels, throws GEAK and GEMM-tuning agents at them, and gates every candidate on an end-to-end A/B before it counts. Around that sit the supporting cast: Magpie for evaluation, TraceLens for trace analysis, Apex for exporting agent trajectories into an RL-flavored training pipeline, and AgentKernelArena, a head-to-head harness that pits Claude Code, Codex, Cursor, and GEAK against each other on the same kernel tasks under identical scoring.

Hyperloom's optimizer pulls its performance target straight from InferenceX. On the current main branch, it scraps InferenceX, and target_analyzer writes a competitor target that the agent then tries to clear. An earlier version of their CI went further — computing a % gain over InferenceX, counting how many models "beat" InferenceX, and posting the scoreboard to a Teams/Slack channel — before that machinery was trimmed for the open-source release.

We see this as a good thing. Our vision for InferenceX is to accurately track performance of open-source frameworks, and it definitely serves this role if agents use it as a reward signal and continuously try to improve. These efforts will benefit the ML community with better model serving performance.

of the engineering is anti-cheat. GEAK had to stop silently scoring the unpatched baseline kernel instead of the agent's actual patch, and add a GEAK_PROTECT_TEST_FILES mode that strips the agent's edits to the test harness so it can't fake correctness by rewriting the reference. Apex ships a tamper detector that flags hardcoded print("PASS") and a banned-library list so an agent can't "optimize" a Triton kernel by quietly routing to a pre-tuned MIOpen or hipBLASLt call. It's the same grit-over-intelligence story: the models will reward-hack a benchmark the instant you let them, and a surprising share of the work is building the guardrails that keep the speedups real.

And it works. GEAK's learned notes already log verified end-to-end wins on shipping silicon. ~+21.8% e2e from an MXFP8 decode-bound dense-linear rewrite on MI355X, with honest caveats where grouped-MoE GEMM stalls out around a 1.1x ceiling. None of this defeats the CUDA moat on its own. But it is the clearest signal yet that AMD is industrializing the exact workflow that, a year ago, required a room full of CUDA engineers.

Introducing AgentX, a new InferenceX Agentic Scenario

While the current 8k1k + 1k1k scenarios are a great proxy for evaluating baseline chip performance, they fail to evaluate the entire system (routers, KV cache transfer, KV cache offloading, schedulers, etc) since they are all single-turn, random data requests.

For the past few months, we have been working with industry leaders to develop an agentic benchmark: AgentX. To achieve our goal of making the benchmark as reflective as possible of real world serving, we collected ~3 months of internal SemiAnalysis Claude Code, Codex, etc. traces, which we then replay offline at varying concurrencies using AIPerf. Some of the statistics of the dataset are shown below. Notably, the median ISL and OSL are 140k and 396, respectively (quite different from 8k1k). With a median cache hit rate of 99.2%! Note, this is assuming an infinite cache so is not, of course, actually attainable in most real-world serving.

The dataset also includes realistic subagent usage as well as dynamic workflows, which further stresses the KV cache by continuously introducing uncached context for the server to process. The tool use time / user think time (captured by inter-turn latency) is also reflected in the dataset, further stressing the KV cache and highlighting the efficacy of KV offloading techniques, which have the possibility of increasing KV cache TTL.

Source: SemiAnalysis HuggingFace

We are very excited about this scenario. We have worked with WEKA, Inferact, RadixArk, LMCache, Mooncake, NVIDIA, AMD, and other industry leaders in ensuring it accurately reflects real traffic that may be experienced by a TaaS provider or frontier lab like OAI/Ant.

Results have been rolling in for the past month and are available on InferenceX. We will write an in-depth article on results and methodology soon.

Single-Node is Out, Disaggregated is In

As discussed, the competitive frontier has moved from single node aggregated to multi node disaggregated inference. In our InferenceX v2 article, we discussed the notion of AMD's "software composability problem" when it came to distributed inference. That is, they had made strides in individual optimizations such as disagg, FP4, WideEP, DP-attention, etc, but combining many of these together broke the stack. While there is still certainly progress to be made, AMD has made significant strides in this area.

bound, while FFN is stateless and batch-scalable, these phases can be split and mapped onto the best-suited hardware. Then, inference leadership becomes a distributed-systems problem rather than a single-kernel one. WideEP, disaggregated prefill, KV transport, expert routing, and scheduler/orchestration are all part of that same new moat. AMD can close single-node gaps, but unless those optimizations compose cleanly in vLLM and SGLang, it is still fighting yesterday's war.

The timeline below puts the gap in perspective. The open CUDA ecosystem has been shipping disagg+WideEP since early 2024, while AMD's first publicly available PD-disagg + WideEP recipes only landed in January 2026 via InferenceX. AFD's real payoff requires superpod-class interconnects, which are the kind of rack-scale hardware AMD does not ship until MI455X. And at GTC 2026, Nvidia announced the Groq 3 LPX, an SRAM-based LPU integrated into the Nvidia inference rack specifically for disaggregated decode FFN.

Source: SemiAnalysis

The New Moat: Sparse Models, Disaggregation, and WideEP

Recent Model Trends: Active Experts Stay Tiny While Total Experts Explode

The model landscape makes the composability problem more urgent. Active expert counts stay pinned in the 4–10 range while total experts scale out to 128, 256, 384, and now 512. Many of the newest giant MoEs also collapse the KV cache to the equivalent of one or two heads via MLA-style low-rank designs. That means the frontier is getting sparser, more

float. Composable distributed inference is.

Disaggregated Prefill and WideEP

Disaggregated prefill separates compute-heavy prefill from memory-heavy decode onto different nodes, removing the interference between the two phases and letting operators tune prefill vs. decode node counts independently. The trade-off is KV-cache transfer over RDMA. We estimate roughly 290 MB for an 8,192-token DeepSeek-R1 FP8-KV prefill.

Currently on InferenceX, AMD's disaggregated inference configs often perform either worse or only slightly better than their single node counterparts. For DeepSeek-R1, iso-interactivity throughput for MoRI SGL outperforms the aggregated SGL config by 2×–3×.

Source: SemiAnalysis InferenceX

WideEP spreads experts across more GPUs and nodes instead of replicating full expert islands. On DeepSeek-R1 (256 routed experts), moving from EP8 on a single 8-GPU node to EP64 across 64 GPUs cuts experts per GPU from 32 to 4, freeing HBM for KV cache, enabling larger concurrent batches (and thus more tokens per expert per GEMM), and scaling aggregate HBM bandwidth with cluster size. The payoff is not subtle: NVIDIA reports up to 2.28× higher per-GPU output throughput from Wide-EP (DeepSeek-V3.2, NVFP4, GB200 NVL72, EP16/EP32 vs EP4/EP8). As we emphasized in InferenceX v2,

Put differently, ATOM and AITER are useful only insofar as they improve the mainstream ecosystem. We have argued that ATOM may help single-node performance but still lacks production features such as NVMe or CPU KV-cache offload, tool parsing, and WideEP. The right strategy is not to build an AMD-only island. It is to upstream the kernels, harden CI, publish recipes, and make FP4 + WideEP + disagg + cache offload work together in the default open-source stacks. That is how the CUDA moat gets eroded now.

Disaggregated Inference Accuracy

Before March 2026, DeepSeekR1 disaggregated configs with DP-attention(DPA) failed accuracy evals with near 0 GSM8K scores, while no-DPA sweeps passed GSM8K at >95%. This was caught with InferenceX eval runs. This shows that no-DPA disaggregated path can preserve correctness, but once DPA is used, the stack breaks, showing AMD's composition problem again. This issue has, however, been fixed.

One issue that has not been fixed is EP at certain batch sizes in SGLang with the MoRI backend. On DeepSeek-R1, decode with concurrency 64 drops GSM8K to ~80% against a ~94% baseline that holds at every other concurrency. The catastrophic version of this bug, where the same low-concurrency path silently produced fluent-but-wrong output and scored 0 on GSM8K, was traced to an uninitialized reduce buffer in AITER's FP4 MoE kernel and patched in June; the residual ~80% cliff at concurrency 64 is still open, with AMD attributing it to numerical corner cases in the quantization kernels that only that batch size exercises and declining to prioritize a fix.

Source: SemiAnalysis InferenceX

Composition Works... Sometimes

AMD has now demonstrated that several of these optimizations can compose: SGLang's MI355X nightly images cover DeepSeek-V4 disaggregation with DP-attention, EP8, and MTP, while four-node DeepSeek-V4 and Kimi WideEP16 configurations have been validated on hardware. But composition remains model- and topology-specific rather than dependable by default. Kimi's DP8/EP8 path was excluded after a GPU memory fault, its WideEP16 int4 decode path currently must disable HIP graph capture, and the WideEP16 CI changes remain under review. The gap is no longer that the individual pieces never work together. It is that AMD cannot yet assume they will work together across models, quantization formats, parallelism strategies, speculative decoding, and network topologies without special cases.

Disaggregated Inference Has to Become Everyone's Job

Disaggregated inference cannot sit in a small specialist corner of AMD. AMD's own ROCm tutorial says single-node optimization is starting to hit limits, and that distributed inference is becoming more important; ROCm followed that with official Mooncake PD-disaggregation docs in December 2025 and MoRI-based distributed inference docs in January 2026. Once the official stack is publishing xP+yD topologies, RDMA requirements, 1P2D recipes, and KV-transfer frameworks, DI is not a side quest. DI is the product. It must

That is the AMD software story so far. The single-node path is getting fixed. MoRI is promising. The overlap stack is finally showing up. But the market is already on to SBO, PD disaggregation, load balancing, and wider EP configurations that must compose cleanly under real production constraints. AMD no longer needs one more hero optimization. It needs the entire distributed inference stack to start moving as one system.

Where AMD Stands: The Distributed Stack, Piece by Piece

MoRI: A Real Win, Built by One Tiger Team 🐯

MoRI, to AMD's credit, is a real win. MoRI is a modular RDMA framework with two dedicated components: MoRI-EP for expert dispatch/combine and MoRI-IO for KV transfer. ROCm has now published official MoRI-based distributed SGLang documentation for MI355X clusters. It is built from first principles by a China-based engineering team. We support the direction, but it needs much more open CI and testing.

AMD's own MoRI roadmap makes the scale problem concrete. The H2 2026 roadmap lays out an ambitious program across three subsystems: a tiered distributed KV cache (HBM→DRAM→NVMe via SPDK/GDS) with scheduler co-design, a next-generation SHMEM v2, EP v2 kernels in FlyDSL/C++ with elastic expert-parallelism and fault tolerance, "mega kernel" codesign with FlyDSL, and rack-level Helios enablement. Topped off with SGLang and vLLM upstreaming. The whole list is owned by the same five or six engineers whose handles dominate MoRI's merged pull requests. It is the right roadmap. It is also, as of publication, entirely still ahead of the team — resting on the same handful of shoulders.

The Newest Silicon Shows the Gap: Helios (gfx1250)

The clearest measure of how far the distributed stack still has to go is AMD's newest hardware. Across every layer of the Helios (gfx1250 / MI455X) stack the same shape repeats: single-model arch-enablement and KV-transfer plumbing are landing fast, while WideEP, validation, and the wave32-tuned high-value kernels are not there yet.

Start at the foundation. PyTorch's own gfx1250 support merged in mid-July (ROCm/pytorch #3421, a cherry-pick of an upstream change that landed, was reverted a day later, and re-

deferring a test runner to a follow-up that does not exist. It is a skeleton: the highest-value paths (Composable Kernel GEMM and SDPA, FP8 grouped GEMM, int4) are filtered out or hard-error, attention is a “Tech Preview,” and wave32 kernels are unwritten. AMD’s code even brands the part “CDNA5” while describing a GFX12.5, wave32, WMMA execution model, the clearest in-tree measure of how far Helios sits from the CDNA wave64 lineage its kernels were written for.

Source: https://github.com/ROCm/pytorch/pull/3421/changes#diff-9b18bbaca027737173099a4fd1766ad9ce03c982f430852a09cc6c76aa365bb4R191

One layer up, the raw MoE kernels are the fastest-moving piece of the stack. AMD’s aiter and FlyDSL libraries have absorbed well over a hundred gfx1250 PRs since spring: wave32/WMMA GEMMs, MXFP4/A8W4 quant, MLA-v4 attention, TDM data-movement atoms. Roughly a quarter of them closed without merging. The centerpiece is the AMD-native “MegaMoE” path: a fused MoE expert-parallel flow whose dispatch/combine op-layer was vendored into aiter and whose gfx1250 grouped-GEMM kernel(global→local expert remap, fused route/scatter, and a masked grouped GEMM tuned for DeepSeek-V4’s 384-expert MoE) is still-open, validated for a8w4 (FP8-activation, MXFP4-weight) only. This is the WideEP MoE primitive the stack needs; it just isn’t finished, and the gpt-oss, R1, and V4 kernels each land on their own track.

The break is at the integration layer, where those kernels meet the frameworks. SGLang’s July 22 gfx1250 nightly only builds and publishes the image. No test job, no accuracy gate, built on a generic runner and merely mirrored through an MI300, because no gfx1250 runner exists upstream. That image builds MoRI but pins a June commit predating gfx1250 support, so MoRI-EP WideEP does not run: the kernels carry no matrix-core or wave64-specific ISA, but MoRI’s build gates only gfx942/gfx950 and the wave32 support AMD has

transfer half is architecture-agnostic host-side RDMA and would move KV cache on gfx1250 today, yet the actual gfx1250 recipes for DeepSeek-V4 and R1 are single-node tensor-parallel with neither WideEP nor disaggregation.

vLLM's gfx1250 bring-up tells the same story from the other side: four models on the FFM simulator and early silicon with volatile results, dedicated perf work only for gpt-oss, whose "ATOM-parity" tuning pass was itself reverted "until AITER is ready". MoRI was also removed from the build outright. It is a credible single-node FP4 bring-up; but it is not the distributed stack

Source: Source: https://github.com/vllm-project/vllm/pull/46516

Even AMD's own engine follows the pattern. ATOM's Helios work is disaggregation-first and expert-parallel-later: ATOM adds DeepSeek-V4 MoRI-IO write-push KV transfer plus a UALink scale-up fabric backend — the Helios-generation interconnect — so the KV-transport half of disaggregation is being built for exactly the model and fabric Helios will run, though its e2e validation ran on eight MI355X nodes and a UALink vPOD rather than MI455x. But it carries no EP dispatch/combine, no MoRI-EP, no WideEP; ATOM's shipped DeepSeek-V4 recipe still serves the model at TP=8 on a single node, and every one of its MoRI-EP PRs targets last-generation gfx942/gfx950. Bottom to top:

PyTorch, kernels, MoRI, frameworks, ATOM, Helios has arch enablement and the first half of disaggregation, and no WideEP anywhere.

At Nvidia's GTC 2025 NCCL session, we asked if Nvidia would provide support to the AMD communication library fork due to this big refactor in the upcoming Nvidia's communication library. Nvidia explicitly said it would not help AMD's communication team adapt to the upcoming NVIDIA library refactor, and that NVIDIA doesn't participate in AMD's communication development at all

A year later, at Nvidia's GTC 2026 Dynamo session, we asked a question: NIXL had already accepted upstream contributions from Trainium's Neuron fork, so would it accept them from AMD's RIXL fork too? The maintainers said yes, on the record, in front of a room.

AMD had been carrying RIXL downstream, burning engineering hours maintaining a parallel copy of infrastructure everyone else gets for free. The only thing standing between AMD and upstream was whether anyone would try. NVIDIA had just removed the excuse.

So we spent April making sure AMD knew it. We put the question to Stephen Bates publicly on LinkedIn, offered Anush help to connect AMD directly to the NIXL maintainers, and asked Anush about MoRI/RIXL upstreaming in person during TensorWave's beyond summit and we helped AMD form the relationship with NVIDIA so that they can upstream the patches.

On May 15, AMD's Andy Luo opened a PR for ROCm/HIP build support for gfx942 (MI300X, MI325X) and gfx950 (MI350X, MI355X), behind a default-off use_rocm Meson flag. It merged June 4 after several rounds of review with NVIDIA-side maintainers.

Andy's PR cleared review by proving zero blast radius on the CUDA path. On the AMD side it validated end-to-end VRAM transfers over NIXL+UCX on MI300X and MI355X. The stacked follow-up, #1647, hit 341 Gb/s cross-node RDMA on two MI355X nodes using AMD's own AINIC RoCE NICs and libionic driver, with no Mellanox NIC in the path. Today, RIXL has been fully ported to NIXL. Dynamo core has since started taking AMD patches too.

Source: https://github.com/ROCm/RIXL

NIXL is the transport layer for disaggregated serving, so if AMD wants PD disagg and WideEP to be dependable rather than model-specific, KV transfer has to work on Instinct in the frameworks people actually deploy, not just inside MoRI. Nobody at AMD should have needed us to point that out. The gaps that remain say the same thing: nixlbench HIP support and the ROCm Dockerfile/CI surface landed as separate follow-ups, an AMD GPU CI runner is still an offer rather than a merged workflow, and rocSHMEM is a future plugin.

The Overlap Stack Is Late, and the Market Has Already Moved On

Two-batch overlap (TBO) should have been table stakes a long time ago. SGLang put overlapping two batches and expert parallelism on the 2025 H1 roadmap, and the public docs now expose `--enable-two-batch-overlap` with claims of up to 2× throughput. AMD's own November 2025 DeepSeek-on-MI300X writeup still described dual-batch overlap as an essential feature that was "still being developed," with future work centered on dual-batch overlap and expert load balancing. MORI EP support for two-batch overlap only merged into SGLang on February 20, 2026. That is late.

The maturity gap showed up immediately. Days after merging, a maintainer flagged the PR for breaking CI and asked for it to be reverted and reopened after fixes. That is the broader ROCm inference story in miniature: the ingredients are increasingly there, but too many key optimizations still arrive late, land fragile, and need another turn before they become something customers can trust by default.

Part 3: AMD Competitive on Total Cost of Ownership & up to 105% Discount for