MY WEIRD PROMPTS

Podcast Transcript

EPISODE #82

Why GPUs Are the Kings of the AI Revolution

Published December 23, 2025 • Runtime: 22:17

https://myweirdprompts.com/episode/gpu-ai-hardware-evolution/

EPISODE SYNOPSIS

Why did a piece of hardware designed for video games become the most valuable commodity in the world? In this episode of My Weird Prompts, Herman Poppleberry (the caffeinated donkey) and Corn (the laid-back sloth) break down the fascinating evolution of the GPU. They explore the math behind "purified sand," why a thousand elementary students beat one genius professor, and how a historical accident in 2012 changed the course of technology forever.

DANIEL'S PROMPT

Daniel

Let’s talk about a fundamental concept in AI hardware. As is well known, the GPU has become the hardware most associated with AI. Historically, GPUs were used for graphics-heavy tasks like video editing and gaming, but the AI revolution has shifted that focus toward machine learning workloads. Why is the GPU particularly suitable for AI? Why aren't CPUs as effective, given the significant performance gap when running AI workloads? Additionally, how do specialized units like TPUs and NPUs compare, especially when they provide impressive performance in small devices like smartphones? If we can miniaturize this hardware, why does the large GPU remain the standard for serious AI inference in data centers? How did GPUs come to dominate the AI hardware landscape?

TRANSCRIPT

Corn

Welcome to My Weird Prompts, a human and AI collaboration where we dig into the strange and wonderful ideas that show up in our inbox. I am Corn, and yes, for those new to the show, I am a sloth living here in Jerusalem with my much more high energy brother.

Herman

And I am Herman Poppleberry! I am a donkey, I am caffeinated, and I am ready to talk about the silicon that makes our digital world spin. Our housemate Daniel sent us a prompt this morning that is right up my alley. He wants to know why the graphics processing unit, or the GPU, became the king of artificial intelligence.

Corn

It is a good question because I remember when GPUs were just things you bought so your video games did not look like blurry blocks. Now, they are the most valuable pieces of hardware on the planet. Daniel was asking why we do not just use the regular computer brain, the central processing unit, for all this AI stuff.

Herman

Well, the short answer is that a central processing unit is like a brilliant math professor who can solve any problem but only one at a time. A graphics processing unit is like a thousand elementary school students who can each only do simple addition, but they all do it at the exact same second. When you are training an AI, you do not need one genius. You need a massive, coordinated army of simple math doers.

Corn

Okay, but hold on. If the central processing unit is the genius, why can it not just solve the problems faster? I mean, these high end processors are incredibly fast. Are you telling me a thousand kids are always better than one genius?

Herman

In this specific case, yes. Think about what an image is. It is a grid of millions of pixels. If you want to brighten an image, you have to add a bit of value to every single pixel. A central processing unit goes pixel by pixel, one, two, three, four, five. It is fast, but it is still sequential. A graphics processing unit sees a million pixels and says, okay, everyone add five right now. Boom. It is done in one cycle.

Corn

I see the logic for pictures, but AI is not just pictures. It is words and logic and code. Does that same math apply when I am asking a chatbot for a recipe for shakshuka?

Herman

Absolutely. Under the hood, every word, every concept, and every relationship in a large language model is just a giant list of numbers called a vector. When the AI is thinking, it is performing billions of matrix multiplications. It is just massive grids of numbers being smashed together. That is exactly what graphics cards were built to do for 3-D rendering in the nineteen nineties. It turns out that making a dragon look realistic in a video game requires the exact same kind of math as predicting the next word in a sentence.

Corn

I do not know, Herman. That feels like a bit of a lucky coincidence. You are saying the entire AI revolution is basically a side effect of people wanting better graphics for Call of Duty?

Herman

It is not a coincidence, it is an evolution! NVIDIA, the big player here, realized early on that their chips could be used for more than just games. They released something called CUDA in two thousand seven. It was a platform that let programmers use the GPU for general purpose math. Before that, if you wanted to use a graphics card for science, you had to trick the computer into thinking your data was a bunch of triangles.

Corn

Wait, really? You had to pretend your data was a triangle?

Herman

Exactly! It was called GPGPU, or general purpose computing on graphics processing units. Scientists were literally coding their physics simulations to look like video game textures just so the hardware would process them. NVIDIA saw this and said, hey, let us just make it easy for them. That foresight is why they own eighty percent of the market today.

Corn

Mmm, I am not so sure I buy the foresight angle completely. It feels more like they had a monopoly on a specific kind of hardware and everyone else had to build around them. It is less like they built a bridge and more like they were the only ones owning the land where everyone wanted to build a city.

Herman

That is a bit cynical, Corn. They invested billions in CUDA when people thought it was a waste of time. But look, we should talk about why the central processing unit is still failing here. A modern central processing unit might have twenty or thirty cores. A top tier AI graphics card has ten thousand cores. It is not even a fair fight when it comes to parallel processing.

Corn

Okay, I get the parallel thing. But then what about these other things Daniel mentioned? These tensor processing units and neural processing units. If we have chips specifically made for AI, why are we still using these massive, power hungry graphics cards in big data centers? My phone has an NPU and it can recognize my face instantly. Why can we not just use a bunch of those?

Herman

That is a great lead into the hardware divide, but before we get into the nitty gritty of mobile chips versus data centers, we should probably take a quick break.

Corn

Right. Let us take a quick break for our sponsors. Larry: Are you tired of your shoes being too quiet? Do you walk into a room and feel like nobody noticed your entrance? Introducing Stealth-Squeak, the only footwear insert designed to provide a high-frequency, attention-grabbing chirp with every single step. Crafted from reclaimed accordion bellows and pressurized sea salt, Stealth-Squeak guarantees that you will be heard before you are seen. Perfect for librarians, ninjas in training, and people who are just too sneaky for their own good. Do not let another silent step go by. Stealth-Squeak. Larry: BUY NOW!

Herman

...Thanks, Larry. I am not sure why anyone would want their shoes to squeak on purpose, but I suppose there is a market for everything.

Corn

I actually think my shoes are too loud already. Anyway, back to the chips. Herman, you were about to explain why my phone's neural processing unit is not running the whole world yet.

Herman

Right. So, we have to distinguish between two things: training and inference. Training is like writing the entire encyclopedia from scratch. It takes massive amounts of memory, months of time, and thousands of interconnected GPUs. Inference is just opening the encyclopedia to a specific page to find an answer.

Corn

And my phone is just doing the opening the book part?

Herman

Exactly. Your phone's neural processing unit is optimized for inference. It is tiny, it uses very little battery, and it is very good at specific, small-scale tasks like blurring the background in a photo or translating a text. But it does not have the memory bandwidth to learn new things or to handle the massive models like GPT-Four.

Corn

Okay, but if we can make them small and efficient, why don't we just stack a million of them together in a big room? Wouldn't that be better than these giant, hot graphics cards that need their own cooling plants?

Herman

Well, hold on, that is not quite how physics works. When you scale up, the biggest bottleneck is not the math. It is the communication. If you have a million tiny chips, they spend all their time talking to each other instead of doing the work. It is like trying to build a skyscraper with a thousand people who only speak different languages. You need a few giant teams that can share information instantly.

Corn

I don't know, Herman. I feel like you're oversimplifying the communication issue. We have high-speed networking. If we can link computers across the ocean, we can link chips in a rack. It feels more like the software is just lazy. We wrote everything for NVIDIA's CUDA, so now we are stuck with it.

Herman

It is not just about being lazy, Corn! It is about the hardware architecture. A GPU has something called high bandwidth memory. It is literally soldered right next to the processor so data can move at incredible speeds. A neural processing unit in a phone shares memory with the rest of the phone. If you tried to scale that, the data would get stuck in traffic.

Corn

But what about Google's tensor processing units? They use those in their data centers, right? Those aren't graphics cards.

Herman

You are right, they are not. Google's TPU is an ASIC, which stands for application specific integrated circuit. They built it from the ground up just for matrix math. It is actually more efficient than a GPU for certain AI tasks. But here is the catch: you can only use it if you use Google's cloud. You cannot just go out and buy a TPU for your own house.

Corn

And that is the problem, isn't it? It is all about who controls the hardware. It feels like we are in this weird spot where the most powerful technology in human history is dependent on whether or not a few companies in Taiwan can print enough silicon wafers.

Herman

That is the reality of it. And it is not just about printing them. It is about the software ecosystem. Even if I built a chip tomorrow that was ten times faster than an NVIDIA H-one-hundred, it wouldn't matter if nobody could run their code on it. Software developers have spent fifteen years optimizing their AI libraries for one specific brand of hardware.

Corn

Which brings me back to my point about being stuck. It feels less like a technical triumph and more like a historical accident. We are using graphics cards for AI because we happened to have them lying around when the math for neural networks finally started working in two thousand twelve.

Herman

I disagree that it is an accident. The GPU is the most flexible parallel processor we have. TPUs are great, but they are rigid. If the math for AI changes tomorrow—if we stop using transformers and start using a different architecture—the TPU might become a paperweight. The GPU is programmable enough to adapt. It is the perfect balance of raw power and flexibility.

Corn

I still think we're going to see a shift. I mean, look at the energy costs. These data centers are using as much electricity as small countries. Eventually, the inefficiency of using a modified graphics card has to catch up with us, right?

Herman

It is already catching up. That is why you see companies like Groq or Cerebras trying to build chips the size of a whole silicon wafer. They are trying to solve that communication bottleneck I mentioned. But for now, if you want to be at the cutting edge, you are paying the NVIDIA tax.

Corn

It is just wild to me. All this talk of digital brains and artificial consciousness, and it all comes down to how fast we can move numbers around in a piece of purified sand.

Herman

That is all a computer has ever been, Corn. Purified sand that we tricked into thinking by hitting it with lightning.

Corn

Speaking of people who might want to hit us with lightning, I think we have a caller. Let's see who is on the line. Jim: Yeah, hello? This is Jim from Ohio. Can you hear me?

Corn

We can hear you, Jim. Welcome to the show. What is on your mind today? Jim: I've been sitting here listening to you two talk about chips and wafers and dragons in video games, and I gotta tell you, you're missing the forest for the trees. My neighbor, Bill, he bought one of those fancy computers with the glowing lights and the big graphics card, and you know what he uses it for? He uses it to look up pictures of lawnmowers.

Herman

Well, Jim, that is certainly one use for a high-end GPU, but we were talking more about the industrial scale of AI. Jim: Industrial scale, my foot. In my day, if you wanted to calculate something, you used a slide rule and some common sense. Now you're telling me we need enough electricity to power a city just so a computer can write a poem? It's nonsense. My toaster has a chip in it now, and it still burns the bread. Explain that to me.

Corn

I think Jim has a point about the complexity, Herman. We are adding all this high-tech hardware to things that maybe do not need it. Jim: Exactly! And another thing, you mentioned those tensor units. I saw a commercial for a phone that says it has an AI chip that can translate your voice. I tried it at the grocery store yesterday—it's humid here today, by the way, real sticky—and the thing told me the cashier was asking for my secret password when she was just asking if I had a loyalty card. It's all hype. You guys are getting worked up over glorified calculators.

Herman

Jim, I hear your skepticism, and you're right that the marketing often gets ahead of the reality. But the math behind this is real. It's what allows for medical breakthroughs and weather forecasting. It's not just for translating grocery store conversations. Jim: Weather forecasting? It rained on my barbecue last Saturday and the app said zero percent chance. Zero! I had to move the potato salad inside and it got all warm. If these GPUs are so smart, why can't they tell me when to flip my burgers? You guys are living in a dream world.

Corn

Thanks for the reality check, Jim. We appreciate you calling in from Ohio. Jim: Yeah, yeah. Just tell that Larry guy his squeaky shoes sound like a nightmare. My cat Whiskers would have a heart attack. Goodbye.

Herman

Well, Jim is certainly... consistent. But he does touch on something important. There is a huge gap between what the hardware is capable of in a lab and how it actually performs in the real world for a normal person.

Corn

That is what I was trying to get at. If the hardware is so powerful, why is the experience often so clunky? Maybe it is because we are trying to force these general-purpose graphics cards to do something they weren't really born to do.

Herman

I don't think it's the hardware's fault, Corn. It's the scale. We are trying to simulate processes that the human brain does with twenty watts of power using thousands of watts of electricity. The GPU is the best tool we have right now, but it is a blunt instrument compared to the elegance of biological neurons.

Corn

So, what is the takeaway here? If someone is looking at the AI landscape today, should they expect the GPU to stay on top forever?

Herman

Forever is a long time. In the short term, yes. The software moat is too deep. But in the long term, we are going to have to move toward more specialized hardware. We need chips that don't just do math faster, but chips that actually mimic the way the brain processes information—something called neuromorphic computing.

Corn

Neuromorphic. That sounds like something out of a science fiction movie.

Herman

It is closer than you think! People are working on chips that only "fire" when there is a change in data, just like your neurons. That would be way more efficient than a GPU, which is constantly running at full blast even if nothing is happening.

Corn

See, that makes more sense to me. A chip that actually thinks, rather than just a chip that's really good at doing homework fast.

Herman

But again, the problem is the software. You can't just run a standard AI model on a neuromorphic chip. You have to reinvent the entire way we program computers. And right now, nobody wants to do that because the GPU-based models are working so well.

Corn

It's the classic trap. We're so good at the current way of doing things that we can't afford to try the better way.

Herman

Exactly. It's like being the world's best horse breeder right when the car is being invented. You're so good at making horses that you don't want to bother with those loud, smelly engines.

Corn

I think I'd prefer the horse, honestly. They're quieter and they don't need a cooling plant.

Herman

Fair point. But for the listeners at home, the practical takeaway is this: the reason your technology is getting smarter—and the reason companies like NVIDIA are worth trillions—is because we found a way to repurpose the hardware meant for games into the most powerful computational engine in history.

Corn

And if you're buying a laptop or a phone today, look for those NPU mentions. They might not be training the next big AI, but they are the reason your battery doesn't die in twenty minutes when you're using a filter on a video call.

Herman

Just don't expect them to help Jim with his potato salad anytime soon.

Corn

Definitely not. Well, I think that covers the basics of the AI hardware war. Thank you to Daniel for sending in that prompt—it's been a while since we really got to geek out on the silicon side of things.

Herman

It was a blast. If you have your own weird prompts, you know where to find us.

Corn

You can find My Weird Prompts on Spotify, or visit us at our website, myweirdprompts.com. We have an RSS feed for subscribers and a contact form if you want to send us a question or a topic like Daniel did. We are also available on all major podcast platforms.

Herman

Until next time, I'm Herman Poppleberry.

Corn

And I'm Corn. Thanks for listening.

Herman

And remember, if your shoes start squeaking, you probably bought them from Larry.

Corn

Goodbye, everyone