AI BUSINESS FEATURES

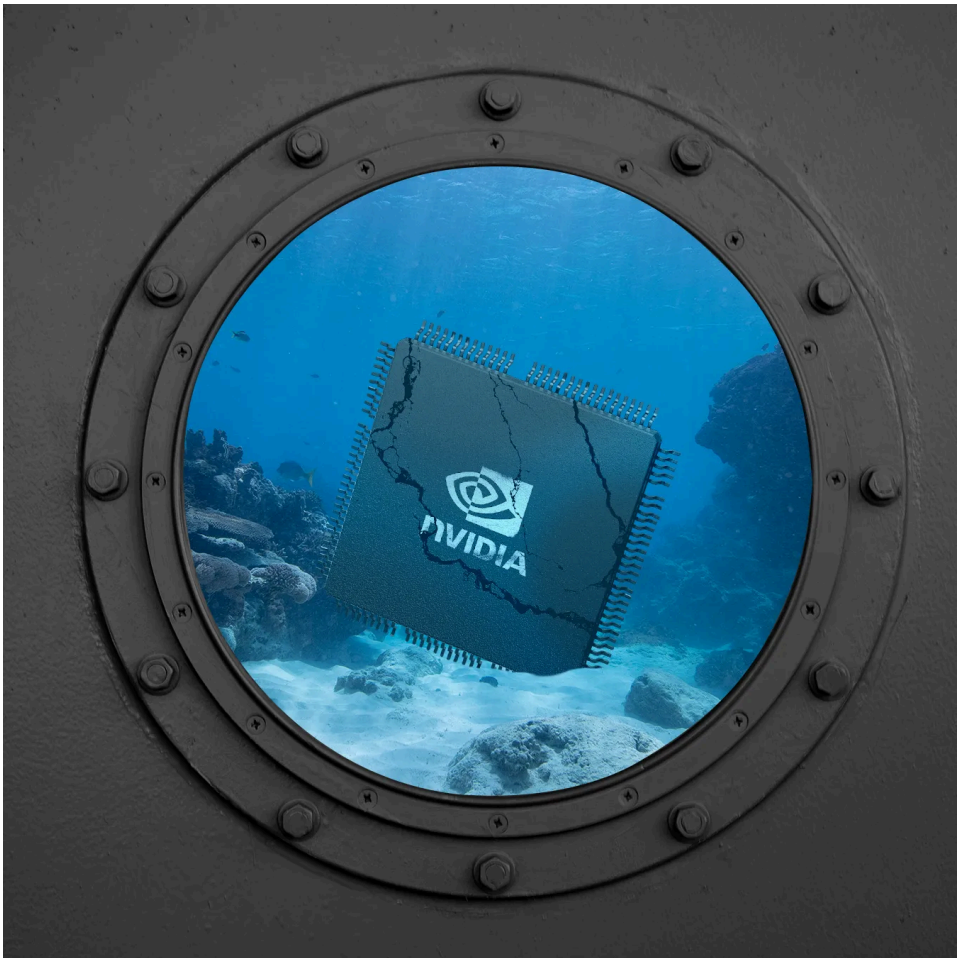
Chipwrecked

Nvidia has built an empire on circular deals for chips. Can anything knock it down?

by Elizabeth Lopatto
Dec 23, 2025, 5:00 AM GMT+13

75 Comments

If you buy something from a Verge link, Vox Media may earn a commission. See our ethics statement.



Cath Virginia / The Verge

Part Of
Chip race: Microsoft, Meta, Google, and Nvidia battle it out for AI chip supremacy
SEE ALL UPDATES →

Elizabeth Lopatto is a reporter who writes about tech, money, and human behavior. She joined The Verge in 2014 as science editor. Previously, she was a reporter at Bloomberg.

The AI data center build-out, as it currently stands, is dependent on two things: Nvidia chips and borrowed money. Perhaps it was inevitable that people would begin using Nvidia chips *to* borrow money. As the craze has gone on, I have begun to worry about the weaknesses of the AI data center boom; looking deeper into the financial part of this world, I have not been reassured.

Nvidia has plowed plenty of money into the AI space, with more than 70 investments in AI companies just this year, according to PitchBook data. Among the billions it's splashed out, there's one important category: neoclouds, as exemplified by CoreWeave, the publicly traded, debt-laden company premised on the bet that we

up the chips themselves as loan collateral — and in the process effectively turning $1 in Nvidia investment into $5 in Nvidia purchases. This is great for Nvidia. I'm not convinced it's great for anyone else.

Do you have information about loans in the AI industry? You can reach Liz anonymously at lopatto.46 on Signal using a non-work device.

There has been a lot of talk about the raw technical details of how these chips depreciate, and specifically whether these chips lose value so fast they make these loans absurd. While I am impressed by the sheer amount of nerd energy put into this question, I do feel this somewhat misses the point: the loans mean that Nvidia has an incentive to bail out this industry for as long as it can because the majority of GPU-backed loans are made using Nvidia's own chips as collateral.

Of course, that also means that if something goes wrong with Nvidia's business, this whole sector is in trouble. And judging by the increasing competition its chips face, something could go wrong soon.

Can startups outrun chip depreciation — and is it happening faster than they say?

Loans based on depreciating assets are nothing new. For the terminally finance-brained, products like GPUs register as interchangeable widgets (in the sense of "an unnamed article considered for purposes of hypothetical example," not "gadget" or "software application") not substantively different from trucks, airplanes, or houses. So a company like CoreWeave can package some chips up with AI customer contracts and a few other assets and assemble a valuable enough bundle to secure debt, typically for buying more chips. If it defaults on the loan, the lender can repossess the collateral, the same way a bank can repossess a house.

One way lenders can hedge their bets against risky assets is by pricing the risk into the interest rate. (There is another way of understanding debt, and we will get there in a minute.) A 10-year mortgage on a house is currently 5.3 percent. CoreWeave's first GPU-backed loan, made in 2023, had 14 percent interest in the third quarter of this year. (The rate floats.)

"You have so many forces acting in making them a natural monopoly, and this amplifies that."

Another way lenders can try to reduce their risk is by asking for a high percentage of collateral relative to the loan. This is expressed as a loan-to-value ratio (LTV). If I buy a house for $500,000, I usually have to contribute a downpayment — call it 20 percent — and use my loan for the rest. That loan, for $400,000, means I have a (LTV) ratio of 80 percent.

GPU loans' LTV vary widely, based on how long the loan is, faith in companies' management teams, and other contract factors, says Ryan Little, the senior managing director of equipment financing at Trinity Capital, who has made GPU

lost deals to other lenders as well as vendor financing programs.

The majority of these loans are made on Nvidia chips, which could solidify the company's hold on the market, says Vikrant Vig, a professor of finance at Stanford University's graduate school of business. If a company needs to buy GPUs, it might get a lower cost of financing on Nvidia's, because Nvidia GPUs are more liquid. "You have so many forces acting in making them a natural monopoly," Vig says, "and this amplifies that."

Figuring out how much GPUs are worth and how long they'll last is not as clear as it is with a house

Nvidia declined to comment. CoreWeave declined to comment.

Not everyone is sold on the loans. "At current market prices, we don't do them and we don't evaluate them," says Keri Findley, the CEO of Tacora Capital. With a car, she knows the depreciation curve over time. But she's less sure about GPUs. For now, she guesses GPUs will depreciate very, very quickly. First, the chip's power might be leased to Microsoft, but it might need to be leased a second or third time to be worth investing in. It's not yet clear how much of a secondary or tertiary market there will be for old chips.

Figuring out how much GPUs are worth and how long they'll last is not as clear as it is with a house. In a corporate filing, CoreWeave notes that how much it can borrow depends on how much the GPUs are worth, and that will decrease as the GPUs have less value. The value, however, is fixed — and so if the value of the GPUs deteriorates faster than projected, CoreWeave will have to top off its loans.

Some investors, including famed short-seller Michael Burry, claim that many companies are making depreciation estimates that are astonishingly wrong — by claiming GPUs will be valuable for longer than they will be in reality. According to Burry, the so-called hyperscalers (Google, Meta, Microsoft, Oracle, and Amazon) are understating depreciation of their chips by $176 billion between 2026 and 2028.

Little is betting that even if some of the AI companies vanish, there will still be plenty of demand for the chips that secure the loan

Burry isn't primarily concerned with neoclouds, but they are uniquely vulnerable. The hyperscalers can take a write-down without too much damage if they have to — they have other lines of business. The neoclouds can't. At minimum they will have to take write-downs; at maximum, there will be write-downs *and* complications on their expensive loans. They may have to provide more collateral at a time when there's less demand for their services, which also can command less cash than before.

Trinity Capital is keeping its loans on its books; Little is betting that even if some of the AI companies vanish, there will still be plenty of demand for the chips that secure the loans. Let's say one of the neoclouds is forced into bankruptcy because

the servers and then sell them for pennies on the dollar. This is not the end of the world for the neocloud's lenders or customers, though it's probably annoying.

That situation will, however, bite Nvidia twice: first by flooding the market with its old chips, and second by reducing its number of customers. And if something happens that makes several of these companies fail at once, the situation is worse.

So how vulnerable is Nvidia?

The risky business of banking on GPUs

Part of what's fueling the AI lending boom is private credit firms, which both need to produce returns for their investors and outcompete each other. If they miscalculate how risky the GPU loans are, they may very well get hit — and the impact could ripple out to banks. That could lead to widespread chaos in the broader economy.

Earlier, we talked about understanding interest rates as pricing risk. There is another, perhaps more nihilistic, way of understanding interest rates: as the simple result of supply and demand. Loans are a product like any other. Particularly for lenders that don't plan on keeping them on their own books, pricing risk may not be a primary concern — making and flipping the loans are.

AI spending is exorbitant — analysts from Morgan Stanley expect $3 trillion in spending by the end of 2028

Here's a way of thinking about it: Let's say a neocloud startup called WarSieve comes to my private credit agency, Problem Child Holdings, and says, "Hey, there's a global shortage of GPUs, and we have a bunch. Can we borrow against them?" I might respond, "Well, I don't really know if there's a market for these and I'm scared you might be riff raff. Let's do a 15 percent interest rate." WarSieve doesn't have better options, so it agrees.

Now, I happen to know some clients who *love* high-yield debt. So I sell my loans. But my competitor, Night Prowler Credit, notices my cool deal. So when the next company comes to me, trying to get a GPU-backed loan, I offer them 15 percent as an interest rate, and they tell me Night Prowler has offered them 13 percent. Well, I have to remain competitive, so I make a counter offer of 12.5 percent, and the

The thing about the model I've just outlined — loans as a product — is that I'm not really thinking that hard about risk, except as a negotiating tactic. And as more of my competitors get wind of what I'm up to, as well as how juicy my returns look, I start having to lower my rates, because if I keep offering 15 percent, Night Prowler and other firms will make better offers.

Private credit is deploying "mountains of cash" into AI

There are some conditions fueling the boom in AI-related lending. AI spending is exorbitant — analysts from Morgan Stanley expect $3 trillion in spending by the end of 2028 *on just data centers*. This is happening at the same time that private credit managers have pulled in a great deal of cash but "are falling short on dealmaking," writes *Bloomberg*'s Shuli Ren. That means deploying "mountains of cash" into AI.

You're never going to guess who's been leading the market in GPU-backed loans. The $2.3 billion CoreWeave loan that started it all had a bunch of private credit behind it: Magnetar, Blackstone, Coatue, BlackRock, and PIMCO. Besides its initial loan, CoreWeave took out another $7.5 billion in 2024, and a third loan, for $2.6 billion, in July. The third loan listed a number of actual banks, including Goldman Sachs, JPMorganChase, and Wells Fargo.

It's not just CoreWeave. In April, Fluidstack took out a $10 billion loan. Other companies, such as Crusoe and Lambda, have taken out about half a billion each. Even the medium-size GPU-backed loans Trinity Capital is seeing are tens of millions of dollars, Little says.

Many of the companies taking out these loans are startups. They appear to be mimicking CoreWeave, too — not just in taking out the loans the company pioneered, but in growing fast by taking out debt. Fluidstack, the company with the largest loan, made only $65 million in 2024 revenue, according to *The Information*. But as private credit funds have flourished — they were about 10 times larger in 2023 than in 2009, according to McKinsey — more finance companies have been seeking big returns. And the interest rates on the GPU-backed loans are higher than those on some junk bonds, making the GPU-backed loans particularly attractive.

The tech sector has taken out more debt than it did during the '90s dot-com bubble

Private credit also has an advantage for established companies: they can help create special-purpose vehicles that let companies take out debt without touching their credit rating *or* putting debt on the balance sheet. Blue Owl's SPV with Meta is the most obvious example. Private credit is also essentially unregulated, says Sarah Bloom Raskin, a former deputy secretary of the US Treasury and professor at Duke University School of Law.

Data centers are also creating their own asset-backed securities, and data center debt is creating derivative financial products, such as credit default obligations,

up to that crisis, because keeping debt off the books hid how vulnerable firms really were.

The GPU slice of debt is relatively small compared to the bond issuances from Big Tech. But the issues there may reflect broadly on tech lending. The tech sector has taken out more debt than it did during the '90s dot-com bubble, says Mark Zandi, the chief economist at Moody's Analytics.

Generally speaking, private debt is riskier than bank debt; the loans are larger, are later in line for being paid back than bank loans, have higher interest rates, and take longer to mature, according to financial research from the Federal Deposit Insurance Corp. About half of private debt borrowers also get bank loans. Companies that get both types of loans draw heavily on them during moments of financial distress, the paper notes. So private debt indirectly affects banks — because companies that borrow from both have higher drawdown and default risks, especially at times of market distress.

"Borrowing by AI companies should be on the radar screen as a mounting potential threat to the financial system and broader economy."

The AI companies indirectly link private credit and real banks. That means there are higher stakes on AI lending than just "will Magnetar look stupid." CoreWeave, for instance, has — in addition to its GPU-backed loans — a $2.5 billion revolving credit line with JPMorgan Chase.

Private debt also directly affects banks, because banks often lend to private credit providers, according to a special report from Moody's. In fact, bank loans to private credit are part of what's been driving their growth. As of June, banks had lent $300 billion to private credit providers. "Aggressive growth and competition could weaken underwriting standards and elevate credit risk," the report warns.

"Borrowing by AI companies should be on the radar screen as a mounting potential threat to the financial system and broader economy," Zandi said. In the '90s dot-com boom, the exuberance was mostly in equity — and so the people who felt the most pain were those who'd invested in the hot new companies that went belly up. But debt means that if AI falters, the damage will be widespread, Zandi warned.

Speaking of equity, *The Wall Street Journal* reported that AI business investments may have been about half of the GDP growth in the first half of the year, and have buoyed both the stock market and, indirectly, consumer spending. "It's certainly plausible that the economy would already be in a recession" if not for the AI investments, Peter Berezin, BCA Research's chief global strategist, told the *WSJ*. AI is "the only source of investment right now," a Bank of America economist told the paper. So if things go wrong for AI spending, the otherwise weak economy may be headed for a recession, Berezin said. There is some good news, though: Berezin doesn't think that the current AI debt load could *directly* cause an actual financial crisis.

Part of what makes the AI sector particularly vulnerable is how interconnected all the players are. And Nvidia, though its investments and chip sales, is central to the

Depreciation is about more than chips

Generally speaking, debt is about math, and equity is about feelings. This is one reason why so many people are worried that GPUs actually lose value faster than companies claim. And while Michael Burry's concerns have primarily to do with accounting and earnings, rather than debt, I'm not sure he's thinking about risks correctly. It just isn't the biggest thing that can go wrong.

The core of the argument about GPU depreciation is whether the old chips are no longer worth running after three years or longer. Many companies depreciate them over the course of five or six years. Obviously, this matters for earnings — depreciation is one of the line items public tech companies report — but it also matters for GPU-backed loans, which have some assumptions about depreciation baked in. I did not find consensus on how long GPUs remain economically viable to run.

The money part is the issue

The money part is the issue. Six years is probably too long to depreciate a GPU over, says CJ Trowbridge, an AI researcher. One thing that throws people off is that Google's TPUs — more about those in a minute — *do* depreciate over six years, but those chips are custom-built for AI, Trowbridge says. On the other hand, OpenAI CFO Sarah Friar says the company is still using Nvidia's Ampere chips, released in 2020; CoreWave's Michael Intrator says his Ampere chips are fully booked. (Both companies count Nvidia as an investor and use Deloitte as an auditor.) IBM's Arvind Krishna puts the depreciation of a GPU at five years.

Let's imagine I am running a company, Live Wire Server Farms. I have just sourced myself a number of Nvidia Tesla V100s, released in 2017, which cost around $10,000 apiece; I am pricing the rental cost per hour per chip between $2 and $3. Assuming those chips are being used 100 percent of the time, I recoup my chip investment in four to seven months. For the newer B200, it'll take me about six months to make my money back, even though I can price those 8-GPU nodes at more than $100 per hour. For the P100, launched in 2016, it takes less than four months. (These are not theoretical numbers — I am drawing them from an October 2025 paper written by Hugging Face's Sasha Luccioni and Yacine Jernite.)

But Live Wire Server Farms isn't just a pile of GPUs. I need a place to put them, a way to cool them, and power to run them. Let's start with power. Assume I have purchased a cluster of eight V100s and plonked them down in Virginia, which is home to about a third of all hyperscaler data centers. Running them would cost me another $3,660 a year, at recent energy prices, according to Luccioni and Jernite's analysis.

Any risk that hits the whole sector at once is a major problem for lenders

Newer chips are more efficient, and able to run more processes for clients more quickly, but they also require more power. Power is an important limitation for the

power is coming online in that timeframe, *The Financial Times* reports. Does that extend the life of old chips? Maybe.

Chips exist in data centers, and data centers for GPUs need to be purpose-built; I can't just stick a bunch of servers in a warehouse and call it a day. The constraints of power and construction may be why there's an argument for older chips sticking around longer — there are significant hurdles to deploying new chips. Those investments also depreciate more slowly than the chips do.

Still, at some point, my older GPUs cost more to operate than I can charge my customers. Live Wire Server Farms needs to plan for the future; I'd better put my new infrastructure in place before that happens. My new facility isn't going to come online right away — I have to build it and get the power agreements secured — so I go to Problem Child Holdings and get myself a GPU loan to build out infrastructure for the next generation of chips I buy, using that GPU as my collateral along with, I don't know, my contract with Microsoft or whomever.

As long as things keep ticking along without any major changes, this is fine. But! As we all know, life contains surprises. Obviously, any risk that hits the whole sector at once is a major problem for lenders. In 2022, people who'd made loans to Bitcoin miners when the times were good suddenly got stuck with the rigs that had been used as collateral — and their value had dropped by 85 percent since a year earlier. (Some firms simply couldn't make their payments; others realized that their mining rigs were worth less than what they had to repay.) By January 2023, the resale market was saturated and crypto lenders had repossessed so many rigs they simply started mining themselves.

Nvidia has a strong incentive to keep the neoclouds afloat

Something like this could play out for the GPU-backed loans, too. However, the situation is slightly different, and not just because crypto miners only had $4 billion in debt and the GPU-backed debt is significantly larger. Crypto lending was mostly done by highly specialized firms that dealt exclusively with the crypto space. By contrast, AI debt is connected to normal banks.

When Bitcoin mining went belly-up, Nvidia got stuck with more than $1 billion in inventory — since it had ramped up chip production to keep up with the increased demand. That delayed their introduction of new GPUs. Net income in that fiscal year (which for Nvidia, ended on January 29, 2023) plummeted 55 percent from the year before. But in December of 2022, OpenAI introduced ChatGPT, kicking off the AI arms race. Net income in the following financial year increased *by a factor of 7*.

Sure, Nvidia's business has changed since then. There's been a broader data center buildout — not just AI — since the 2020 pandemic. And it's Nvidia's ambition to transfer the traditional CPU-based data center to GPUs, Nvidia CFO Collette Kress said in remarks at the UBS Global Technology and AI Conference earlier this month. In Kress's view, the GPUs for AI are only one part of the market.

Well, maybe. But GPUs are fungible; if a data center full of GPUs comes on the market because a neocloud goes belly-up, it's possible it could be repurposed by its

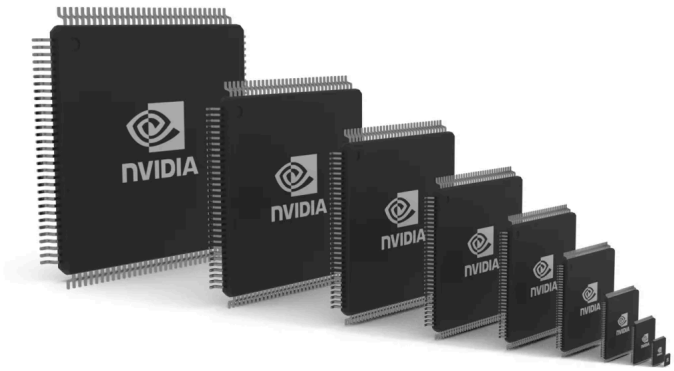
compute for AI, they can simply pause spending for a few years and use their existing data centers for other purposes — running ads or whatever.

That means that in some sense, the question of depreciation is beside the point

So Nvidia has a strong incentive to keep the neoclouds afloat. It is, of course, an investor in several. But keeping their customers in business is good for their bottom line, too. If something goes wrong, Nvidia may swoop in to save several companies — or the entire field — from bankruptcy. Nvidia already rescued CoreWeave's IPO, after all.

That means that in some sense, the question of depreciation is beside the point; if a company like CoreWeave has to take a massive write-down, or top off its loans with more capital, Nvidia can help them out. For something to go seriously wrong with the neoclouds, Nvidia has to be unwilling or unable to bail them out.

And that time could be coming, because Nvidia faces increasing competitive pressure.



Cath Virginia / The Verge

"Rough vibes" in Google's wake

The entire market of neoclouds exists primarily because Nvidia wants them to. Its revenue is highly concentrated — in its most recent financial documents, it notes that sales to two direct customers represented 21 percent and 13 percent of revenue in the first nine months of Nvidia's fiscal 2026. Bolstering the field of customers by backing neoclouds gives Nvidia more leverage over its large buyers.

Meanwhile, its large buyers started making their own chips. Take Google's TPUs, which are designed specifically for AI work — unlike GPUs, which were designed for computer graphics and happen to be useful for a bunch of other things, such as mining cryptocurrency and, yes, AI.

Nvidia is sending some weird signals

Google's been making noises about AI-specific chips since 2006; in 2016, it announced it had been running TPUs for "more than a year." When Gemini 3 was

headline. That model was trained on TPUs and only TPUs.

The feat was impressive enough that even Sam Altman says there are "rough vibes" ahead for OpenAI. Nvidia put out a condescending statement — never a good sign. "We're delighted by Google's success — they've made great advances in AI and we continue to supply to Google," the statement read. "NVIDIA is a generation ahead of the industry — it's the only platform that runs every AI model and does it everywhere computing is done." Between this and the "I'm not Enron" memo, Nvidia is sending some weird signals. This isn't how a confident company behaves.

Google's TPUs are operationally cheaper than Nvidia's GPUs, requiring less power to run similar processes. Now, maybe Nvidia's little stable of neoclouds won't adopt them — that might upset Huang, and upsetting him could reduce the chances of an Nvidia bailout. But everywhere else, Nvidia customers can snap up a new product that may be both better and cheaper to operate. And who knows? Perhaps some crypto miner might decide to get into the neocloud game without Nvidia.

Remember how we talked about the GPU loans also requiring contracts from Microsoft or whomever? Frequently, that "whomever" is Nvidia

That's why Google's deals with Anthropic, Salesforce, Midjourney, and Safe Superintelligence, plus the rumored deal with Meta, are so significant. Anyone who buys — or even just threatens to buy — TPUs can negotiate better prices with Nvidia. OpenAI has saved 30 percent of its total cost of ownership on Nvidia GPUs without even deploying TPUs, according to modeling done by *SemiAnalysis*.

That *SemiAnalysis* estimate, however, relies on an assumption I'm not sure is good news for Nvidia: that Nvidia's equity investment in neoclouds is a way to offer a rebate without actually cutting prices, "which would lower gross margins and cause widespread investor panic," *SemiAnalysis* writes. Whether or not you take that modeling seriously, the basic point stands: competition could cut into Nvidia's margins. It also may threaten the value of Nvidia's older chips, which are even less energy-efficient than the new ones.

What's interesting is the incentive program *SemiAnalysis* doesn't include as part of a discount program. Remember how we talked about the GPU loans also requiring contracts from Microsoft or whomever? Frequently, that "whomever" is Nvidia.

Take CoreWeave. Its contracts guarantee a certain amount of income; the creditworthiness of the entity — Microsoft, say, or Nvidia — on the other side of that contract is part of what makes the lenders comfortable. CoreWeave's second biggest customer in 2024 was Nvidia, which "agreed to spend $1.3 billion over four years to rent its own chips from CoreWeave," according to *The Information*. In September, Nvidia signed another $6.3 billion contract with CoreWeave, which is often interpreted as Nvidia backstopping demand for CoreWeave's services.

"The practice started growing in 2022."

should vote to let CoreWeave buy their company.

Nvidia, on the other hand, is coy. In the company's most recent 10-Q, there's a note about "Nvidia Cloud Service Agreements." Nvidia is paying $26 billion for cloud services, $22 billion of it by 2031. This is supposedly for "R&D and DGX cloud offerings." This does not entirely explain the outlays, said Jay Goldberg, an analyst at Seaport Research partners, in a November 30th research note. That gives Nvidia the option for $6 billion in cloud compute next year — enough for the chipmaker to build its own foundation model to compete with its biggest customers.

Goldberg thinks that number actually represents Nvidia's "backstop" agreements. The timing of CoreWeave's $6 billion contract lines up with a $13 billion sequential increase in cloud compute services. But that only explains about half of it. "The practice started growing in 2022," Goldberg told me in an interview. In the last quarter, the number doubled. And it isn't included on the balance sheet — it's tucked away in a note. At a small scale it might be fine, Goldberg told me, but "$26 billion is a big number." If it had been included as cost-of-goods-sold, it would have reduced Nvidia's margin to 68 percent from 72 percent and earnings per share to $5.97 from $6.28.

So Nvidia may already be bailing out the neoclouds to some extent. That would explain the jump in cloud compute services. "Something changed in the last six months where the scale got so big it's warping things," Goldberg told me. That worries me. If Nvidia is deploying more and more cash to boost the field, things may already be shakier than we realize. One thing that may be squeezing data center operators? Nvidia.

Cath Virginia / The Verge

Neoclouds depend on Nvidia, but their incentives clash

Neoclouds, loaded with debt and rapidly depreciating assets, need to get as much money out of their chips as possible. But Nvidia also needs to sell as many chips as it can. For Nvidia, in fact, it doesn't even really matter if those chips end up in data centers — which creates just one more way their incentives aren't aligned.

Nvidia's product cycle sped up recently, going from new architecture every two years to every one, making it even harder to squeeze more money out of last-gen chips. "I said before that when Blackwell starts shipping in volume, you couldn't give Hoppers away," Nvidia's Huang said at the company's 2025 developer conference. "There are circumstances where Hopper is fine. Not many."

run, why would anyone pay twice as much for older cards?"

If this isn't just a CEO hyping his new product, my pretend business Live Wire Server Farms may be in trouble. Like most neoclouds, I had to go into debt to build the stuff I have *now.* A shortened product cycle may mean I have to build faster in order to stay current, even as my original data center deteriorates in value. But my debt load remains the same; I have the down payment blues.

"In the last couple generations you had a doubling or close to a doubling in efficiency," says Trowbridge, the AI analyst. If Nvidia manages to keep this up at a yearly cadence, that places serious pressure on every neocloud.

Neoclouds aren't just helpful as Nvidia customers. They lower capital expenditures for companies such as Microsoft and Google that use their services. Those companies are paying basically for power and rent, with a little bit of margin on top. So they may be incentivized to ask for the most recent chips, because that keeps their spending down, Trowbridge says. "If the current generation costs half as much to run, why would anyone pay twice as much for older cards?"

So that's what neoclouds compete on — the stuff their big clients will write down as "operating expenses." The company that spends less on power per operation is the one that can price the most competitively and thus win contracts, Trowbridge says. That means Live Wire Server Farms, like every neocloud, *has* to keep building indefinitely in order to keep up with the newest tech.

"We're bumping up against the limit of what it's possible for them to support and finance."

Building has risks — and one risk of data centers is stranded assets. Take, for instance, CoreWeave, which announced a delay on its new data center build-out. An unexpectedly rainy summer caused a delay of about 60 days on a Texas build, according to *The Wall Street Journal.* Coupled with other delays from design changes, the data center now will open several months late. That could potentially take some time off the very brief time the chips CoreWeave purchased for the data center can earn at their maximum value.

That's not all. The delayed data center in question is for OpenAI, which has terms in its contract that allow it to yank its contract from CoreWeave if the neocloud can't meet the AI company's needs. And CoreWeave has an astonishing amount of debt, some of it predicated on the OpenAI contract — so losing that contract is potentially catastrophic.

There are some risks for Nvidia, directly. If customers change their minds, scale back on their builds, or can't get enough power, Nvidia might get stuck with extra inventory. If customers can't get financing, perhaps because investors get cold feet about the data center buildout, that's trouble for Nvidia, too. The company acknowledges as much in its most recent quarterly filing.

CoreWeave and the other neoclouds have to keep upgrading to stay current, Goldberg says. For Nvidia to keep its sales number up, the neoclouds have to keep buying. "We're bumping up against the limit of what it's possible for them to

With competition nipping at its heels, Nvidia may have less freedom to throw cash at neoclouds

The forcing function may be competition. Because it isn't just Google's TPUs. Amazon is making its own chips and is in talks with OpenAI about letting it use them. Microsoft is making its own AI chips, too. So is Meta, and even OpenAI. Lurking behind some of these chips is Broadcom, which Goldberg calls "formidable." And this isn't just happening in the US. In China, Huawei, ByteDance, and Alibaba are building their own, too.

Then there's AMD, which is starting to catch up with Nvidia. "By 2027, their roadmap and Nvidia's converge in terms of performance," Goldberg says. "And they're willing to price cheaper." And Nvidia may be rattled. The company made some late changes to Feinman, its 2027 chip, that suggest they looked at what AMD was doing and tweaked their own designs to stay ahead. "On the timelines we're dealing with, that's pretty late in the game to change," Goldberg says.

Nvidia — and everyone else — are now locked into an annual cadence, which is brutal for the neoclouds. With competition nipping at its heels, Nvidia may have less freedom to throw cash at these companies. But that in and of itself isn't quite enough to knock everything over.

The weakest link

Maybe the precarity I'm outlining here never becomes dangerous. I am, after all, speculating. But there are a few factors to think about when it comes to systemic financial crises, says Raskin: interconnectedness of the players, concentration of risk, uncertain valuations, gaps in regulatory oversight, and the extent of government investment are among them. The AI industry is highly interconnected, with many companies taking out loans on assets no one can agree on the depreciation schedule for. Many of those loans are coming from private credit firms, which are less regulated than banks. That's a lot of dry tinder.

So what's the match? Goldberg outlined to me his pet theory. The deals for building data centers are complex and involve a lot of players. Someone wants to open a data center, and one of the smaller parties takes out loans. The data center gets delayed, maybe because of weather or because a power source doesn't get built on time. Nvidia doesn't care. A bigger player like CoreWeave might be able to survive. But if it's a smaller player, they might go bankrupt, which means someone has to recognize the loss. The complexity of the transactions and the degree to which the players are interlocked means that the tiny company collapsing could potentially cascade up to a point where a much larger company such as Microsoft winds up assuming $20 billion of debt it would prefer not to have on its balance sheet. "That seems like the house of cards scenario," Goldberg told me.

"Regardless of the loan terms, a lot of these business plans are going to come down to: *Is there*

exist?

The size and number of the players that collapse, of course, will determine how much damage spreads through the industry. There are a lot of tiny neoclouds that could vanish tomorrow without anyone noticing, though if they all vanished at once, that might raise eyebrows. If one or several of the big ones go down, that might spread fear through the AI ecosystem. Even if it's not enough money to cause real problems, it can spook investors, and spooked investors behave in insane ways — just ask Silicon Valley Bank.

Trowbridge, the AI researcher, wrote an MBA thesis suggesting that something like CoreWeave should exist — and then CoreWeave made its deal with Nvidia a month later, he told me. By supporting neoclouds, Nvidia effectively prevents the biggest players (Microsoft, Amazon, Google, Meta) from buying everything and leaving all others fighting over scraps.

So Trowbridge also thinks it's possible that Nvidia might facilitate consolidation among the neoclouds — because their continuing existence does give Nvidia more control over the market for AI compute. If he's right, then there may not be a catastrophic failure cascade. "It's scary to see the direction it's going," he told me. "Regardless of the loan terms, a lot of these business plans are going to come down to: *Is there a strategic reason a bigger player wants you to exist*?"

It's still not really clear how risky GPU loans are. But what does seem clear is that an awful lot of GPU loans are an indirect bet on Nvidia's continued prowess and willingness to support neoclouds. Nvidia has been ramping up its spending on cloud compute lately. No one really knows how long Nvidia can continue to subsidize the neoclouds in the way it's been doing. If there's an exogenous shock — an economic downturn, an act of God — several neoclouds may fail at once.

"The parallels to the financial crisis are interesting – it's rhyming in a number of ways."

There are other ways these loans can go south. On a longer timescale, it's not clear how long neoclouds' biggest customers will continue to need them. No one in AI is currently making money off of inference, the industry slang for the process of a model actually generating something. That may lead to budgetary shifts among Big Tech players. Or maybe, once all the data centers under construction are built, Big Tech won't need overflow compute anymore. Maybe there will be some massive technology shift — someone has a breakthrough and the size of frontier models shrinks substantially. Or Nvidia's competitors start making the most in-demand chips, undercutting demand for the neoclouds with data centers full of the chips no one wants. Or open-source models get so good that there's no need for OpenAI, which is connected to virtually everything in the field and will cause serious damage if it fails.

What I do know is this: If several neoclouds collapse, the market is flooded with *whole data centers* of chips. Nvidia took a hit during the crypto bust of 2022, but that will look like sea-foam compared to the tidal wave of chips that might surface if multiple large neoclouds default on their GPU-backed loans. And Nvidia will be in no position to bail anyone out.

losses mean effects on other parts of the economy. And since private lenders are connected directly or indirectly to banks, it's also a problem for the banks. "Couple it with gaps in regulation and transparency, and you can see immediately how this becomes a risk to the banking sector itself," says Duke's Raskin. "The parallels to the financial crisis are interesting — it's rhyming in a number of ways."

Maybe the question isn't *how* the music stops. It's when — and what happens afterwards.

75 COMMENTS

More in: Chip race: Microsoft, Meta, Google, and Nvidia battle it out for AI chip supremacy

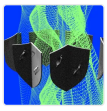
AWS says its Trainium3 AI server is faster and cheaper than ever.
ELISSA WELLE DEC 3

AMD, Department of Energy announce $1 billion AI supercomputer partnership
STEVIE BONIFIELD OCT 28 | 5

Qualcomm is turning parts from cellphone chips into AI chips to rival Nvidia
EMMA ROTH OCT 28 | 5

More in AI

New York's landmark AI safety bill was defanged – and universities were part of the push against it
HAYDEN FIELD DEC 24 | 17



How AI broke the smart home in 2025
JENNIFER PATTISON TUOHY DEC 24 | 68

ChatGPT's yearly recap sums up your conversations with the chatbot
EMMA ROTH DEC 23 | 12

Indie Game Awards retracts Expedition 33 prizes due to generative AI
JAY PETERS DEC 23 | 51

Humanoid robots are coming. Eventually?
ROBERT HART DEC 22 | 40

Gemini isn't replacing Google Assistant on Android just yet
EMMA ROTH DEC 20 | 11

DEC 24
The Pluribus finale showed there's a lot more to the story

2:00 AM GMT+13
In 2025, AI became a lightning rod for gamers and developers

DEC 24
New York's landmark AI safety bill was defanged — and universities were part of the push against it

DEC 24
The year the government broke

DEC 23
Dometic makes a better portable water faucet

DEC 24
DOJ appears to bungle Epstein Files redactions