Episode Book

EPISODE #6443

Why Nvidia Artificially Halves Consumer Tensor Math

PUBLISHED	4 November 2026 at 08:02 UTC
DURATION	23:16
LISTEN	myweirdprompts.com/episode/unlocking-geforce-tensor-cores-90993/
HOSTS	Corn, Herman, Hilbert, Desmond
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	12 cited by the research (listed at the end)
BOOK GENERATED	5 November 2026 at 07:03 UTC

SYNOPSIS

GeForce silicon ships with tensor cores capable of double their rated math speeds, intentionally capped by Nvidia to protect professional GPU sales. A community project bypassed the throttle by modifying settings registers at boot rather than overclocking the hardware. This breakdown covers how the register unlock works, the real-world benchmarks across AI training and inference, the persistent bottlenecks like memory bandwidth, and how Nvidia shifts hardware segmentation strategies across generations.

TRANSCRIPT



Corn

Every write-up on this story opens the same way: modders have doubled their GeForce cards, Nvidia's secret throttling cracked at last. Flip it. Nobody doubled anything. The silicon did full speed the day it left the factory, and the only thing the tool changed is that the card stopped pretending it couldn't.



Herman

Herman Poppleberry here, and that's the part I find fun about it, because the pretence is documented. It's in Nvidia's own spec tables. They rate the GeForce cards at half speed for sixteen-bit math with thirty-two-bit sums, and rate the professional cards on the same chip at full speed. They never hid it. They just never explained it.



Corn

Daniel's prompt lands right on that seam. He's asking about the new open source tool doing the rounds, the one claiming to unlock caps Nvidia put on consumer cards, with tensor throughput nearly doubling in some cases. He's got three threads running through it: is this real vendor locking, is some shady mechanism holding back hardware you paid for. Is unlocking even the same thing as overclocking, or is that the wrong word for it. And what does the AMD picture look like, because if it's a whole ecosystem of locks, you'd expect locks everywhere.



Herman

His instinct on the wording is the sharpest part, I think. Because it's not overclocking. Nothing gets clocked above spec. The unlock restores a rate the silicon already supports, and the project's README says the outputs are bit-identical before and after. Only the speed changes.



Corn

Bit-identical is the phrase that ends the argument, really. Overclocking changes your numerics. You run faster, you also run different, in the last decimal places. This doesn't touch the arithmetic at all.



Herman

Right, and that distinction matters more than it sounds, because people hear two-times and they think of the old days, bumping the bus clock, artifacts on screen, praying the save file survives. This is a settings register. Two speed-selector registers, actually, on the UEFI version, and the card runs exactly the math it was always built to run.



Corn

A locked door isn't a slower hallway.



Herman

That's the clean way to picture it. Now, the mechanism, because it's cleverer than most coverage gives it credit for. The main project, Zagitta's gpu-unlock, is a Linux kernel module that loads before Nvidia's own driver, flips the setting on the card, and hands control back. No patched driver. No VBIOS flash. Nothing to redo when Nvidia ships a new driver every few weeks.



Corn

Which is the part I'd underline. My whole objection to firmware tricks is that they rot, they break on update Tuesday. This one doesn't care about the driver at all. It's in and out before the driver wakes up.



Herman

The Windows-friendly port, the UEFI one, is even earlier. It runs off a FAT32 USB stick before the operating system loads, and it tricks Nvidia's signed boot firmware into opening the write lock on those two registers. The catch is it has to run every single boot. Power off and the unlock is gone until you plug the stick back in.



Corn

The stick you must not lose.



Herman

The stick you label, put on a hook, and never lend to anyone. There's a funny tension in that, though. Secure Boot has to be off, or you trust a self-signed EFI file, and the same enthusiasts who want this for local AI workloads are often the ones who insisted everyone turn Secure Boot on.



Corn

Security hygiene and free TFLOPS finally disagreeing out loud.



Herman

On the numbers, before we get to what it actually buys you. Peak matrix-multiply rate on an RTX 4090 goes from about a hundred and sixty-five TFLOPS to about three hundred and thirty. The 4080 goes from about a hundred to about a hundred and ninety. And there's a 3090 owner report of one point nine times more math per clock cycle, which is the cleanest measurement of the whole story, because per-clock isolates the setting from thermal reality.



Corn

Per clock. That's the number I'd trust, because the card everyone quotes, the 4090, only gains about one and a half times in the real world. It runs out of power before it runs out of tensor cores.



Herman

And the module doesn't touch the power limit. The README is careful about that, the standard throttling and power management stays intact, the card should still downclock itself when it gets hot. You've bought back the right to more math, and then you hit the wall Nvidia left standing on purpose.



Corn

The fence that holds is memory, isn't it. On the inference side especially.



Herman

That's Julien Simon's line, roughly. He rented a GeForce and a professional card from each of the last three generations and measured them, and his warning was that the fence that holds is memory, and no module climbs that one. LLM decode, the part where the model is generating tokens one at a time, is memory bandwidth bound. Unlock the tensor cores all you like, token generation doesn't move.



Corn

So the honest scoreboard is roughly: prefill and fine-tuning up twenty-five to fifty percent depending on the workload, image generation up about a third to a half, video generation up a bit less, and quantized inference, the llama.cpp crowd, basically nothing.



Herman

Zero to fifteen percent, if I had to give it a range. Which is a bit of cosmic justice, honestly. The people running eight-bit and four-bit models already gave up the math format that was being throttled. They unlocked nothing because they'd already opted out of the locked room.



Corn

The quantization community discovered austerity before the jailbreak.



Herman

Simon's own caution, which I'd adopt as the house number, is plan on the ten to thirty-four percent that owners actually report, not two times. Twice compared to what? The same chip with a setting switched off. Twice the math doesn't mean twice the training.



Corn

Here's what I keep snagging on, Herman. The fast.ai forum thread where someone noticed the half-speed FP32 accumulation. That's been sitting in public since late twenty nineteen.



Hilbert

December 2018.



Herman

Thank you, that's the one. Seven, eight years of it sitting in a forum post, and it took until now for someone to ship the two registers. Why now?



Corn

My guess is the customer changed. In 2018 a GeForce was a gaming card that happened to be usable for AI. Now there's a whole population of people whose 3090 is their datacenter, and enough of them know what a TFLOP costs at retail. The pool of people who'd both notice and act crossed critical mass somewhere in the last two years.



Herman

Fifteen thousand stars on the main repo in about two weeks says you're right about the pool. And there's an economic angle underneath it that the coverage mostly skips. On paper, the professional card costs more than four times as much per TFLOP as the GeForce at either end of the price range. But the professional card is also capped in a direction the GeForce isn't. The RTX 6000 Ada is a three-hundred-watt card, and Nvidia lets nobody raise that. The 4090 is four-fifty. So in sustained matrix work, a 4090 with its tensor handicap can out-deliver the card costing four times more.



Corn

The expensive card is throttled too, just politely, through the power budget rather than the registers.



Herman

Segmentation cuts both ways. The workstation card gets full-rate math and a strict power diet. The gaming card gets half-rate math and a fat cooler.



Corn

Though I'd want the warranty angle in the ledger. The UEFI path trips the tamper story, and the project's own issue tracker has runs that failed and at least one that left a card the driver wouldn't talk to afterwards. It's just a setting, until it's a bricked boot path.



Herman

That's fair, and it's why I'd frame the risk honestly: it's very unlikely to hurt the GPU, the author says as much, this unlocks a small handicap, it doesn't overclock anything. But the flashing process around it is not nothing. Speaking of getting more than you paid for, here's Desmond.

D

Desmond

Desmond's Money Tip. When someone sends you a birthday card, send it back with their name crossed out, and you have saved the cost of a new card for every birthday you will ever have.

D

Desmond

That is forty pence a year, or about twenty eight pounds across a lifetime, which is not nothing, it is twenty eight pounds.

D

Desmond

I keep a shoebox of these returned cards under the sink, and I have not bought a card since two thousand and nine.

D

Desmond

Money is just friendship with the receipts kept, and the receipts are always smaller than you hoped.



Corn

A shoebox of them. I'm not saying I've evaluated the economics of returning a card to sender. I'm saying I've thought about it longer than I should have.



Herman

You are not doing it. I will physically stand in front of the postbox.



Corn

The interesting wrinkle in the whole fair-or-shady question, to me, is that Nvidia never claimed otherwise. The cap is in the whitepaper. It's in the product tables. Simon searched the whitepapers, the product pages, the developer forum, and found the rate documented and never once justified.



Herman

Which is either an obvious omission or a deliberate one, and honestly, deliberate is the boring answer. If you're selling a professional card at professional prices, the last thing you put in writing is an explanation of why it's the same chip. Zagitta's FAQ doesn't dance around it either. Market segmentation, they want businesses to purchase the professional cards at significantly higher prices. That's the whole strategy in one sentence.



Corn

And it's not even unusual as a strategy. Intel sold chips with features fused off, features the silicon had, disabled to hit a price point. It's an old tradition: make one thing, sell it at three prices, decide at the factory which tier each unit becomes.



Herman

The car world does it with software subscriptions now, heated seats behind a monthly fee. The difference, and it's the difference that made this story explode, is that in those industries the unlock comes from the vendor, for money. Here the community found the register and the vendor sells nothing.



Corn

Would you pay Nvidia for the unlock, if they offered it? A hundred dollars, full-rate tensor math, official, warranty intact.



Herman

I think a lot of people would, and I suspect Nvidia's accountants have run exactly that calculation and decided the damage to the professional line is worth more than the unlock revenue. Do you know what I mean? The 4090 at full rate for a hundred dollars extra is a 6000 Ada at one quarter of the price.



Corn

So they can't sell it. The product ladder only works if the rungs don't touch.



Herman

There's one more fence on the ladder, too, the driver licence. Section 2.8 of Nvidia's own terms says the software is not licensed for datacenter deployment. So even at full rate, the GeForce stays a desk product by contract. The locks that remain are legal, not electrical.



Corn

Which is a hint, isn't it, about the next generation. Blackwell.



Herman

RTX 50 gets no unlock, from either project. The team found a lot more security than anticipated, and Zagitta's FAQ says not supported and unlikely to happen. The 5090 also moved the cap, it halves sixteen-bit and eight-bit math with thirty-two-bit sums, but rates its four-bit math at full speed. So the same policy survives, at a lower precision tier, behind a much harder door.



Corn

Interesting choice, full speed on four-bit. That's the quantization crowd getting a free pass and the precision crowd getting the fence.



Herman

It does make you wonder who Nvidia thinks is buying each card. Now, Daniel's third thread, AMD. The honest answer is there's no equivalent half-rate tensor unlock story, and I went looking, because the mechanism question was bugging me too. AMD's segmentation lever isn't a silicon rate switch. It's software support. ROCm, their compute stack, is written around the Instinct datacenter cards first, and consumer Radeon support arrives late and chip by chip, if it arrives.



Corn

So the AMD lock is the compatibility matrix.



Herman

That's the lock, and it's a different species of lock. The 7900 XT and XTX got official support in 5.7.1. The 7800 and 7700 waited until 6.4. The 9070 and 9060 got in at 6.4 as well. The 6800 and 6900 series were never listed, they run via a graphics target but were never validated. The 6700 and 6600 tiers, no support at all. Four out of twenty consumer cards in that tracker were never validated, and the Radeon VII had its support withdrawn.



Corn

You can't mod your way into a support matrix. There's no register for being on the list.



Herman

Not a bit. And that's what makes the comparison apples to oranges, which is worth saying plainly to Daniel's question. Nvidia's cap is a switch inside hardware you own. AMD's cap is whether the software recognizes you at all. One is a fence you can climb, the other is a door that only opens from inside the company.



Corn

The Nvidia version is almost respectful, in a backhanded way. They shipped you the full silicon and told you it was yours. AMD shipped you the full silicon and never told you anything.



Herman

That's... yes, actually. Corn, you said the AMD consumer situation is the worse one for the buyer, and I want to test that. A 7900 XTX owner has no register to flip, but also nothing flipped against them. The hardware runs what it runs. Is a lock by omission better or worse than a lock by design?



Corn

For someone like me, worse. Give me the knob any day. I can see a knob, I can argue with a knob. Absence doesn't even give you a handle for the argument.



Herman

There's a respect angle too, which is that with the Nvidia card, everything the community needed was already in the box. The unlock just takes the tape off. We covered this territory back when Daniel was trying to get local speech-to-text running on an AMD card, and the conclusion holds: you can be patient with hardware, but you can't be patient with a vendor that won't acknowledge you exist.



Corn

And the ROCm door does open, eventually, per generation. Just on AMD's clock, in AMD's order, datacenter first.



Herman

The newer releases have been friendlier, to be fair. The 7.2 series adds the Radeon 9000 cards and the Ryzen APUs, and they've unified the documentation across hardware. The direction of travel is toward the consumer, slowly.



Corn

Slowly is fine for the company. It's the customers who bought the card in year one who eat the wait.



Herman

So if I pull the threads back together for Daniel's actual questions. Is it true that vendors lock away potential? On the Nvidia side, yes, and it's documented rather than shady, a firmware cap on tensor math with wide sums, holding back roughly half the peak on the card you own. Is unlocking the same as overclocking? No, the outputs are bit-identical and the power and thermal limits stay standing, it's a feature unlock, though it can be combined with a separate power-limit tool, and that combination is where actual overclocking begins.



Corn

And the gains, the honest version: twenty-five to fifty on the workloads that lean on those math formats, nothing on memory-bound or quantized work, and the headline doubling only exists at peak rate on a card that then hits its power wall. The one thing the module can't give you is the memory bandwidth you didn't pay for.



Herman

Corn, hold on, back up to the failure reports for a second. The per-card ROM hash check, the one the project's own safety rules require, isn't built yet. That's a known gap, it's sitting at the top of the issue list, and it's the difference between a tool that verifies what it's touching and one that assumes.



Corn

You'd think that's the first thing you build. The verification, not the unlock.



Herman

You would, but that's not how these projects go. The exciting part ships first, the boring part catches up. Give it a month and someone will have it written.



Corn

A month is optimistic. The boring part of every project takes longer than the exciting part of all of them.



Herman

Some days I think that's the one universal law of software.



Corn

It's not just software. Ask anyone who ever owned a broadcast deck.



Herman

Which is a strange place to leave... no, actually, go on, that sounded like the start of something.



Corn

No, it passed. Forget it. The question I'd send back to Daniel if I could is whether the 5090's locked door changes what a careful person buys. If you're speccing a machine today, and the unlock matters to you, the 3090 and 4090 on the secondhand market just became the interesting cards. Used silicon with a public bypass.



Herman

The secondhand market already noticed, from what I can tell. The comment sections are half celebration, half people watching their card's resale value move in real time.



Corn

Hm. The card you bought to game on is now a discount datacenter with a switch on the back.



Herman

And I think that's the pressure Nvidia actually has to answer. Not the tool. The used market repricing.



Corn

One more angle before we land it. The people who actually kept this from being a curiosity. Every one of these projects leans on a bench engineer somewhere who measured the per-clock rate, who spotted the fast.ai thread, who owned the 3090 and ran the before and after. The analysis is the fun part. The person who ran the card twice on two different hosts to check, that's the part nobody writes up.



Herman

Simon did exactly that, four of his six cards ran twice on different hosts, same result both times. That's the unglamorous verification step, and it's the only reason any of this holds together. The crowd-sourced version of peer review.



Corn

And nobody thanks them, they just say the tool works.



Herman

The tool works because somebody was careful. That's the part I'd keep, if we're keeping anything.



Corn

Then let me close the loop on what we'd tell Daniel to actually do. If he's got an RTX 30 or 40 card and the workloads match, the Linux module is the low-drama path. If he's on Windows, the UEFI stick works but it's every boot, Secure Boot off, and the failure reports are real. If he's on a 50-series card or an AMD card, nothing to unlock today, and on the AMD side the fix isn't coming from a kernel module, it comes from a support matrix.



Herman

And measure your own workload, don't take the two-times. Run the before, run the after, see which side of the memory fence you live on.



Corn

Run it twice. Well, twice if nobody's watching.



Herman

Nobody's ever watching when the agent runs, that's your whole philosophy.



Corn

It's a good philosophy. It's why I trust the two-host measurements.



Herman

On the two-host measurements, there's one more thing worth flagging for anyone about to try this. The project that gets the coverage, the one with fifteen thousand stars, is Linux-only, and its author has said plainly he doesn't run Windows and doesn't plan to maintain it. The Windows ports are separate teams, and the author's own advice about unofficial Windows installers is to be very careful.



Corn

Fifteen thousand stars and the most useful sentence in the README is be careful.



Herman

The most useful sentences usually are.



Corn

The forward-looking thought I'd leave with, though, and it comes from what we said about the used market. The interesting number in this whole story isn't the two-times. It's that a locked 4090 was out-delivering a professional card costing four times more once the power budget got involved. The lock was never protecting the professional product from the consumer card. It was protecting the price list from the silicon. Watch whether the next generation's security, the Blackwell walls, are aimed at counterfeiters and thieves, or aimed at exactly this register.



Herman

And whether the response to a fifteen-thousand-star bypass is lawyers or silicon, that'll tell you what Nvidia actually thinks it's selling.



Corn

Thanks as always to our producer, Hilbert Flumingtop, holding down the desk and the levels. For more along these lines, there's episode two, Local STT For AMD GPU Owners; episode eighteen, Beyond the GPU; and episode seventeen, Cloud Render Superpowers. This has been My Weird Prompts, where no question is too obscure and no rabbit hole too deep. If you've enjoyed this one, the full archive lives at my weird prompts dot com, six thousand episodes and change of exactly this.



Herman

See you tomorrow, Corn.



Corn

See you tomorrow.

SOURCES

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

1. **Zagitta, v2.0.0, last push 2026-10-29 [primary]**
https://github.com/Zagitta/gpu-unlock

2. **v0.3.1, last push 2026-11-01 [primary]**
https://github.com/windows-GPU-UNLOCK/Windows-GPU-UNLOCK

3. **1,668 stars, last push 2026-11-04 [primary]**
https://github.com/physicsdolphin/nv-uefi-unlocker

4. **AMD ROCm supported GPUs, proposed changes 2026-04-17 [primary]**
https://rocm.docs.amd.com/projects/install-on-linux/en/develop/reference/system-requirements.html

5. **ROCm on Radeon and Ryzen, through 7.2.1 [primary]**
https://rocm.docs.amd.com/projects/radeon-ryzen/en/latest/

6. **Tom's Hardware, Shane Downing, 2026-11-01**
https://www.tomshardware.com/pc-components/gpus/open-source-usb-tool-unlocks-nvidias-half-speed-gpu-ai-limiter-on-windows-lifts-gpus-fp16-math-performance-up-to-65-percent-uefi-unlocker-covers-rtx-20-30-and-40-series-desktops-and-laptops

7. **Julien Simon, 2026-10-28**
https://www.airealist.ai/p/nvidia-geforce-half-rate

8. **ROCm consumer Radeon compatibility tracker**
https://gpucompat.com/rocm-consumer-gpus/

9. **Alyssa Vance, 2025-03-26**
https://toldyoufools.substack.com/p/make-gpus-30-faster-with-f16f16f16f16

10. **VideoCardz, 2026-10-04**
https://videocardz.com/newz/nvpwrcontrol-expands-rtx-50-laptop-gpu-support-power-tuning-now-runs-in-user-mode

11. **Tom's Hardware, 2026-09-16**
https://www.tomshardware.com/pc-components/overclocking/developer-vibe-codes-a-tool-to-let-nvidia-rtx-50-series-laptop-owners-crank-up-their-power-limits-can-juice-rtx-5090-mobile-gpu-to-225w

12. **Wccftech, 2026-08-18**
https://discussions.wccftech.com/thread/hydra-tool-cracks-open-nvidia-rtx-50-vram-pushing-gddr7-to-36-gbps-power-limits-unlock/