

Token Factory

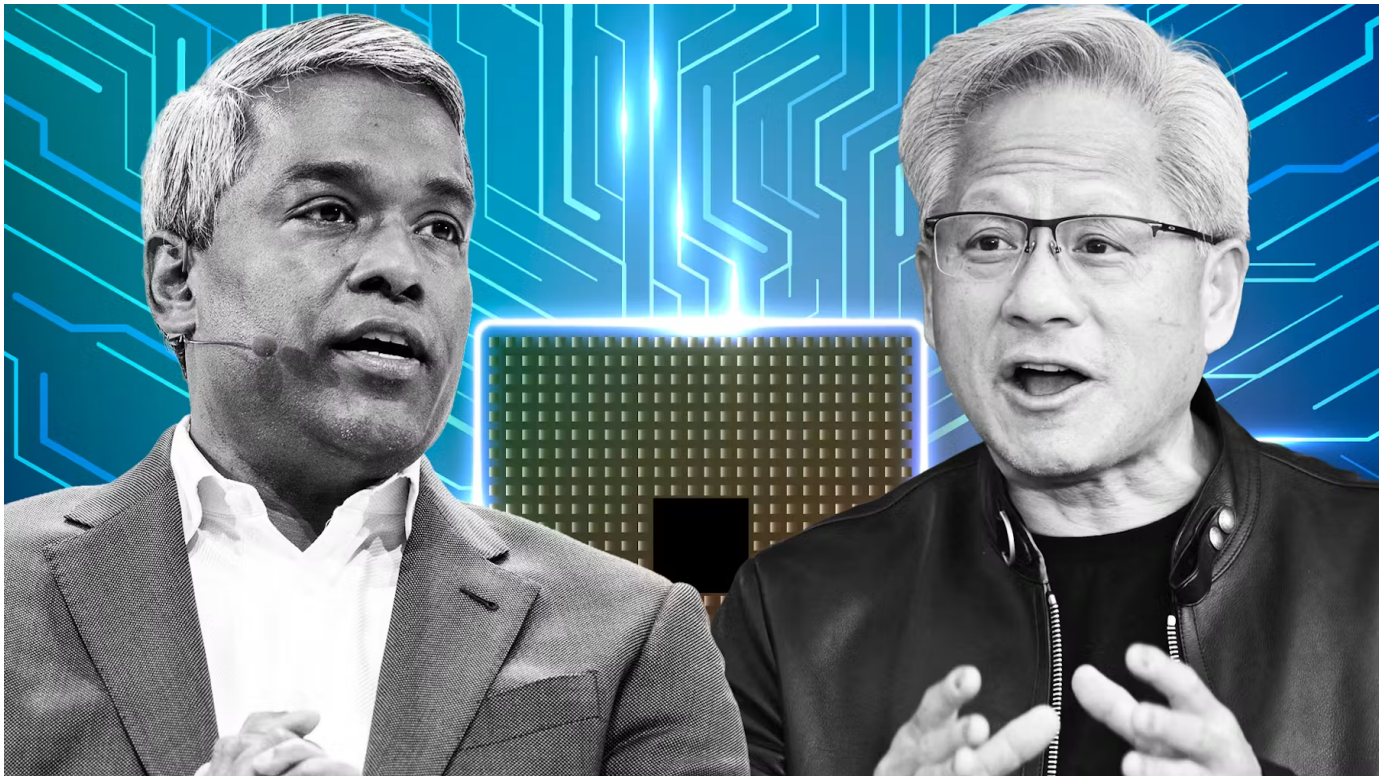
[Learn more](#)

Featured Partner

NEBIUS

AI Infrastructure

Exclusive: Behind Google's TPU Ground War to Lure Nvidia's Most Loyal Customers



Thomas Kurian, left, and Jensen Huang. Photos via Getty; art by Mike Sullivan

By Amir Efrati, Catherine Perloff and Phoebe Liu

Jul 13, 2026, 8:00am PDT



Google is kicking off a ground war with Nvidia.

You're reading the **AI Infrastructure** Newsletter

[View all newsletters](#)

While Google is one of the biggest buyers of Nvidia AI server chips, it has also been developing its own competing AI chip business that represents a threat to Nvidia's dominant market share.

Before this year, Google's tensor processing units were housed almost entirely in Google facilities and could only be rented from Google Cloud. Now Google is trying to sell those TPUs to young cloud providers, known as neoclouds, that have done only one thing so far: develop facilities to rent out Nvidia graphics processing units.

One such neocloud that Google recently tried to sell TPUs to is Nscale.

Like many neoclouds, Nscale probably wouldn't have much of a business without Nvidia's support. The chip giant has allocated part of its scarce supply of GPUs to the two-year-old company and, as of late last year, it was Nscale's largest preferred shareholder after investing capital in it. Microsoft and Google are slated to rent Nvidia GPUs from Nscale facilities in Europe in the coming year, and Nscale just bought one of the largest available sites for AI data centers in the U.S., in West Virginia.

Google's pitch to Nscale and other neoclouds on why they should buy TPUs in addition to GPUs comes down to economics.

Google told them TPUs are more efficient, reliable and stable than GPUs, especially given the various changes Nvidia has made to its chip hardware systems between different generations. Such changes have caused problems for buyers, most notably last year when they were trying to get Nvidia's Grace Blackwell systems up and running at scale, as we have reported in detail.

Plans for AI data centers that would come online next year usually center around Nvidia's next-generation Vera Rubin chip system. Nvidia chip buyers have gotten small shipments of these new GPUs, but they won't likely know the reliability of such systems until next year when they try to install large clusters.

Consistent Design

TPUs have had a more consistent design than GPUs, Google has argued, and require less complicated configurations for the networking cables that connect servers to one another so they can handle large models and process queries faster.

For its part, Nvidia said last fall in a post on X that it was “delighted by Google’s success” and also that Nvidia’s technology was “a generation ahead of the industry” and “offers greater performance, versatility, and fungibility than” custom-designed AI chips from Google and others. See our [explanation of the technical differences between GPUs and TPUs](#).

Google has discussed financially backstopping TPU data centers that neoclouds might want to develop, which would help them raise the funds to buy chips and other equipment, said two people with knowledge of the discussion. And Google also has offered to rent the TPUs back from the neoclouds so it can use them for AI development, Google Search and advertising, or rent them out to Google Cloud customers.

Nvidia also rents back billions of dollars’ worth of its GPUs from neocloud partners, but that arrangement makes less sense for its business because it has much less need for AI chips compared to Google, one of the world’s biggest AI model developers and cloud providers.

There’s no indication Nscale is moving forward with a TPU deal for one of its future data centers, and an Nscale spokesperson said “all active contracted clusters and active discussions are for GPU capacity,” not TPUs.

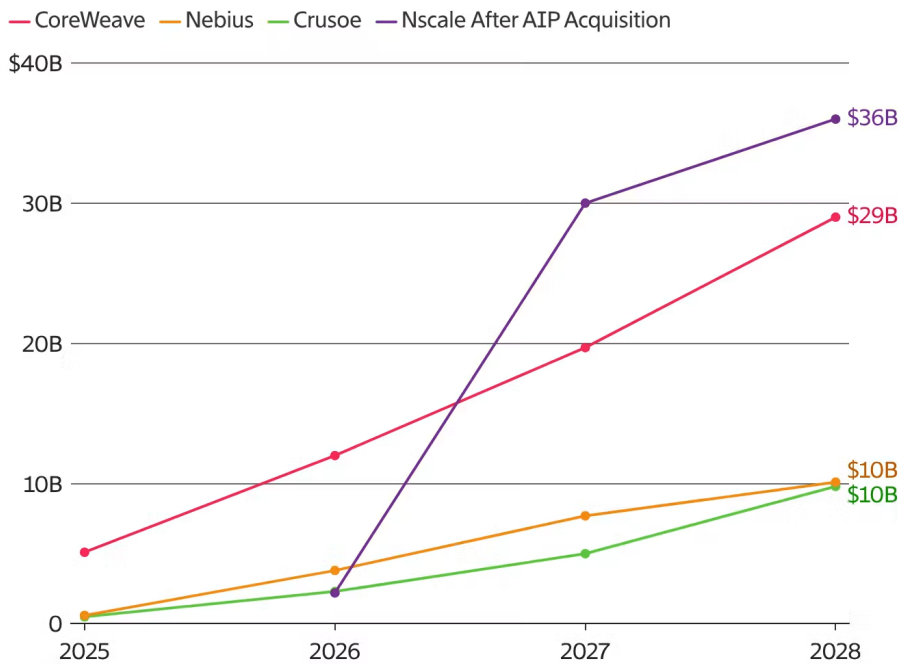
But we do know Nvidia heard about the talks with Google and has since discussed some financial incentives with the neocloud. One person close to the deal said these incentives appeared aimed at getting Nscale less interested in working with Google. The Nscale spokesperson disputed that characterization, saying “Nvidia did not offer Nscale any financial incentives in exchange for not using TPUs.” (An Nvidia spokesperson declined to comment.)

The episode reminded us of when Nvidia CEO Jensen Huang sprung to action after finding out OpenAI was talking to Google about using TPUs. Nvidia and OpenAI have since gotten closer, including through an equity investment and other financial arrangements.

Three other high-profile neoclouds backed by Nvidia—Nebius, Lambda and CoreWeave—[have said](#) they have no interest in buying TPUs anytime soon because the market for them is limited.

Neocloud Revenue Dreams

Projected revenue for upstart providers, as of March 2026



Source: The Information reporting, Goldman Sachs and FactSet

Google currently has a few major TPU users, including Anthropic (which trained its industry-leading models on the chips), Meta Platforms and Apple, as well as trading firms like Citadel Securities and startups led by former Googlers, such as Ilya Sutskever's Safe Superintelligence. Google has discussed backstopping AI data centers that large TPU customers such as Anthropic want to develop and lease themselves.

But dozens of other small neoclouds that are less under Nvidia's sway might be more receptive to the TPU pitch, because they might not want to be just another in a long line of undifferentiated GPU cloud providers.

Why Bother?

In the meantime, Google and private equity giant Blackstone recently announced the creation of a neocloud that will rent out TPUs to AI developers next year. That came after Google last year started working with Fluidstack, a neocloud that didn't have such strong Nvidia ties, about a deal to operate TPUs for Anthropic. Google has also been talking to potential finance partners to borrow money for special purpose vehicles that would buy TPUs and lease them to customers.

You might wonder why Google would go to all this trouble. We have reported that some Google managers privately said that they believe TPUs can be as big as 10% of Nvidia's AI server chip business and that huge Nvidia customers like Meta are interested in less expensive alternatives. Increasing the pool of buyers for TPUs will also improve the economics of producing them.

By getting AI developers, neoclouds or their lenders to finance TPU production for their own facilities, Google can still monetize these chips without every dollar showing up as a capital expenditure—a line item investors are increasingly scrutinizing.

One less-discussed and important factor in all of this is how much capacity Google and its TPU co-design partner, Broadcom, can get from Taiwan Semiconductor Manufacturing Co., which produces all TPUs and GPUs. Nvidia in recent years has become TSMC's biggest customer, accounting for a large share of its manufacturing capacity. Higher demand for TPUs could strengthen Google's hand in getting more of that TSMC production.

Ted Ladd, a spokesman for Google, said in a statement that his company has a “multi-decade partnership with Nvidia” and that GPUs are the “most popular chips on Google Cloud, used by our mutual customers, ranging from large enterprises like Toyota to cutting-edge startups like Thinking Machines. We believe in giving customers choice and flexibility, and many use both GPUs and TPUs for their workloads.”

Nvidia's Hefty Revenue Share

Nvidia CEO Jensen Huang is trying other ways to pull more neoclouds into the company's orbit and stave off Google's TPUs. As Phoebe reported earlier this month, Nvidia is offering to backstop GPU neoclouds by guaranteeing it would rent back GPUs from their facilities if they can't find other companies to rent them. In exchange, it would take a cut of their revenue, though we didn't know how much at the time.

We have a little more clarity now. A representative of one neocloud that has talked to Nvidia about this program said Nvidia promised to pay the neocloud a certain rate for any Vera Rubin GPU that would otherwise sit unused. But that rate is about half the rental amount the neocloud believes it can get from other customers. And if the neocloud finds customers that pay the higher rate, Nvidia's cut would be half of the difference between that higher retail price and the guaranteed rate.

Paying out that revenue share to Nvidia would effectively cut the neocloud's already slender gross margin in half.

Doing such a deal might still be worth it, this representative said. Such a backstop would help convince banks to loan money for facilities that otherwise would never get built, particularly in areas where finding an investment-grade (hyperscale) customer is more difficult. —*Valida Pau contributed to this article.*

AI Is Eating Fiber

Big AI developers like Anthropic and some of the cloud providers that serve them frequently need to move large amounts of data between data centers as they train new AI models or run existing ones. That can be an expensive headache because while petabytes of data are in transit, the GPUs that train the models may sit idle.

Lumen, one of the world's largest owners of fiber-optic cables, says it's speeding up these data transfers. Last week, Lumen bought networking software company Alkira for \$475 million. Combining Alkira's software with Lumen's fiber would allow cloud providers to program how they move data between data centers more easily without needing to reserve access to fiber-optic cables through other companies in advance, said Chris Stansbury, Lumen's chief financial officer.

The bet is part of Lumen's plan to reap rewards from the AI boom after several years of intense financial struggle.

"We've moved beyond 'Just give me pipes,'" Stansbury said.

The new offering comes as the biggest tech firms that own their own fiber networks, including Microsoft, Meta and Amazon, are buying more cables from leading fiber manufacturer Corning—possibly because their AI work is sparking more data transfers. Lumen has said it has a deal to reserve 10% of Corning's fiber production.

Separately, Nvidia, the most important supplier to these tech firms, in May announced an up to \$3.2 billion equity investment in Corning to help the company open new manufacturing facilities.—*Phoebe Liu and Laura Bratton*

New From Our Reporters