



MONEY > INVESTING

Why Google's Custom AI Chip Strategy Can't Dethrone Nvidia

By [Jon Markman](#), Contributor. ⓘ Analyzing tech stocks through the prism of... 

Follow Author

Published Apr 23, 2026, 11:47am EDT

    0

 Add Us On Google ⓘ



INDIA - 2025/07/29: In this photo illustration, the Nvidia logo is seen displayed on a smartphone with a TSMC(Taiwan Semiconductor Manufacturing Company) logo in the background. (Photo Illustration by Avishek Das/SOPA Images/LightRocket via Getty Images) ... More
SOPA IMAGES/LIGHTROCKET VIA GETTY IMAGES

Google Is Building Its Own AI Chips To Beat Nvidia. Google's \$4 Billion AI Chip Strategy Misses The Real Bottleneck.

Last week, The Information reported that [Google is in late-stage talks with Marvell Technology](#) to co-develop two new custom AI chips. One is a memory processing unit to pair with Google's existing Tensor Processing Units. The other is an entirely new TPU designed specifically for AI inference, the kind of real-time work an AI model does every time it answers a question, not the massive training runs that built it in the first place.

If the Marvell deal closes, Google will have four design partners working on its custom silicon supply chain: Broadcom, MediaTek, Marvell, and its own internal teams. Google [now expects TPU shipments to hit 4.3 million units in 2026](#), scaling to 35 million by 2028. Custom AI chip sales are projected by TrendForce to grow 45% in 2026, compared to 16% for standard GPUs. Google is offering cloud providers financial guarantees to make its TPUs the default for production AI workloads. It is launching a new software layer called TorchTPU to help developers move off the Nvidia ecosystem.

K.Mizra LLC, Ex. 2013, p. 2
Google LLC v. K.Mizra LLC
IPR2026-00304

Most coverage frames this as the moment Google decides to take Nvidia seriously. That framing is not wrong. It is just measuring the wrong battle.

The battle that actually matters is happening inside one company in Taiwan that neither Google nor Nvidia controls, and neither of them can replace.

What Inference Is, And Why It Matters Now

Training is the phase where a model learns. Massive clusters of GPUs chew through billions of data points over months. It is expensive, computationally intense, and has been where most of the AI spending went from 2022 to 2025.

MORE FOR YOU

Google May Add Play Store Alerts For Removed Android Apps



U.S. Building Ebola Treatment Center In Kenya For Exposed Americans (Live Updates)



The Rolling Stones Chart A New No. 1 Hit After Nearly 65 Years Together



Inference is what happens after the model is trained. Every time someone asks ChatGPT a question, every time Claude writes a line of code, every time a company's internal AI agent looks up a document, that is inference. Each request is small. But if you multiply it by billions of users and hundreds of enterprises, inference starts to cost more than training ever did.

K.Mizra LLC, Ex. 2013, p. 3
Google LLC v. K.Mizra LLC
IPR2026-00304

This is why the chip industry is pivoting. Training chips were built for raw power. Inference chips need to be built for cost, speed, and efficiency per query. The companies that win inference will not necessarily be the ones with the fastest chip. They will be the ones who can deliver inference at the lowest cost per million tokens, which is a different engineering problem.

Forbes Daily: Join over 1 million Forbes Daily subscribers and get our best stories, exclusive reporting and essential analysis of the day's news in your inbox every weekday.

[Sign Up](#)

By signing up, you agree to receive this newsletter, other updates about Forbes and its affiliates' offerings, our [Terms of Service](#) (including resolving disputes on an individual basis via arbitration), and you acknowledge our [Privacy Statement](#). Forbes is protected by reCAPTCHA, and the Google [Privacy Policy](#) and [Terms of Service](#) apply.

Google sees this clearly. Its Marvell deal targets exactly this gap. The memory processing unit it is co-developing is designed to fix the memory bandwidth bottleneck that slows down inference on existing TPUs. The new inference-specific TPU is a ground-up redesign for this workload. MediaTek's contribution, a chip codenamed Zebrafish, is targeting a 30% cost reduction on per-inference operations. The whole strategy is coherent and the engineering is serious.



And none of it matters if the chips cannot actually be fabricated.

The One Building That Makes All Of This Work

K.Mizra LLC, Ex. 2013, p. 4
Google LLC v. K.Mizra LLC
IPR2026-00304

Every advanced AI chip on the planet, with a handful of irrelevant exceptions, is manufactured by [Taiwan Semiconductor Manufacturing Company \(TSMC\)](#). Nvidia's Blackwells. Google's TPUs. Amazon's Trainium. Microsoft's Maia. AMD's MI series. If the chip is cutting-edge and designed for AI, it comes off a wafer in Hsinchu, Taiwan. TSMC controls over 90% of global advanced chip fabrication.

Here is what that means in practice for Google's ambitious multi-partner strategy.

In January 2026, [Nvidia surpassed Apple as TSMC's largest customer](#), a position Apple had held for more than a decade. Creative Strategies analyst Ben Bjarin projects Nvidia will generate around \$33 billion in TSMC revenue in 2026, representing roughly 22% of the foundry's total revenue. Apple is now second, at \$27 billion and 18%.

The more important number is in advanced packaging, the CoWoS process used to stitch together logic and memory into a working AI chip. [Nvidia has reserved over half of TSMC's CoWoS capacity for 2026](#), with [some reports citing over 60%](#), booking 800,000 to 850,000 wafers. TSMC itself is doubling its CoWoS capacity from roughly 35,000 wafers per month in late 2024 to a projected 130,000 per month by the end of 2026. Even with that expansion, [Nvidia's bookings extend into 2027](#).

TSMC CEO C.C. Wei has been explicit about what this means structurally:

- Building a new fab takes approximately three years
- Reaching full production capacity takes another two years
- That is a five-year minimum between deciding to expand and serving new customers at scale

Google cannot fab its way out of this. Neither can Amazon, Microsoft, or anyone else. Every AI chip designer on the planet is now queuing for the same machines, in the same buildings, operated by the same company, whose next round of capacity will not come online for years.

Why This Changes The Frame

Here is the trap most investors fall into with AI chip stories: they think the competition is between Nvidia and Google, or Nvidia and AMD, or Nvidia and Amazon's Trainium. They read "Google develops custom TPU" as "Google is coming for Nvidia's lunch."

What is actually happening is different. Google, Nvidia, Amazon, Microsoft, and every other designer is competing for slots at TSMC. The designer who gets more slots ships more chips. The

K. Mizra LLC, Ex. 2013, p. 5
Google LLC v. K. Mizra LLC

IPR2026-00304

designer who gets fewer slots ships fewer. The winner of the design race is still constrained by the gatekeeper of the fabrication race. And that gatekeeper has made one thing obvious over the last two years: it favors its biggest customer, and its biggest customer is Nvidia.

Why does Nvidia get the best slots? Three reasons:

- It writes the biggest upfront checks. Multi-billion-dollar prepayments lock in production years in advance.
- Its chips are the most complex and highest-revenue-per-wafer. TSMC earns more per square millimeter on a Blackwell than on almost anything else it makes.
- Jensen Huang built the relationship with Morris Chang, TSMC's founder, three decades ago. [Huang told Chang at the time that Nvidia would one day be TSMC's biggest customer.](#) That prediction is now reality.

Google is now trying to do what Nvidia did, but on a compressed timeline. The Marvell talks, the MediaTek partnership, the extended Broadcom agreement, all of it is Google building the same kind of structural redundancy Nvidia spent years earning. It is the right move. It is also a multi-year project, during which Nvidia will keep getting the best capacity, the fastest nodes, and the earliest access to new processes.

The investor mistake is reading "Google challenges Nvidia" as "Nvidia's moat is cracking." The actual lesson is smaller and more precise: **the moat has always been at the fab, not at the designer, and the fab has already told you who it favors.**

What To Actually Watch

The people who understand this industry stopped watching chip benchmarks a year ago. They watch allocation reports.

The numbers that matter are not clock speeds or FLOPS. They are:

- **TSMC's quarterly capacity disclosures**, which tell you how many wafers are actually rolling off the advanced nodes
- **CoWoS allocation by customer**, which tells you who is getting the packaging that turns silicon into a working AI accelerator
- **TSMC's capex guidance**, which tells you how fast the capacity wall is moving

K.Mizra LLC, Ex. 2013, p. 6
Google LLC v. K.Mizra LLC
IPR2026-00304

- **The gap between ordered and delivered wafers**, which is the real measure of supply tightness

Every AI chip story of 2026 will eventually get filtered through these numbers. Google's TPU ambitions, Amazon's Trainium scaling, Microsoft's Maia roadmap, Anthropic's rumored custom silicon, even Nvidia's Rubin architecture. None of them exist without TSMC, and TSMC's pace is set by physics, not marketing.

There is a reasonable argument to be made that the most valuable position in the entire AI stack is not the best chip designer or the best model or even the best cloud provider. It is the one company that controls the physical bottleneck everyone else has to pass through. TSMC does not need to win the AI race. It just needs to keep being the only place where the race can actually happen.

That is the real story behind Google's four-partner strategy. The design wars are real. The winners will make real money. But the fab wall is taller than any design advantage, and it is getting taller every year as the capacity expansion lags the demand curve. Google is not building a Nvidia killer. It is building insurance against the possibility that Nvidia takes an even bigger slice of a pie that neither of them controls.

The bottleneck is not in Mountain View or Santa Clara. It is in Hsinchu. Watch that building, not the benchmarks.

[Editorial Standards](#)[Reprints & Permissions](#)

Visit [Jon's website](#).

[Follow Author](#)

Join The Conversation

Comments 0

One Community. Many Voices. Create a free account to share your thoughts. Read our community guidelines [here](#).

[See All Comments \(0\)](#)

K.Mizra LLC, Ex. 2013, p. 7
Google LLC v. K.Mizra LLC
IPR2026-00304



Ad removed. [Details](#)

Forbes

© 2026 Forbes Media LLC. All Rights Reserved.

[AdChoices](#) [Privacy Statement](#)  [Your Privacy Choices](#) [Cookie Preferences](#) [Digital Terms of Sale](#) [Terms of Service](#)

[Contact Us](#) [Send Us Feedback](#) [Report a Security Issue](#) [Jobs At Forbes](#) [Reprints & Permissions](#) [Forbes Press Room](#)

[Advertise](#)

BREAKING Stocks making the biggest moves midday: MGM Resorts, Zoom Com

CREATE FREE ACCOUNT



Ferguson Home

Ceiling Fans w/Lights - Lighted Ceiling Fan - Amazing Selection & Prices

Discover stylish lighting solutions for every room in your home at Ferguson Home. Discount Prices on Ceiling Fans with Lights! Free Shipping over \$49

[Learn more >](#)

TECH

How Google makes custom chips used to train Apple AI models and its own chatbot, Gemini

PUBLISHED FRI, AUG 23, 2024, 8:00 AM EDT

CNBC PRO

YOUR FIRST LOOK AT TOMORROW,
AND LAST LOOK AT TODAY.

Stocks @ Night Newsletter

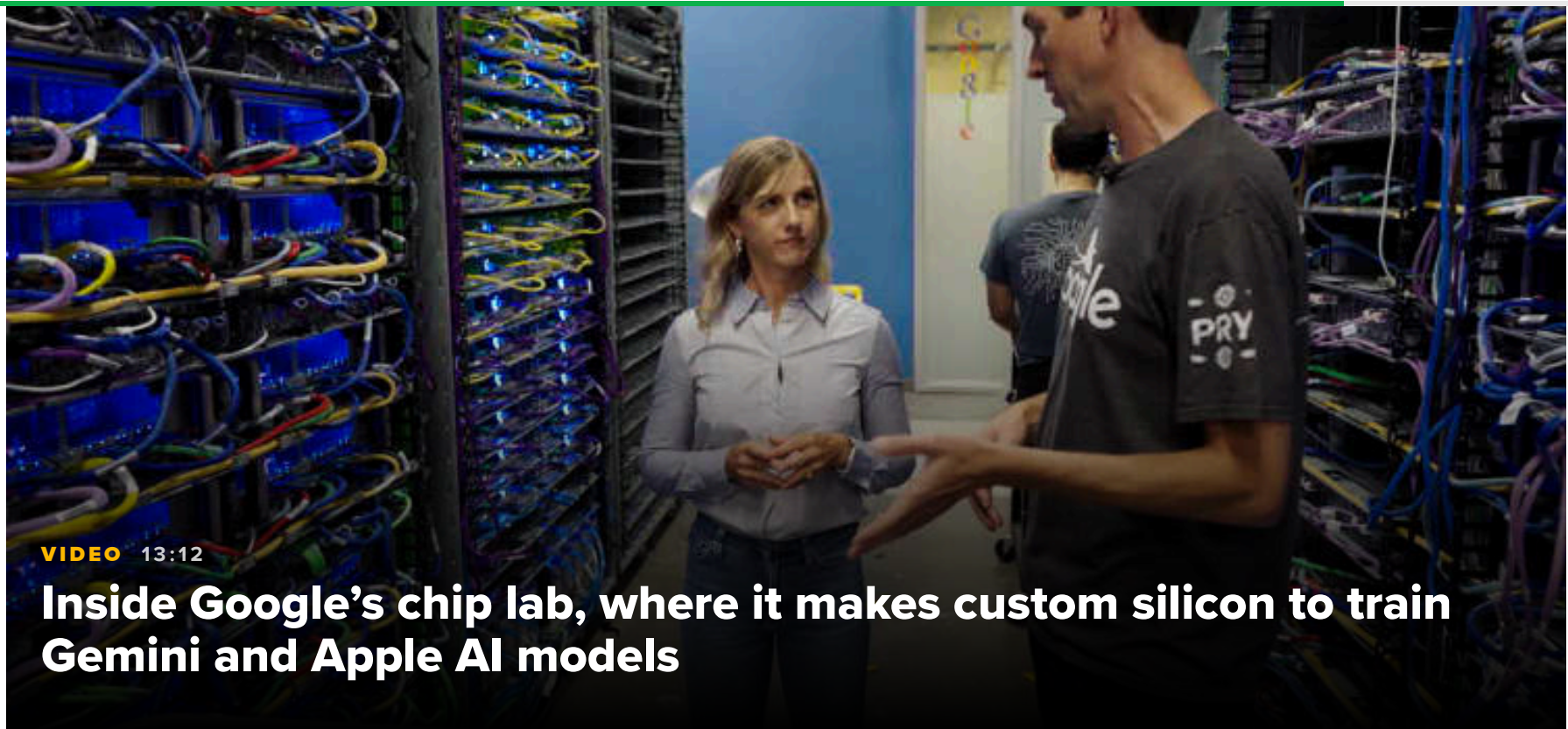
Email

JOIN FOR FREE

☐ BY SUBSCRIBING TO THIS NEWSLETTER, YOU AGREE TO OUR [TERMS OF SERVICE](#) AND [PRIVACY POLICY](#), TOGETHER WITH THIS [DISCLAIMER](#), AND AGREE TO RECEIVE GENERAL MARKETING EMAILS FROM CNBC.

stocks making the biggest moves midday: MGM Resorts, Zoom Com

CNBC



Inside a sprawling lab at [Google](#)  headquarters in Mountain View, California, hundreds of server racks hum across several aisles, performing tasks far less ubiquitous than running the world's dominant search engine or executing workloads for Google Cloud's millions of customers.

K.Mizra LLC, Ex. 2013, p. 10
Google LLC v. K.Mizra LLC
IPR2026-00304

Stocks making the biggest moves midday: MGM Resorts, Zoom, Comcast



Originally trained for internal workloads, Google's TPUs have been [available to cloud customers since 2018](#). In July, [Apple](#)  revealed it [uses TPUs](#) to train AI models underpinning Apple Intelligence. Google also relies on TPUs to train and run its Gemini chatbot.

“The world sort of has this fundamental belief that all AI, large language models, are being trained on [Nvidia](#) , and of course Nvidia has the lion's share of training volume. But Google took its own path here,” said Futurum Group CEO Daniel Newman. He's been covering Google's custom cloud chips since they launched in 2015.

Google was the first cloud provider to make custom AI chips. Three years later, [Amazon](#)  Web Services announced its [first cloud AI chip](#), Inferentia. [Microsoft](#) 's first [custom AI chip](#), Maia, wasn't announced until the end of 2023.

But being first in AI chips hasn't translated to a top spot in the overall [rat race](#) of generative AI. Google's [faced criticism](#) for [botched product releases](#), and Gemini came out more than a year after OpenAI's [ChatGPT](#).

Google Cloud, however, has gained momentum due in part to AI offerings. Google parent company [Alphabet reported](#) cloud revenue rose 29% in the most recent quarter, surpassing \$10 billion in quarterly revenues for the first time.

K.Mizra LLC, Ex. 2013, p. 11
Google LLC v. K.Mizra LLC
IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com

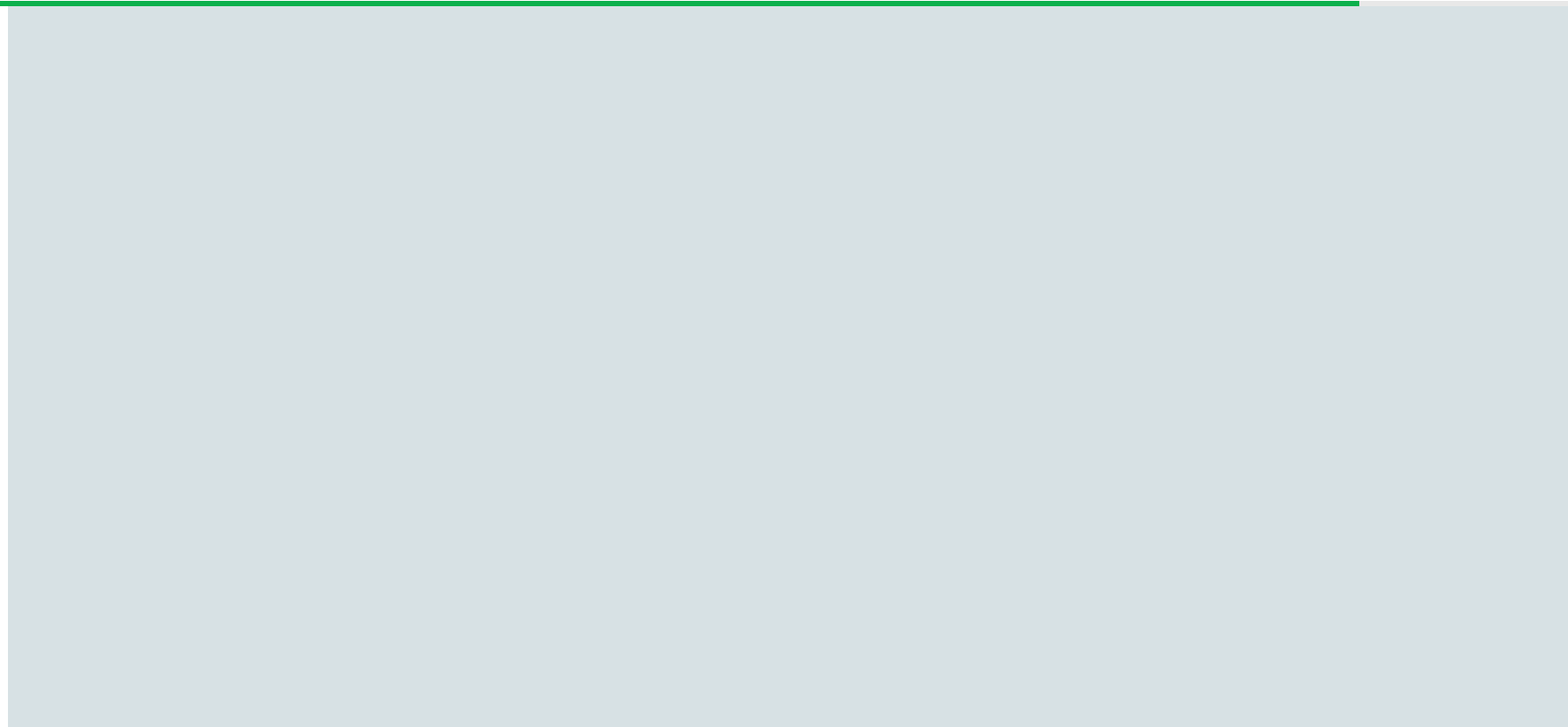


even ahead of the other two clouds for its AI prowess,” Newman said.

‘A simple but powerful thought experiment’

In July, CNBC got the first on-camera tour of Google’s chip lab and sat down with the head of custom cloud chips, Amin Vahdat. He was already at Google when it first toyed with the idea of making chips in 2014.

stocks making the biggest moves midday: MGM Resorts, Zoom Com



Amin Vahdat, VP of Machine Learning, Systems and Cloud AI at Google, holds up TPU Version 4 at Google headquarters in Mountain View, California, on July 23, 2024.

Marc Ganley

“It all started with a simple but powerful thought experiment,” Vahdat said. “A number of leads at the company asked the question: What would happen if Google users wanted to interact with Google via voice for just 30 seconds a day? And how much compute power would we need to support our users?”

K.Mizra LLC, Ex. 2013, p. 13
Google LLC v. K.Mizra LLC
IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



“We realized that we could build custom hardware, not general purpose hardware, but custom hardware — Tensor Processing Units in this case — to support that much, much more efficiently. In fact, a factor of 100 more efficiently than it would have been otherwise,” Vahdat said.

Google data centers still rely on general-purpose central processing units, or CPUs, and Nvidia’s graphics processing units, or GPUs. Google’s TPUs are a different type of chip called an application-specific integrated circuit, or ASIC, which are custom-built for specific purposes. The TPU is focused on AI. Google makes another ASIC focused on video called a Video Coding Unit.

Google also makes custom chips for its devices, similar to [Apple’s custom silicon strategy](#). The Tensor G4 powers Google’s [new AI-enabled Pixel 9](#), and its new A1 chip powers Pixel Buds Pro 2.

The TPU, however, is what set Google apart. It was the first of its kind when it launched in 2015. Google TPUs still dominate among custom cloud AI accelerators, with 58% of the market share, [according to The Futurum Group](#).

Google coined the term based on the algebraic term “tensor,” referring to the large-scale matrix multiplications that happen rapidly for advanced AI applications.

With the second TPU release in 2018, Google expanded the focus from inference to training and [made them available for its cloud customers](#) to run workloads.

K.Mizra LLC, Ex. 2013, p. 14
Google LLC v. K.Mizra LLC
IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



they've been in tight supply," said Stacy Rasgon, senior analyst covering semiconductors at Bernstein Research.

The AI boom has sent Nvidia's stock through the roof, catapulting the chipmaker to a \$3 trillion market cap in June, surpassing Alphabet and jockeying with Apple and Microsoft for position as the [world's most valuable public company](#).

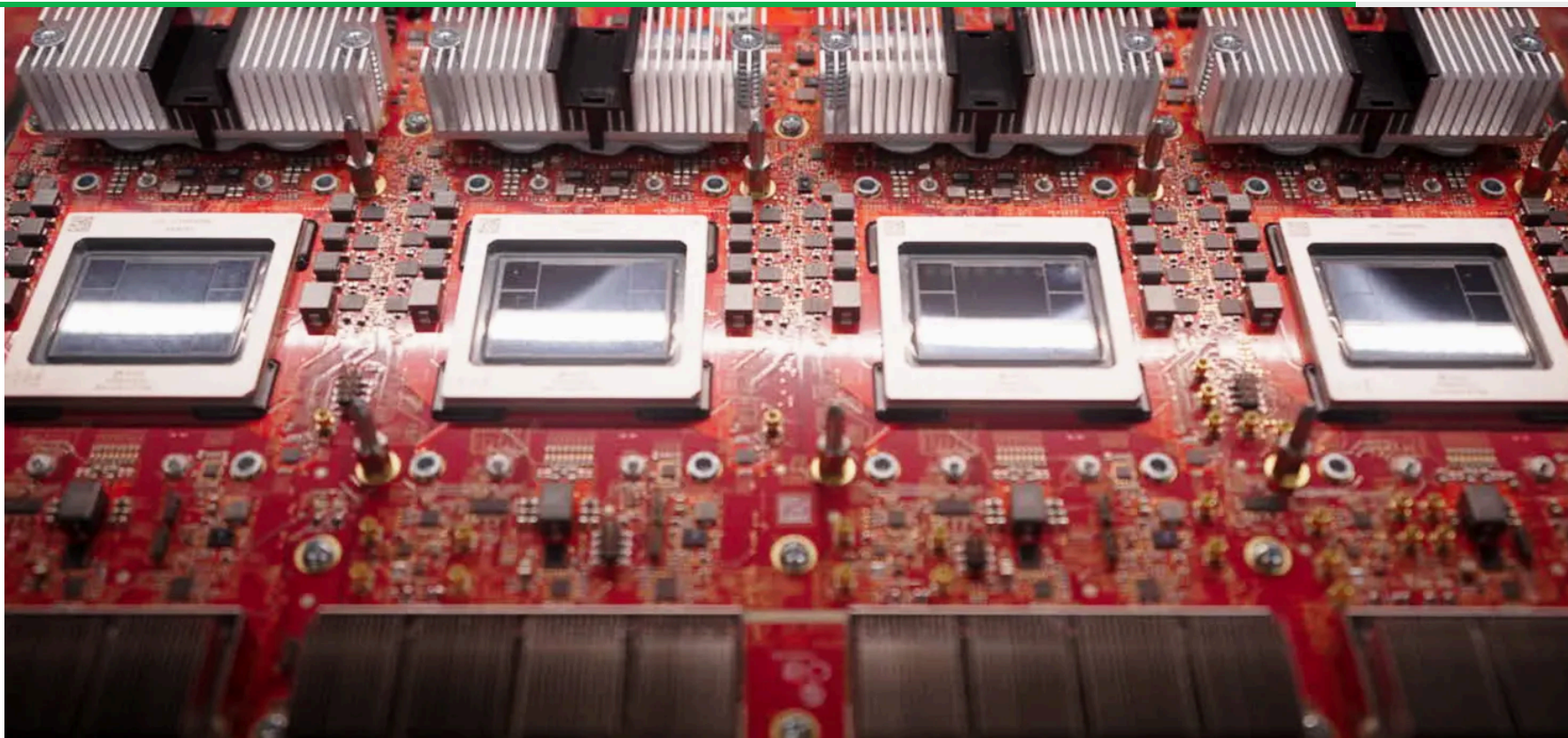
"Being candid, these specialty AI accelerators aren't nearly as flexible or as powerful as Nvidia's platform, and that is what the market is also waiting to see: Can anyone play in that space?" Newman said.

Now that we know [Apple's using Google's TPUs](#) to train its AI models, the real test will come as those full AI features roll out on iPhones and Macs [next year](#).

Broadcom and TSMC

It's no small feat to develop alternatives to Nvidia's AI engines. Google's sixth generation TPU, called Trillium, is set to come out later this year.

stocks making the biggest moves midday: MGM Resorts, Zoom Com



Google showed CNBC the sixth version of its TPU, Trillium, in Mountain View, California, on July 23, 2024. Trillium is set to come out later in 2024.

Marc Ganley

“It’s expensive. You need a lot of scale,” Rasgon said. “And so it’s not something that everybody can do. But these hyperscalers, they’ve got the scale and the money and the resources to go down that path.”

K.Mizra LLC, Ex. 2013, p. 16
Google LLC v. K.Mizra LLC
IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



make these partnerships happen.

“AI chips — they’re very complex. There’s lots of things on there. So Google brings the compute,” Rasgon said. “Broadcom does all the peripheral stuff. They do the I/O and the [SerDes](#), all of the different pieces that go around that compute. They also do the packaging.”

Then the final design is sent off for manufacturing at a fabrication plant, or fab — primarily those owned by the world’s largest chipmaker, [Taiwan Semiconductor Manufacturing Company](#) , which makes 92% of the world’s most advanced semiconductors.

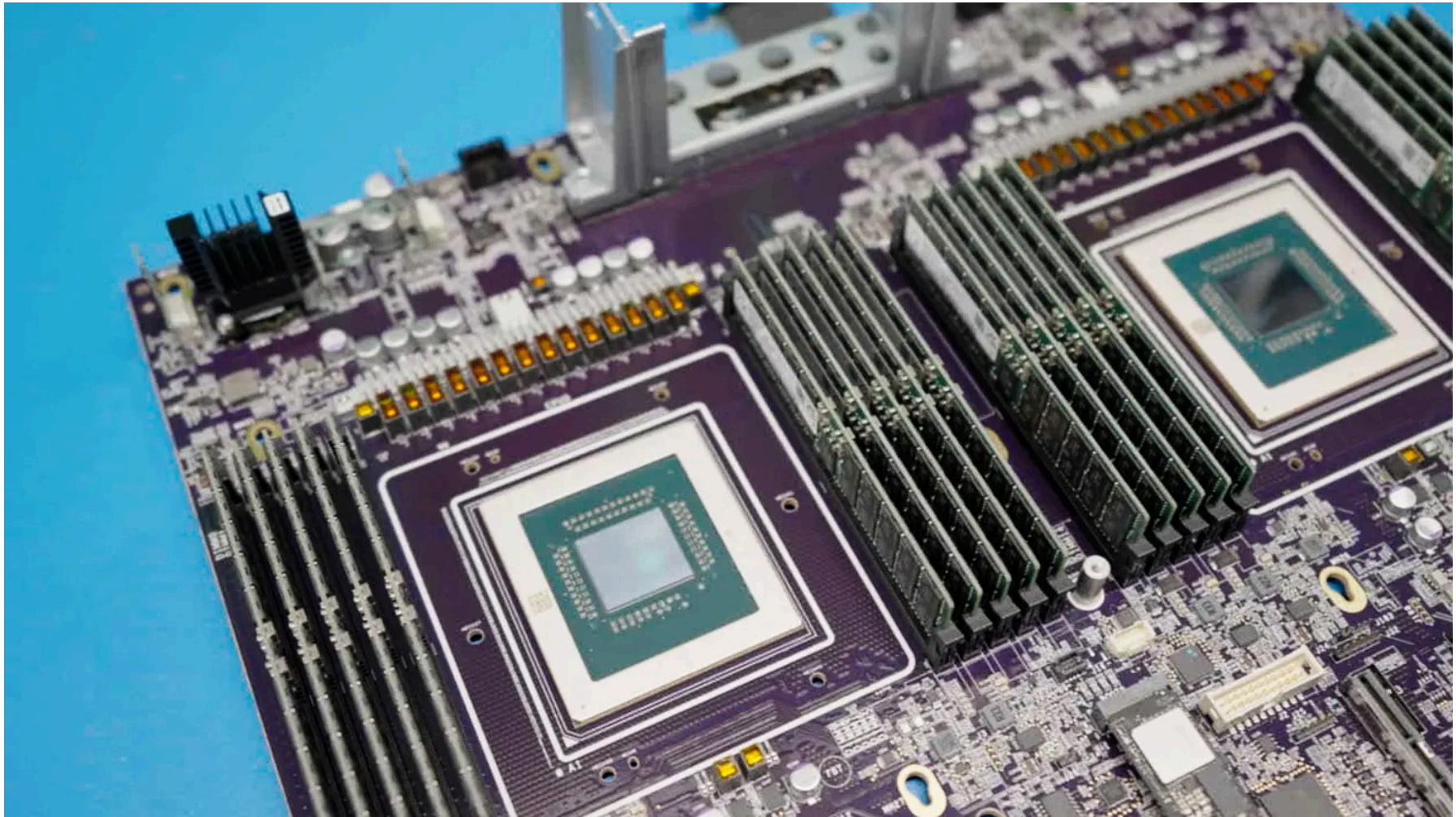
When asked if Google has any safeguards in place should the worst happen in the geopolitical sphere between China and Taiwan, Vahdat said, “It’s certainly something that we prepare for and we think about as well, but we’re hopeful that actually it’s not something that we’re going to have to trigger.”

Protecting against those risks is the primary reason the White House is handing out \$52 billion in CHIPS Act funding to companies building fabs in the U.S. — with the [biggest portions](#) going to [Intel](#), [TSMC](#), and [Samsung](#) to date.

Processors and power

K.Mizra LLC, Ex. 2013, p. 17
Google LLC v. K.Mizra LLC
IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



Google showed CNBC its new Axion CPU,

Marc Ganley

“Now we’re able to bring in that last piece of the puzzle, the CPU,” Vahdat said.

K. Mizra LLC, Ex. 2013, p. 18

“And so a lot of our internal services, whether it’s [BigQuery](#), whether it’s [Google Spreadsheets](#),

IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



[Alibaba](#) launched its server chip [in 2021](#). Microsoft announced its CPU [in November](#).

When asked why Google didn't make a CPU sooner, Vahdat said, "Our focus has been on where we can deliver the most value for our customers, and there it has been starting with the TPU, our video coding units, our networking. We really thought that the time was now."

All these processors from non-chipmakers, including Google's, are made possible by [Arm](#) chip architecture — a more customizable, power-efficient alternative that's gaining traction over the traditional x86 model from [Intel](#) and [AMD](#) . Power efficiency is crucial because, by 2027, AI servers are projected to use up [as much power every year as a country](#) like Argentina. Google's latest environmental [report](#) showed emissions [rose nearly 50%](#) from 2019 to 2023 partly due to data center growth for powering AI.

"Without having the efficiency of these chips, the numbers could have wound up in a very different place," Vahdat said. "We remain committed to actually driving these numbers in terms of carbon emissions from our infrastructure, 24/7, driving it toward zero."

It takes a massive amount of water to cool the servers that train and run AI. That's why Google's third-generation TPU started using direct-to-chip cooling, [Wired](#), 2023, p. 19

Google LLC v. K.Mizra LLC
IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



generative AI tools and making its own chips.

“I’ve never seen anything like this and no sign of it slowing down quite yet,” Vahdat said. “And hardware is going to play a really important part there.”

Choose CNBC as your preferred source on Google and never miss a moment from the most trusted name in business news.

stocks making the biggest moves midday: MGM Resorts, Zoom Com



stocks making the biggest moves midday: MGM Resorts, Zoom Com

CNBC



Potentiometers



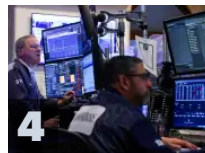
Rotary, slide, or trimmer
options with a variety of
configurations

[Learn more](#)

TRENDING NOW

K.Mizra LLC, Ex. 2013, p. 22
Google LLC v. K.Mizra LLC
IPR2026-00304

Stocks making the biggest moves midday: MGM Resorts, Zoom Communications

**Anthropic confidentially files IPO prospectus with SEC****Stocks making the biggest moves midday: MGM Resorts, Zoom Communications, Nvidia, Viasat, IBM & more****CLUB We're picking up more shares of an out-of-favor stock with a bright future****The stock market just did something eerily similar to the dot-com bubble top in 2000**

Sponsored Links

FROM THE WEB

K.Mizra LLC, Ex. 2013, p. 23
Google LLC v. K.Mizra LLC
IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



Here's How Much A 1-Day Bathroom Remodel Costs In Denver

Jacuzzi Bath Remodel

[Learn More](#)

stocks making the biggest moves midday: MGM Resorts, Zoom Com



stocks making the biggest moves midday: MGM Resorts, Zoom Com



stocks making the biggest moves midday: MGM Resorts, Zoom Com



stocks making the biggest moves midday: MGM Resorts, Zoom Com



[Subscribe to CNBC PRO](#)

[CNBC Councils](#)

[Closed Captioning](#)

[Internships](#)

[Site Map](#)

[Help](#)

[Subscribe to Investing Club](#)

[Select Personal Finance](#)

[Digital Products](#)

[Corrections](#)

[Podcasts](#)

[Contact](#)

[Licensing & Reprints](#)

[Join the CNBC Panel](#)

[News Releases](#)

[About CNBC](#)

[Careers](#)

News Tips

Got a confidential news tip? We want to hear from you.



CNBC Newsletters

Sign up for free newsletters and get more CNBC delivered to your inbox

get more CNBC
Ex. 2013, p. 28
Google LLC v. K.Mizra LLC

IPR2026-00304

stocks making the biggest moves midday: MGM Resorts, Zoom Com



Advertise With Us

PLEASE CONTACT US

[Ad Choices](#) | [Privacy Policy](#) |  [Your Privacy Choices](#) | [CA Notice](#) | [Terms of Service](#)

© 2026 Versant Media, LLC. All Rights Reserved. A Versant Media Company.

Data is a real-time snapshot *Data is delayed at least 15 minutes. Global Business and Financial News, Stock Quotes, and Market Data and Analysis.

Market Data Terms of Use and Disclaimers

Data also provided by [REFINITIV](#) 

K.Mizra LLC, Ex. 2013, p. 29
Google LLC v. K.Mizra LLC
IPR2026-00304