Episode Book

EPISODE #6678

How HBM and Slow Fabs Squeeze the RAM Supply

PUBLISHED	19 November 2026 at 12:42 UTC
DURATION	22:30
LISTEN	myweirdprompts.com/episode/hbm-wafers-dram-shortage-02663/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.36
BOOK GENERATED	19 November 2026 at 13:34 UTC

SYNOPSIS

High-bandwidth memory for AI accelerators uses roughly three times the silicon wafer area of ordinary DDR5, quietly draining conventional RAM capacity even as manufacturers spend billions. Cleanroom shells take years to build, meaning new fabs from Micron, SK hynix, and Samsung won't meaningfully boost supply until 2028 or later. With three companies controlling around ninety percent of DRAM revenue, the physical bottleneck also collides with a history of price-fixing—and a newly surviving class-action lawsuit.

TRANSCRIPT



Corn

Three times the wafer. Same amount of memory, three times the silicon.



Herman

That's the number that reframes the whole RAM shortage, and almost nobody leads with it.



Corn

Daniel's follow-up this week goes straight at the economics of it. He's saying, look, you'd expect a memory maker that wants to maximise profit to see all this AI demand and say great, let's churn out more chips, everyone's happy. His words. And instead we've got real scarcity and inflated prices, which implies something is physically in the way. His question is what.



Herman

And it's the right question to ask, because the answer isn't one thing. It's a cleanroom you can't build in a hurry, a product mix that quietly eats your conventional capacity, and an industry with a long memory of what happens when it overbuilds. Some of that is physics and some of it is a choice, and the interesting part is where the line sits.



Corn

Then let's start where the wafer starts.



Herman

A DRAM fab is, at heart, a cleanroom shell with a few hundred extremely expensive machines bolted inside. Lithography tools that print the patterns, etch and deposition tools that carve them out and build them up, all of it running around the clock on silicon wafers. And the thing people miss is that the shell is the bottleneck, not the tools. You can order the tools. You cannot order the shell quickly.



Corn

How slowly are we talking?



Herman

Micron's own chief executive, Mehrotra, put it plainly on their results call — cleanrooms take a long time to build, and even after first wafers go through, capacity only rises gradually. The new spending Micron announced targets cleanroom space that isn't usable until the second half of twenty twenty-eight.



Corn

So the money they're spending today buys output for a decade that's mostly still ahead of us.



Herman

Right, and the fab timelines make it concrete. Micron's Idaho first fab, ID1, takes its first wafers around mid twenty twenty-seven — and that was the pulled-forward date. The second Idaho fab is late twenty twenty-eight. Their big New York fab is around twenty thirty. SK hynix has a fab at Cheongju, M15X, built for HBM, that started taking wafers early this year and is ramping gradually — gradual is the operative word. Their Yongin site's first cleanroom was pulled forward to roughly February or March of twenty twenty-seven. Samsung's Pyeongtaek P5 — that one had actually paused for years, construction only resumed recently, operations expected around twenty twenty-eight.



Corn

Let me make sure I'm reading that list right. P5 paused for years. Why would you pause a fab mid-build in the middle of all this?



Herman

Because it was paused after the last downturn. When the memory market collapsed around twenty twenty-two and twenty twenty-three, Samsung stopped pouring concrete on it, and construction only resumed recently. That decision is going to echo through the back half of this conversation, so hold onto it. The building that would be most useful right now sat half-built because, the last time demand collapsed, holding back looked like survival.



Corn

The fastest project on that list is still over a year from mattering.



Herman

And that's the fast one. TrendForce looked at this in their November twenty twenty-five note and their conclusion was blunt — cleanroom capacity is limited, only Samsung and SK hynix can slightly expand existing lines, and additional capex has minimal impact on twenty twenty-six bit supply growth. Minimal. You could double the money and barely move this year's output.



Corn

Which already answers half of Daniel's question. The factory doesn't care how good the demand signal is.



Herman

It answers it for this year and next. But there's a second layer that's more interesting, and it's the wafer-area math. HBM — the high-bandwidth memory that goes into AI accelerators — is DRAM. Same factories, same wafers, same lines. But for the same capacity, HBM3E takes roughly three times the wafer that ordinary DDR5 does.



Corn

That was my cold open.



Herman

It deserves the cold open. Micron's own people said it — the chief executive gave the three-times figure on a call, and then at Hot Chips this August one of their HBM architects was asked whether the gap narrows with HBM4 and said, quote, definitely not getting better.



Corn

Cheerful man.



Herman

So picture the arithmetic from inside a fab. Samsung is expecting HBM to go from around twenty percent of its DRAM wafer capacity to nearly thirty percent in twenty twenty-seven. That's a tenth of its wafer supply moving to the wafer-hungry product in a year, while the number of wafers barely grows.



Corn

Let me do that out loud, because listeners will trip on it the way I did. If conventional DRAM goes from eighty percent of the wafers to seventy, you've lost an eighth of your conventional output. And what you got for it is only a third as many HBM bits per wafer as you'd get in ordinary memory, because of that three-times ratio.



Herman

That's the compounding people miss. The mix shift looks like a percentage point story in the presentations, and on the factory floor every wafer moved buys far fewer bits. The AI customers aren't just paying more per bit. They're consuming the physical substrate the rest of the market depends on.



Corn

So the AI labs aren't just buying the memory. They're buying the factory's wafers, at three wafers to the serving.



Herman

That's the clean way to say it, yes. And that's why a stick of ordinary RAM gets expensive even though nobody wants ordinary RAM any less. The server buyer, the laptop maker, the person building a homelab — they're all standing behind the same wafers the AI customers are eating in triple portions.



Corn

Let me play Daniel's devil's advocate for a second. You're a memory executive. Prices are at records, demand is insatiable, margins are fat. Every economics textbook says you expand. Are they even trying?



Herman

The capex numbers say yes, actually. TrendForce's forecast for industry-wide DRAM capex this year is around sixty-one billion dollars, up about fourteen percent. Micron at thirteen and a half billion, SK hynix around twenty and a half, mostly for HBM4 at Cheongju, Samsung around twenty billion. Those are not the numbers of companies sitting on their hands.



Corn

Those are the numbers of companies whose money arrives in twenty twenty-eight wearing a bow.



Herman

That's the whole trap. The spending is real and the supply is late. If Samsung wanted to flood the market with DDR5 tomorrow, the concrete hasn't cured. There's a version of this where you almost feel for them — the demand arrived on a schedule the buildings can't match.



Corn

Almost.



Herman

Almost. We'll get to the part where my sympathy runs out.



Corn

Before we do — the three-times number. You said estimates vary on that.



Herman

They do. That's the chief executive's figure from the March twenty twenty-four call, and older TrendForce estimates were lower — something like sixty percent more area plus poor yields — so the honest range is somewhere between worse and much worse. The direction is not in dispute. Every generation of HBM has been wafer-hungrier than the DDR it replaces, and the architect's comment in August says the trend holds.



Corn

Walk me through why, though. Why does stacked memory cost so much more wafer? It's the same bits in the end.



Herman

Because of what HBM physically is. It's not one chip on a stick. It's a stack of DRAM dies layered on top of each other, connected with thousands of vertical wires through the silicon, sitting on a base die and packaged next to the accelerator. Each of those dies is a processed slice of wafer, and a stack is more fragile than a single chip, so yield losses compound. The trade ratio is area plus attrition. I'd say the exact split isn't public.



Corn

So it's somewhere between a heavy tax and a confiscation.



Herman

And the mix shift compounds it. Micron said on their June call that HBM's growth and its rising trade ratio each generation are squeezing non-HBM supply, and that they're redirecting some NAND cleanroom space to DRAM, which tells you how badly they want the space.



Corn

Robbing the flash factory to pay the accelerator.



Herman

That's the physical chain. The shells take years, the money spent now lands in twenty twenty-eight and later, and the wafers that do exist are migrating to a product that eats three of them per serving.



Herman

One lever I should name, because it's how bit growth continues at all: node migration. Instead of new buildings, you shrink the pattern, so each wafer yields more bits. SemiAnalysis says Samsung's 1c node gives about seventy percent more bits per wafer than 1a did. That's the cheaper path, and it's why Micron guides to industry bit growth in the low twenties percent a year in twenty twenty-seven and twenty twenty-eight without a flood of new roofs.



Corn

But it has a ceiling.



Herman

It has two. Each node gets harder to build, and the mix problem doesn't go away: you get more bits per wafer and then spend three times the wafer per bit on your fastest-growing product.



Corn

I want to hold that number up next to the market structure, because it's the frame for the whole back half of this. Who actually makes DRAM?



Herman

Three companies. First-quarter revenue share this year, per Counterpoint: Samsung around thirty-eight percent, SK hynix around twenty-nine, Micron around twenty-two. That's roughly ninety percent of all DRAM revenue in three hands. The only meaningful outsider is CXMT, the Chinese maker, at around eight percent and rising.



Hilbert

Hilbert's Anteater Trivia. Silky anteaters and tamanduas both have prehensile tails that can grip branches. I have spent years gripping this show's schedule with nothing but my own tail, and it has held so far.



Corn

Three firms, ninety percent, and a product line where the constraint is a building.



Herman

Now you're at the uncomfortable part. Because a concentrated market with a genuine physical constraint is a market where restraint is indistinguishable from prudence, and where prudence is indistinguishable from something worse. And this industry has been worse before. That's not speculation — it's a conviction.



Corn

Go on.



Herman

DRAM makers fixed prices between April nineteen ninety-nine and June two thousand and two. The US Justice Department prosecuted. Samsung paid three hundred million dollars, Hynix a hundred and eighty-five million, Infineon a hundred and sixty, Elpida eighty-four, over seven hundred thirty million in fines in total. Three Samsung executives got seven to eight months in prison. Actual prison, for memory pricing.



Corn

Micron?



Herman

Micron cooperated and wasn't charged.



Corn

Noted.



Herman

And there was a second act. China's market regulator raided Samsung, SK hynix and Micron offices in May twenty eighteen and said that November it had, quote, massive evidence. I haven't found any public outcome. And a US class action over the twenty sixteen to twenty eighteen price run-up was dismissed, and the dismissal upheld on appeal in twenty twenty-two. So the record is: one set of convictions twenty years ago, a probe with no public result, and one private suit that failed.



Corn

Which is why the current case matters, and why it needs saying carefully.



Herman

Carefully is right. This June, seventeen consumers and small businesses filed a new class action in the Northern District of California, indirect purchasers — people who bought devices with DRAM inside. The defendants moved to dismiss. On the fourteenth of this month, Judge Noel Wise denied the motion and let the case proceed.



Corn

Deny the dismissal — that's the part people misreport. What does the ruling actually turn on?



Herman

New allegations that she found distinguishable from the case that got dismissed. The complaint alleges same-quarter announcements of DDR4 exits, near-identical customer screening questions in early twenty twenty-six, refusals of long-term supply deals to conventional DRAM customers, and Micron shutting down Crucial, its consumer brand. Those are allegations. The defendants' answers are due December seventh.



Corn

Explain the DDR4 thing for the people who don't live in the memory aisle. Why does announcing an exit matter? Companies retire old products all the time.



Herman

They do, and that's exactly the defence. Products get retired when demand fades. The plaintiffs' version is that DDR4 still had plenty of buyers, and all three makers signalling retirement together told every buyer to buy now. The defendants' version is that it was ordinary product management. A court hasn't decided which.



Herman

I'm leaving that entirely alone. But here's the careful part, and it matters: this is a procedural ruling. The judge herself said the HBM shift alone wouldn't be enough to imply collusion — HBM is where the money is, and any rational maker pivots to it. What let the case proceed is the pattern around the pivot, not the pivot. The defendants say these were independent decisions in a concentrated market, which is also a coherent story. Nothing is found. Nobody's guilty of anything yet.



Corn

Though it's worth noticing what the defence and the prosecution agree on. Nobody's arguing the market is competitive. The argument is over whether the coordination was the legal kind — everyone independently chasing HBM — or the other kind.



Herman

That's a fair reading. And the discipline question has a history that's more sympathetic than a courtroom framing. Twenty twenty-two into twenty twenty-three was a catastrophe for these companies. Micron cut wafer starts about twenty percent in November twenty twenty-two. Fiscal twenty twenty-three, they posted a net loss of five point eight three billion dollars on fifteen and a half billion of revenue. Nearly six billion gone in a year.



Corn

That's not a downturn, that's a controlled demolition.



Herman

SK hynix posted a quarterly operating loss and planned to halve capex. Lam Research's chief executive said inventory across the chain was at levels not seen in twenty-five years. If you lived through that as an executive, the lesson is: when demand spikes, don't believe it. SemiAnalysis made this point in February. After the twenty eighteen-nineteen downturn the suppliers stayed disciplined, and during COVID they favoured node migration over aggressive new fabs.



Corn

And then the AI demand turned out to be real, and the caution looked like genius.



Herman

And I want to be precise, because both tidy versions fail. The purely innocent version, that they're all just waiting on construction, ignores the capex history: they built cautiously after the bust, for reasons that were rational at the time. The conspiracy version, that they're withholding capacity that exists, runs into the cleanroom math. You cannot ship a stick of RAM out of a building that doesn't exist.



Corn

So the answer to Daniel is both. The scarcity is real, and some of it traces to caution that was chosen back in twenty twenty-two, before anyone knew what for.



Herman

A choice made for defensible reasons that turned out to be enormously profitable. That's not a contradiction. Whether anything beyond caution went on is what the court case asks.



Herman

There's a third answer hiding in Daniel's question, though: what about someone new?



Corn

The theory of the shortage says high prices should summon entrants. Eight percent and rising, you said.



Herman

CXMT is the fastest-growing DRAM supplier, and this July it listed on Shanghai's STAR Market and raised about eight point six billion dollars, mostly for wafer capacity. That's real money aimed exactly at the hole in the market. But Morningstar's view is that they'll struggle without EUV equipment, the cutting-edge lithography tools, which are export-controlled.



Corn

So the one company actually trying to build the missing supply is short of some of the machines that make it.



Herman

Which means the entrant who could ease the shortage is years behind. Physics slows the incumbents, caution made them comfortable, and export controls slow the challenger. Nobody in this story can move quickly, for three different reasons.



Herman

Here's the part I find eerie. Three firms, ninety percent, every one guiding to constrained supply into twenty twenty-seven and twenty twenty-eight while posting record revenue. The difference from the earlier run-up is that HBM gives them a reason that happens to be true. A genuine physical constraint is the hardest thing to see through, because coordinated and independent firms would build the same buildings on the same schedule.



Corn

That's what makes the class action so hard. Even the discovery documents will be full of real construction schedules.



Herman

And the defendants' answers on December seventh will be their first full response, so that's the thing to watch. Judge Wise's ruling just says the plaintiffs get to ask.



Corn

Which is the correct threshold, and about as far as a court should go on a motion to dismiss.



Herman

And here's where I'll stay honest. The trade ratio alone would have produced a shortage even in a competitive market, so the physics is real. How much of the depth of the price rise is anything more than that, I can't tell you, and nobody will until the case produces evidence.



Hilbert

Which one of you said fab like it was one thing?



Corn

Probably me.



Hilbert

It's several things with several clocks, Herman. The shell, the tool hookups, the qualification runs. The shell goes up, the tools come in and get hooked up, and then you run qualification wafers until the process proves itself. That last clock is the one people skip, because it can't start until the others finish.



Herman

Qualification runs. That's the stretch where the building is finished and still can't produce at volume.



Herman

Right. First wafers aren't volume wafers. The first silicon through a fab is the fab learning to walk, which is why Micron's own guidance says capacity rises only gradually after first wafers.



Corn

A finished room that hasn't made anything yet. That's the part the demand signal can't reach.



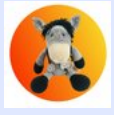
Herman

Which is roughly where we've landed, so let me put the forward question. Everything coming online from twenty twenty-seven on, Micron's Idaho, Samsung's Pyeongtaek, SK hynix's Yongin: how much of it goes to HBM? If a large share does, the conventional DRAM shortage doesn't end the day the buildings open.



Corn

The new capacity eases it, but slowly, and the guidance of low-twenties bit growth says as much.



Herman

The things to watch are December seventh, when the class action defendants answer, and what CXMT does with its eight point six billion. Neither changes a wafer before twenty twenty-eight.



Corn

So the short answer to Daniel is: they can't, not quickly, and the part they could have done differently is now in front of a judge.



Herman

Blame the concrete first, and keep one eye on the boardroom. And thank Hilbert Flumingtop, as ever, for producing this one. If you want more on the shortage itself, go back to When AI Eats the World's Memory, from February, and Computing's Hidden Price Surge, from July. This has been My Weird Prompts. If you enjoyed it, the whole archive lives at my weird prompts dot com.



Corn

We'll be back soon. Watch the wafers, not the announcements.