SDS PODCAST

EPISODE 992: TOKENMAXXING VS AI HARDWARE BOTTLENECKS

Jon Krohn:	00:00	This is episode number 992 on Token Maxing and the AI Hardware Bottleneck. Welcome back to the SuperDataScience podcast. I'm your host, Jon Krohn. The hottest social media trend in AI today seems to be token maxing, wherein folks use agentic tools like Claude Code and OpenAI's Codex to maximize the number of tokens they consume. I don't know why someone would think this is an actual indicator of productivity, but apparently it stems this trend of token maxing stems from companies tracking AI usage via internal dashboards. I hope you and your company aren't engaged in this vanity metric nonsense. And relatedly, today's topic on the podcast is the AI Compute Crunch, how the breakneck rush to train and deploy ever larger language models with people using evermore tokens from them is now slamming into very real, very physical bottlenecks across the global supply chain. You might have noticed this personally over the past several months, if you're a heavy Claude code user, you probably bumped up against the weekly rate limits Anthropic introduced last August.
	01:14	Anthropic was unusually candid about why. They said that a small fraction of power users were running Claude twenty four seven in the background and eating into capacity for everyone else. The company has been blunt that they are, in their own words, very compute constrained. OpenAI also famously paused new Source signups to redirect scarce GPUs. GitHub at one point stopped accepting new subscribers for its Copilot bot and on the developer side on- demand H100 access, a popular Nvidia GPU through AWS Azure or Google Cloud has become genuinely unreliable for any team without pre-reserved capacity. So what's actually going on? In short, demand for AI compute is rising substantially faster than the underlying hardware supply chain can scale. And the bottleneck is no longer any one thing. There are at least four overlapping strategies that have to

be solved simultaneously and solving any one of them doesn't actually unstick the system.

02:11 Bottleneck number one is the most famous GPUs themselves. Lead times on NVIDIA's data center GPUs are now running 36 to 52 weeks and Blackwell allocation, the next generation after those popular hopper style H100s, Blackwell allocation is largely sold out through mid 2026 with reported backlogs in the millions of units. Even those older NVIDIA H100 chips originally launched back in 2022 have actually gotten more expensive to rent on the spot market somewhere around 30% more expensive since November because customers can't get hold of newer hardware and are falling back on previous generations. But here's the wrinkle. GPU fabrication itself isn't really the binding constraint anymore. The real choke point sits one layer up at TSMC, the world's largest dedicated independent semiconductor foundry who have a monopoly on state-of-the-art chips. So yeah, the real choke point is with them with TSMC and something called CoWas, Chip on Wafer on Substrate, the advanced packaging step that bonds GPU dies to their high bandwidth memory stack.

03:19 Whether that phrase means anything concrete to you or not, the main takeaway on this is that TSMC's CEO CC Way has publicly stated that Co-op's capacity chip on wafer on substrate is sold out through 2026 and NVIDIA alone has reportedly locked up roughly 60% of TSMC's total co-ops allocation through 2027. TSMC is racing to nearly quadruple monthly co-ops output by late 2026, but as that TSMC CEO CC Way dryly puts it, there are no shortcuts. Building a new fab, a new fabricator that creates these chips takes two to three years. All right, that's bottleneck number one, the GPUs themselves. Bottleneck number two is the memory side of that same equation, high bandwidth memory or HBM. So I mentioned memory in bottleneck number one and it's the memory itself that forms another bottleneck. So every

Nvidia H100 needs 80 gigabytes of HBM3, so high bandwidth memory third generation and every Blackwell 200 needs 192 gigabytes, so three times as much of another kind of high bandwidth memory called HBM3E, which is fifth generation and 25% faster than the HBM3 third generation high bandwidth memory.

04:41 So total HBM demand, high bandwidth memory demand has roughly quintupled since 2023 and there are only three companies on the planet that make this stuff, SK Hyneks, Samsung and Micron. All three say their HBM supply is largely sold out well into 2026 and Nvidia has reportedly cut consumer RTX 50 graphics card production by 30% to 40% in the first half of this year because the same memory fabs feeding HBM lines are also responsible for memory on consumer devices. For example, the kinds of consumer devices that are used to render video games on people's personal computers. New HBM fabs, similar to the way that with bottleneck number one, I was talking about how fabs take two to three years to make for GPUs. With HBM Fabs, it's kind of the same story. It takes 18 to 24 months to have a new HBM fabricator come online and demand is expected to outpace supply for at least three more years.

05:41 All right, bottleneck number one was GPUs. Number two was memory. Bottleneck number three is one that I find particularly interesting because it's caught a lot of folks off guard and that's CPUs. As workloads shift from training to inference, especially as agentic AI proliferates, that is AI systems that plan multi-step tasks call tools and coordinate work across many GPUs. The CPU to GPU ratio in the data center is climbing dramatically. Analysis from Morgan Stanley suggests that chatbot style systems like the ChatGPT interface that we were first introduced to a few years ago, that needs roughly one CPU for every 12 GPUs, whereas Agentix systems require a one-to-one ratio. Whoa. So 12 times as many CPUs per GPU relative to what we were expecting in the GenAI era up until now

that the GenAI era has morphed into the Agentic AI era. And so this is good news for CPU manufacturers like Intel.

06:44 Intel said on its Q1 earnings call this year that its server CPU shortfall in CFO David Zinsner's words starts with a B meaning billions of dollars of unmet demand for CPUs. Server CPU prices therefore have jumped 10 to 20% in just the past couple of months. That incidentally is a big part of why Intel's market cap has more than doubled over the past six months. All right we're onto the fourth and final bottleneck and this is arguably the most intractable. That's electricity. So through 2024, the binding constraint on building a new AI data center was capital and chips. But now every major hyperscaler has reported that their buildouts are gated not by money or hardware but by grid interconnect timelines, which can run 18 to 36 months and transformer lead times, which is another 18 to 24 months on top of that. Gartner now projects that power shortages will restrict 40% of AI data centers by 2027 and local communities are pushing back hard.

07:54 As of March, at least 12 US states had filed data center moratorium bills. Maine's legislature, for example, actually passed a statewide moratorium in April. The governor vetoed it, but more than 50 local moratoriums. Moratoria, I'm not sure of the plural, but more than 50 local moratoria have already passed across the US. Communities near hyperscale clusters in Virginia, Texas and Georgia are already seeing electricity rate increases of 8% to 15% and that along with concerns about water consumption and land use is fueling the backlash. To put the scale of the demand wave in perspective, the five biggest hyperscalers, Alphabet, Amazon, Meta, Microsoft, and Oracle, they're on track to spend something on the order of $725 billion combined on capital expenditure in 2026, roughly three quarters of which is targeting AI

infrastructure. That's up from about $120 billion total in 2022. That's a sixfold increase in four years.

08:56 Anthropic alone has in just the past few weeks committed over $100 billion to AWS for up to five gigawatts of GPU capacity or AI compute capacity, locked in roughly another five gigawatts from Google and signed a deal to take all of the SpaceX Colossus One site for more than 300 megawatts and over 220,000 NVIDIA GPUs, which you'll notice is exactly the capacity injection that allowed Anthropic to recently double ClaudeCode's rate limits. And yet perhaps the most remarkable pattern across the data I've been digging through is the asymmetry between hyperscaler capital expenditure and the hardware suppliers who actually have to build the underlying gear. While the hyperscalers have tripled their combined CapEx over the past two years, the chip makers, equipment vendors, networking gear suppliers and cooling specialists have only increased theirs by about half. Those hardware suppliers are wary of overbuilding capacity that could sit idle if the AI boom slows.

09:59 And given the timelines involved, even an aggressive ramp up now wouldn't deliver meaningful relief until 2028 or later. Elon Musk has announced plans for what he's called a Terrafab intended to produce more processing power per year than today's entire global semiconductor industry, but even it isn't expected to start production till 2028 at the earliest and at a fraction of the envision scale that year. Now, before you write off the whole AI buildout is doomed, there are real reasons for optimism as well. The key thing here is algorithmic efficiency, which continues to improve dramatically. Google's TurboQuant, for example, announced in March. I've got a link to that for you in the show notes, briefly tanked memory stock prices because it promised to materially reduce how much memory inference workloads need. Custom silicon is also proliferating. So Amazon's Trainum two cluster powering anthropics training is already operational with

over 500,000 chips and the hyperscalers are increasingly looking to compliment but apparently not replace NVIDIA.

10:59 And as I've discussed many times on the podcast, there's still enormous headroom for further LLM efficiency gains in both training and inference through techniques such as mixture of experts approaches, which you can hear more about back in episode number 778. The summary overall from me is this. The gap between AI software's voracious appetite and the messy multi-year physical reality of building chips, packaging them, manufacturing memory and wiring up substations is going to be a major story for at least the next two to three years. Improving an LLM takes weeks. Building a fab takes years. Building transmission lines can take a decade. That mismatch will shape who gets to scale, who gets throttled, and quite possibly which AI companies survive the next economic cycle. But for those of us building applications on top of these models, the picture is actually pretty exciting still. Compute scarcity is forcing the entire industry to get dramatically more efficient, better quantitization, smarter inference scheduling, smaller specialized models, custom silicon that's economical to run and scale.

12:04 Every one of those efficiency gains makes the AI applications you can dream up cheaper to deploy and easier to put into the hands of more people. Just don't bloody TokenMax as you do it. All right. And we have time for a Apple Podcast review here at the end of the episode. We have one here from someone named the Liz with a whole bunch of Zs or Zs, depending on what country you're from. The Liz W. Who says that this is the only AI podcast you need. Thanks, Liz W.

12:41 She, I assume, goes on to say that you know how we all have that one podcast that we go to when we need consistently valuable information but don't have a lot of time. For me, it's the SuperDataScience AI podcast with Jon Krohn. She goes on to say that hands down, my go-

to source for keeping a pulse on what matters. The content is always timely and speaks grounded truth for practitioners. I've learned more from the podcast than any other source available. Fantastic. It continues to go on, but I'll stop there. Thank you very much for that wonderful positive five-star review. Thanks to everyone for all the recent ratings and feedback on Apple Podcasts, Spotify, and all the other podcasting platforms out there, as well as for likes and comments on our YouTube videos. Really appreciate all that. I read all the feedback, which is helpful for adapting the show for the future.

13:33 And that also reminds me that we did ... I'm very grateful to all of you who filled in the brand survey that I posted on social media and mentioned once on the podcast as well. Nearly 600 people filled out that 40 questionnaire around our podcast brand. Yeah, we're digging into those data now and we'll have some kind of results to tell you about in the near future. All right. So yeah, thanks for continuing to support the show, interacting with us and letting other folks know about the podcast through your comments and that kind of thing. If you write written feedback on air on Apple Podcasts, I will read it on air just like I did today.