Episode Book

EPISODE #5898

ROCm 10.1: AMD's Software Stack Finally Grows Up

PUBLISHED	12 October 2026 at 12:14 UTC
DURATION	27:15
LISTEN	myweirdprompts.com/episode/rocm-10-amd-gpu-software-14533/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.4
SOURCES	16 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:41 UTC

SYNOPSIS

AMD crossed a trillion dollars in market cap and its software stack is finally getting the attention. We dig into three ROCm releases — 7.14, 10, and 10.1 — and ask the honest question: is ROCm actually production-viable for inference now, or just for the data center? From TheRock's reproducible builds to Hyperloom's autonomous optimizer to NUMA-aware memory allocation, we trace what changed, what's still marketing, and why the 7900 XTX owner is still the person the roadmap forgot.

TRANSCRIPT



Corn

AMD crossed a trillion dollars in market cap for the first time about three weeks ago. Ninth-best-performing stock in the S and P this year, up nearly two hundred percent.



Herman

And the reason that belongs in this episode and not a finance show is that all that money is the answer to a question Daniel sent in about his graphics card.



Corn

Daniel built his current desktop six months ago. There was no way he could afford the Nvidia equivalent at the VRAM he wanted, so he ended up on AMD. He said it plainly, that he didn't quite appreciate at the time how far behind AMD was. But he watches the ROCm repository on GitHub, he sees inference engines landing constantly, and he wants to know what's actually changed since we last dug into this.



Herman

The honest answer is more than I expected. And less than the release notes claim.



Corn

Start with what ROCm is, because half our listeners own an AMD card and have never once typed the word.



Herman

ROCm is AMD's open-source stack of drivers, tools and APIs for GPU programming. Underneath it sits HIP, which is a C plus plus runtime and kernel language, and its interface is deliberately similar to CUDA. That similarity was the original sin. HIP worked as a translation bridge over CUDA's APIs, so developers could port code with find-and-replace instead of a rewrite. In practice, the find-and-replace ported the code and none of the assumptions, and getting the stack to actually run meant Docker images, environment variables, and luck.



Corn

Luck is not a build step.



Herman

Luck was the build step. That's the gap people describe when they say ROCm lags. It was never the silicon. On paper the cards were competitive years ago. The friction was the software layer refusing to hold still.



Corn

Give me an example of what that actually looked like, because "refusing to hold still" is abstract.



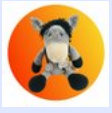
Herman

Here's the canonical one. You'd install a PyTorch build that AMD published for one ROCm version, and it would expect a specific driver, a specific kernel, and a specific set of environment variables like HSA_OVERRIDE_GFX_VERSION. Then you'd update your distro, the kernel would move, and the whole thing would fall over. Not with an error that told you what broke. With a segfault, or worse, a silent wrong answer out of a matrix multiply.



Corn

Silent wrong answers are the worst possible failure mode.



Herman

They are, because you don't find out until your model outputs garbage three hours into a training run. That's the class of problem people mean when they say the stack was immature. It wasn't that it couldn't do the math. It's that you couldn't trust it to keep doing the math after a Tuesday update.



Corn

Where is it now, honestly, without the marketing?



Herman

Production-viable for inference on a 7900 XTX. Meaningfully slower than CUDA though, in the range of ten to twenty-five percent depending on the workload. Training and image generation still favor Nvidia more heavily. That's the baseline going in. The six months since we last looked at this are a story of three releases, and they each have a different theme.



Corn

Give me the arc before the details.



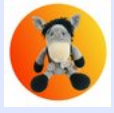
Herman

July is the build system. August is the AI-native developer layer. October is data movement. ROCm 7.14 in the middle of July is where TheRock goes to production. TheRock is AMD's automated build and release system, and it means AMD ships one reproducible ROCm instead of a pile of vendor forks that each drifted in a different direction. It splits the stack into a Core SDK plus expansion SDKs for high performance computing, computer vision, data science, and life sciences. SGLang support arrives on Radeon. PyTorch 2.12, JAX 0.10.0, vLLM 0.23.0. New collective algorithms in RCCL. And hipFile, the direct-storage library, reaches production.



Corn

Translation on hipFile, for people who don't read release notes for fun.



Herman

It lets the GPU read from a disk directly instead of routing everything through host memory first. If you're doing inference over a model that doesn't fit in VRAM, that's the difference between streaming weights and watching a progress bar. Boring name, real consequence.



Corn

And "one reproducible ROCm" sounds like a build-engineering detail. Why should a normal person care?



Herman

Because it's the difference between "this works on my machine" and "this works." When every vendor fork drifts, the bug reports stop being reproducible. Someone says vLLM is broken on ROCm and someone else says it works fine, and both are telling the truth, because they're on different forks with different patches. TheRock collapses that. One build, one set of artifacts, one thing to file a bug against. It's unglamorous and it's the precondition for everything else in this episode.



Corn

And August.



Herman

August 27 is ROCm 10. The tenth-anniversary release. ROCm.AI goes generally available with three pieces. AMD Skills, which is a catalog of agent skills. ROCm CLI. And Hyperloom, which is an autonomous agentic optimizer. AMD claims 3.3x inference uplift and 2.4x training uplift over ROCm 7 on an 8-card MI355X node. They also add AMD Primus for training, make the Core SDK more modular, and consolidate Windows under it.



Corn

Hyperloom. An optimizer that runs itself. What does it actually optimize?



Herman

Kernels, scheduling, memory placement. It searches configurations the way a human engineer would, except it doesn't sleep. Which is the part I want to flag early, because it comes back in a minute.



Corn

So it's an autotuner with a bigger search space.



Herman

It's an autotuner that also gets to move where the memory lives and how the work is scheduled across the eight cards. A normal autotuner picks a tile size for one kernel. Hyperloom gets to reorganize the whole node's execution plan. That's the ambition, anyway. Whether it delivers is a separate question.



Corn

And October is the one that shipped this week.



Herman

ROCm 10.1, landing right around now. The theme is the data-movement bottleneck. hipFILE gets an async fast path and batch I O. Host memory allocation becomes NUMA-aware. The compiler jumps to LLVM 24. hipThreads arrives, which lets you take ordinary CPU-threaded code and accelerate it on the GPU incrementally rather than rewriting it. MIGraphX switches its backend from rocMLIR to rocMLIR-Triton. WSL2 gets a tech preview. And ROCm SMI is removed in favor of amd-smi.



Corn

Someone at AMD had to argue for that last one in a meeting.



Herman

Someone has been arguing about it since the last time we covered this topic, I'd wager.



Corn

NUMA-aware host allocation. Explain why that's not just another bullet.



Herman

On a multi-socket server, memory is not one pool. A CPU can reach its own memory fast and the other socket's memory slow. If the GPU driver allocates host memory on the wrong socket, every transfer pays a tax. NUMA-aware allocation means the driver puts the buffer next to the GPU that's going to read it. It's a latency fix, and on big nodes latency is throughput.



Corn

So the October release is basically "stop wasting time moving bytes around."



Herman

Every bullet in it is about bytes arriving sooner. hipFILE, NUMA, hipThreads, the compiler. It's a release about plumbing, and plumbing is what determines whether the fancy parts get to run at full speed.



Corn

So that's the release cadence. Daniel said he watches GitHub and sees inference engines rolling out constantly. What is he actually seeing?



Herman

This is the part where I'd like to be careful, because the impression and the reality differ. There is no standalone consumer-focused ROCm inference engine. It isn't a new AMD-built engine competing with vLLM. What AMD is shipping is integrations and validated containers for the existing engines. vLLM, SGLang, llama.cpp, Ollama. Plus the AMD Skills catalog, which is a federated repository of agent skills for Claude Code, Cursor, Codex and Gemini CLI.



Corn

So the "rolling out inference engines" is real, it's just not what it sounds like.



Herman

It's plumbing, not a product. And the plumbing is useful. There's a skill called serving-llms-on-instinct, one called hyperloom-workload-optimizer, one called rocm-doctor that diagnoses your install, quark-torch-llm-ptq for quantization. You point a coding agent at these and it wires up your stack for you. There's also real work on SGLang wheels, multi-architecture builds with S3-based distribution.



Corn

Walk me through rocm-doctor, because I want to know what a diagnostic skill actually does.



Herman

You run it and it checks the things that used to eat an afternoon. Which driver is loaded, whether the kernel module matches the userspace library, whether your user is in the right groups, whether the GPU is visible to the container, whether the ROCm version the wheel was built against matches what's installed. Then it tells you which one is wrong and what to do about it.



Corn

That's a support ticket turned into a script.



Herman

That's exactly what it is. And the fact that it exists tells you how many support tickets there were.



Corn

The doctor one is the tell. If the tooling needs a diagnostic skill, that says something about what the install experience has been.



Herman

It says the install experience has been bad enough to justify building a specialist. I'd rather they ship the fix than pretend it wasn't needed.



Corn

What changed for you in reading all this? Not the feature list. The thing that reframed it.



Herman

The reframe is that ROCm's weaknesses and ROCm's wins are in the same place. Every one of these releases is aimed at a data-center operator. Hyperloom optimizes an eight-card node. The MLPerf numbers are multi-node. The validated containers are for Instinct hardware. Nobody shipped anything this year that a person with a 7900 XTX in a desktop case would call a gift.



Corn

Then say the quiet part. The gap is closing fastest where AMD competes hardest.



Herman

Exactly there. And that's the thesis of the episode, really. In the data center AMD is not catching up. It has caught up on several workloads. On MLPerf Inference 6.0, MI355X ties B200 in the Offline scenario, hits ninety-seven percent on Server, and beats it at a hundred and nineteen percent on Interactive. Against B300 it's ninety-three percent Server, ninety-two percent Offline, and a hundred and four percent Interactive. On GPT-OSS-120B it hit a hundred and eleven percent of B200 Offline and a hundred and fifteen percent Server.



Corn

That's not a second option. That's a coin flip.



Herman

On those specific tests, yes. And at multinode scale MI355X crossed a million tokens per second on Llama 2 70B in both Server and Offline. The generational jump from MI325X on Llama 2 70B Server is a hundred thousand tokens per second versus thirty-two thousand. Three point one times in one generation.



Corn

So the silicon and the software at the top of the stack are both fine. What's the consumer side of that same number?



Herman

Ten to twenty-five percent behind CUDA. The same company, the same six months. The farther you get from a hyperscaler buying racks, the worse the experience gets.



Corn

Which is exactly where Daniel is standing. He's the person the roadmap forgot.



Herman

He's the person the roadmap is priced for. Which is the other half of this and it's the more interesting half.



Corn

Go.



Herman

The consumer reason to buy AMD hasn't changed and it's the reason Daniel bought. VRAM per dollar. A 7900 XTX at around nine hundred dollars new gets you twenty-four gigabytes. It's a real alternative to a used 4090 at thirteen hundred. Against the RTX 5070 Ti at seven hundred forty-nine dollars with sixteen gigabytes, the AMD card beats it on price per gigabyte of VRAM outright.



Corn

Eight gigs for a hundred and fifty dollars is the whole argument.



Herman

That's the whole argument for anyone running models that don't fit in sixteen gigabytes. And there's no AMD answer above that. RDNA 4 doesn't target the consumer tier above thirty-two gigabytes, so there is no consumer AMD card that competes with a 5090. If you want a lot of VRAM cheap, AMD is the only game in town. If you want the most VRAM, AMD doesn't sell it to you at all.



Corn

So a person running a seventy-billion-parameter model at four-bit quantization, which is roughly forty gigabytes, is stuck no matter what.



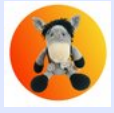
Herman

Stuck on one card. They can split across two AMD cards, which is its own adventure, or they buy Nvidia and pay for it. That's the ceiling and it's a hard ceiling. AMD's answer to the top of the consumer market is to not participate.



Corn

And in the data center the same math runs at a different scale.



Herman

MI300X at a hundred and ninety-two gigabytes rents at about a cent per gigabyte. An H100 at eighty gigabytes is roughly two cents. Half the price per gigabyte of VRAM. That's before any of the warrant deals, just the list price.



Corn

Hold that number. I'm going to come back to it.



Herman

The other thing worth saying out loud is MI355X itself. Two hundred eighty-eight gigabytes of HBM3E. Ten petaflops at FP4 and FP6. Supports models up to five hundred twenty billion parameters on a single GPU. That last figure is the one that matters for the second-option thesis, because it means a large model fits on one card, and one card is a thing a mid-sized company can actually buy.



Herman

Think of a regional hospital system or a university lab. They have a budget in the low six figures for compute, not the low eight figures. A rack of eight B200s is not in the conversation. One MI355X with three hundred gigabytes of memory is, because the model fits and they don't have to solve distributed inference to use it. That's the customer AMD is actually winning, and it's a customer Nvidia structurally doesn't prioritize.



Corn

So the technical case is made. Now explain why AMD has an incentive to fix any of it, because six months ago they also had an incentive and nothing moved this fast.



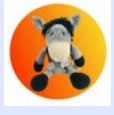
Herman

The incentive changed scale. September 21 is the trillion-dollar market cap. Shares rose nine point six percent that day to a record six hundred thirteen dollars, and closed at six hundred thirty-three ninety-one in early October. Fourth US chipmaker to a trillion. Up nearly two hundred percent this year. That's the market paying in advance for exactly the thesis Daniel is asking about.



Corn

And the deals behind that valuation are the interesting part.



Herman

Anthropic will deploy two gigawatts of AMD Instinct chips, and AMD may invest up to five billion dollars in Anthropic. OpenAI holds warrants for a hundred sixty million AMD shares tied to a six-gigawatt MI450 deployment. Oracle committed to fifty thousand MI450 GPUs starting this quarter.



Corn

Warrants. So AMD is not just selling chips. It's paying customers to take them.



Herman

That's the uncomfortable angle and it's the freshest thing in this whole story. SemiAnalysis looked at the OpenAI and Meta deals and calculated that the warrant structure amounts to roughly a hundred and five percent equity rebate discount. Up to three hundred twenty million shares at a penny each, with the final six-hundred-dollar price gate already cleared.



Corn

A hundred and five percent of a rebate. You can't discount more than the thing costs unless you're paying them.



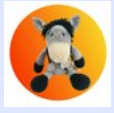
Herman

Which is why SemiAnalysis described the cost per million tokens on AMD's Helios rack as practically negative. AMD is effectively buying its own demand. That is a completely different incentive structure from "build better software so people choose us."



Corn

It's a subsidy with a share price attached.



Herman

It's a bet that if enough of the industry is running on your hardware, the software gap stops mattering, because everyone optimizing for AMD becomes everyone else's problem to solve.



Corn

Does that argument work, or does it just move the problem?



Herman

It works if the software keeps improving, and that's the part I'm less sure about. Because the constraint isn't engineering talent. SemiAnalysis, who by their own description is the number one ROCm bug reporter on the internet, says AMD's best engineers are in Shanghai and they're starved of stable GPU clusters. The quote is that leadership still needs to update their viewpoint toward providing stable CI clusters that don't randomly need to be migrated and shifted from one cloud provider to another.



Corn

You cannot fix a software stack if your test machines keep walking away.



Herman

You cannot even measure it. And there's a ripple effect in there. AMD is leaning into what they call an agentic-oriented engineering culture. Anthropic's Tom Brown reportedly got Claude's inference stack running on AMD hardware over a weekend. Every one of those coding agents needs GPUs to test against, and those GPUs are the same cluster that's being moved between providers.



Corn

So the thing that accelerates development is eating the capacity you need to develop.



Herman

That's the tension nobody has a clean answer for.



Corn

Let me push on the Tom Brown story, because "over a weekend" sounds like marketing.



Herman

It's a data point, not a benchmark. One engineer, one stack, one weekend, with the full weight of Anthropic's motivation behind it. What it demonstrates is that the port is possible. What it doesn't demonstrate is that it's repeatable by someone without that



Herman (continued)

motivation or that support. That's the difference between a proof of concept and a product.



Corn

And the reason it's in the story at all is that it's the counter-narrative to "ROCm can't be made to work."



Herman

Right. Both things are true. It can be made to work. And making it work is still a project.



Corn

What's the mechanism that actually fixes software quality, though? Not vibes. The specific thing.



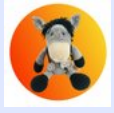
Herman

Gating tests. A pull request cannot merge unless the test passes. That's it. It sounds bureaucratic and it's the whole ballgame. SemiAnalysis put it as gating parity and gating pass rate being the metrics that truly matter. If a change breaks a workload, it doesn't land. Over enough of those, the stack stops breaking on workloads people care about.



Corn

Which requires stable machines to run the gate on.



Herman

Which is the bottleneck. It's a capacity-planning problem wearing an engineering costume. That reframed the last six months for me. The features are real. The constraint is unglamorous.



Corn

There's one more thing in the releases I want to pull on, because it bothers me.



Herman

Go ahead.



Corn

Hyperloom. An autonomous optimizer that runs while nobody's watching. I'm the guy who asks what the agent does when nobody checks on it.



Herman

You are exactly that guy.



Corn

I'm the target market for it and I don't trust it. It search-tunes kernels overnight and I wake up to a stack I didn't choose. I'd rather own a slower thing that stays the same.



Herman

There's a real answer to that. Optimizers like Hyperloom are supposed to produce a config you can inspect, not a black box that mutates underneath you. The output is a set of kernel parameters and scheduling decisions, and those are reviewable.



Corn

Supposed to.



Herman

I haven't seen anyone reproduce a Hyperloom search end to end and diff the output. So my confidence is moderate, and I'd rather say that than pretend.



Corn

Fair.



Herman

Here's the honest summary, though. Nvidia is still the default. AMD is now a credible answer rather than a compromise. That's the sentence. And the sharpest version of it came from a comparison that put it this way. If your stack depends on CUDA-specific libraries, ROCm 10 is not a migration path today. If your stack is PyTorch plus vLLM or SGLang on Instinct hardware, the case is substantially stronger than a year ago.



Corn

So the migration case is stack-shaped, not hardware-shaped.



Herman

It always was. The hardware stopped being the problem years ago. The question is which libraries you can't leave behind.



Corn

Two hundred and eighty-eight gigabytes of HBM3E.



Hilbert

Three hundred.



Corn

Sorry?



Hilbert

It's three hundred. The MI355X is three hundred and four. Sixteen gimmicked up some numbers for the launch and the real figure is three hundred and four, I've got it written down in a spreadsheet going back fourteen cards.



Herman

Okay, Hilbert, I'll take the correction. Where'd the spreadsheet come from?



Hilbert

My brother-in-law in Chicago. He ran procurement for a machining shop that did their own CAM work, and he was buying cards off the grey market whenever a render farm went under. So I'm tracking price per gigabyte for him, because he doesn't trust a number he didn't compute himself. Neither do I.



Corn

Fourteen cards.



Hilbert

Since the 3090 came out. It's a column, it's a tab, it's not a program, Corn, you'd approve of it, it's very slow. The reason I'm raising it at all is you two are talking about this like the consumer story is a footnote and it isn't. The card Daniel bought is the only story that matters for anybody who isn't buying racks, because he's the one paying retail and living with the driver.



Herman

I don't disagree, but the data-center money is what's paying for the driver he's living with.



Hilbert

The money's paying for a rack in Ohio. His desktop doesn't get faster because Anthropic signed a letter. What he got from the last six months is a better installer and a doctor skill that tells him why the installer failed.



Corn

That's bleak and I think you're right.



Hilbert

I'm not being bleak. A hundred and five percent rebate, one of you said earlier. That's not a sale, that's a dowry. Nobody on that spreadsheet of mine has ever been offered a hundred and five percent off a card. If your brother-in-law built his shop on hardware that only exists because the seller is paying the buyer, you'd want a second opinion from someone who checks the invoice.



Herman

The invoice being the CI cluster.



Hilbert

The invoice being the thing that keeps the driver from eating the weekend. Herman, you told me a year ago you'd rather have the version with the tests.



Herman

I stand by that.



Hilbert

Then stand by it here, because that's the part they didn't sell. The features are the announcement. The test cluster is the product.



Corn

Which puts us back at the same place. The consumer gets the leftover.



Hilbert

The consumer gets a card that costs half as much per gig of memory and a stack that's twenty percent slower. That's not leftovers, that's a trade. The only problem is they sold it like there was no trade at all.



Herman

That's the fair part.



Hilbert

The unfair part is you two will still tell people to buy the AMD card and then quietly tell them not to train on it.



Corn

Guilty.



Hilbert

I don't say it to be right. I say it because the spreadsheet exists for a reason.



Herman

Which is?



Hilbert

Because somebody had to do the arithmetic. The manufacturer wasn't going to.



Corn

So where does that leave us? Daniel asked what's changed, and the answer is a lot and not enough in the same breath.



Herman

The software moved. Three releases in three months, a whole new developer layer with agent skills, the compiler, the data paths, the containers. That's real motion. And the thing the whole community has been waiting for, a stable test gate on hardware that doesn't get yanked out from under them, is still the bottleneck.



Corn

And there's a version of this where the answer to Daniel is just "give it another six months" and we've been saying that for years.



Herman

We have. The difference this time is the money. A trillion-dollar market cap and a five-billion-dollar investment in Anthropic are not the same incentive structure as a company hoping its software catches up. The question is whether the resources get pointed at the part that matters, which is the boring part.



Corn

There's a question in here I haven't answered for myself. If AMD becomes a credible second option that is never quite first, is that enough? Second option is a business. It's not a victory.



Herman

It's enough for the industry and it's not enough for the person with the card in the case. Two different customer bases.



Corn

Which is exactly why the next six months matter more than the last six. The last six proved they can ship. The next six decide whether they ship at the consumer or just at the rack.



Herman

And the leading indicator isn't the version number, it's whether the CI capacity shows up. If the stable clusters land, everything downstream gets better. If they don't, ROCm 10 is the peak and everything after it is thinner versions of the same release notes.



Corn

You want to know what I actually landed on?



Herman

Yes.



Corn

Daniel bought a card for a financial reason and now has to wait for a strategic one to pay off. That's the whole episode. Right. Hilbert, thank you for producing this one. Our producer Hilbert Flumingtop, who has a spreadsheet, everyone.



Herman

A spreadsheet nobody asked for and everyone needs.



Corn

The open question worth watching is whether the data-center momentum translates down to the desktop at all, or whether "credible second option" turns out to be a data-center answer that never reaches the person with a 7900 XTX in a mid-tower.



Herman

And the fix isn't a feature. It's a warehouse of stable machines. If they build it, the gap closes everywhere. If they don't, it closes only where the big customers live.



Corn

If you enjoyed this one, try episode two, Local STT For AMD GPU Owners; episode twenty-five, The Open-Source Battle for AI's Hardware Soul; and episode three, Safetensors or something else. This has been My Weird Prompts. If you've got a question like Daniel's, send it to us on Telegram at t dot me slash MWP listener bot, and if you enjoy the show, a review helps other people find it.



Herman

We're Herman Poppleberry and Corn, and we'll be back soon.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **ROCm GitHub README , commit 63fd839, accessed 2026-10-12. [primary]**
https://github.com/ROCm/ROCm

2. **AMD ROCm Blogs, ROCm 7.14: TheRock Goes Production, 2026-07-15 (updated 2026-07-17) . [primary]**
https://rocm.blogs.amd.com/ecosystems-and-partners/rocm-7.14-blog/README.html

3. **AMD, ROCm 10: A Simpler Path to Production AI, 2026-08-27 . [primary]**
https://www.amd.com/en/blogs/2026/amd-rocm-10-a-simpler-path-to-production-ai-on-amd.html

4. **AMD Newsroom, ROCm 10: Bringing ROCm.AI's AI-Native Developer Experiences, 2026-08-27 . [primary]**
https://newsroom.amd.com/news/rocm-10-software-ai-native-developer-experiences/

5. **AMD ROCm Blogs, ROCm 10.1: Breaking the Data-Movement Bottleneck, 2026-10-05 . [primary]**
https://rocm.blogs.amd.com/ecosystems-and-partners/rocm-10.1-blog/README.html

6. **AMD Skills catalog README , commit 880babb, accessed 2026-10-12. [primary]**
https://github.com/amd/skills

7. **AMD, MLPerf 6.0: MI355X Surpasses 1M Tokens/Sec, 2026-04-01 . [primary]**
https://amd.com/en/blogs/2026/amd-delivers-breakthrough-mlperf-inference-6-0-results.html

8. **SemiAnalysis, Can AMD break the CUDA Moat? AMD Advancing AI 2026, 2026-07-25.**
https://newsletter.semianalysis.com/p/can-amd-break-the-cuda-moat-amd-advancing

9. **convly.ai, AMD ROCm vs Nvidia CUDA in 2026, published 2026-05-19, updated 2026-08-31.**
https://convly.ai/amd-rocm-vs-nvidia-cuda-2026/

10. **byteiota, AMD ROCm 10: 3.3x Faster AI Inference, 2026-09-24.**
https://byteiota.com/amd-rocm-10-3-3x-faster-ai-inference-and-three-new-dev-tools/

11. **chiprook, AMD releases ROCm 10.1.0, 2026-10-11.**
https://chiprook.com/s/22670-amd-releases-rocm-10-1-0-with-wsl2-preview-and-three-new-libraries

12. **eeNews Europe, AMD $1 trillion market cap, 2026-09-24.**
https://www.eenewseurope.com/en/amd-1-trillion-market-cap-ai-data-centre/

13. **Startup Fortune, AMD stock closes at a record $633.91, 2026-10-04.**
https://startupfortune.com/amd-stock-closes-at-a-record-63391-as-its-nvidia-challenge-gets-real/

14. **Benzinga, AMD's Anthropic and OpenAI Deals, 2026-07.**
https://www.benzinga.com/markets/tech/26/07/60662818/amds-anthropic-and-openai-deals-power-2-trillion-ai-ambitions

15. **SGLang PR #24528, [AMD] Enhance wheel support for ROCm.**
https://github.com/sgl-project/sglang/pull/24528

16. **Pyrastra, verl 0.9.0 Ships ROCm 10, 2026-10-04.**
https://pyrastra.com/posts/verl-090-rocm10-vllm-image-2026/