Nvidia's Big Tech Rivals Put Their Own A.I. Chips on the Table

Chafing at their dependence, Amazon, Google, Meta and Microsoft are racing to cut into Nvidia's dominant share of the market.

Jan. 29, 2024,

Chris Gash

In September, Amazon said it would invest up to $4 billion in

Anthropic, a San Francisco start-up working on artificial intelligence.

Soon after, an Amazon executive sent a private message to an executive at another company. He said Anthropic had won the deal because it agreed to build its A.I. using specialized computer chips designed by Amazon.

Amazon, he wrote, wanted to create a viable competitor to the chipmaker Nvidia, a key partner and kingmaker in the all-important field of artificial intelligence.

The boom in generative A.I. over the last year exposed just how dependent big tech companies had become on Nvidia. They cannot build chatbots and other A.I. systems without a special kind of chip that Nvidia has mastered over the past several years. They have spent billions of dollars on Nvidia's systems, and the chipmaker has not kept up with the demand.

So Amazon and other giants of the industry — including Google, Meta and Microsoft — are building A.I. chips of their own. With these chips, the tech giants could control their own destiny. They could rein in costs, eliminate chip shortages and eventually sell access to their chips to businesses that use their cloud services.

While Nvidia sold 2.5 million chips last year, Google spent $2 billion to $3 billion building about a million of its own A.I. chips, said Pierre Ferragu, an analyst at New Street Research. Amazon spent $200 million on 100,000 chips last year, he estimated. Microsoft said it had begun testing its first A.I. chip.

But this work is a balancing act between competing with Nvidia while working closely with the chipmaker and its increasingly powerful chief executive, Jensen Huang.

Mr. Huang’s company accounts for more than 70 percent of A.I. chip sales, according to the research firm Omdia. It supplies an even larger percentage of the systems used in the creation of generative A.I. Nvidia’s sales have shot up 206 percent over the past year, and the company has added about a trillion dollars in market value.

What’s revenue to Nvidia is a cost for the tech giants. Orders from Microsoft and Meta made up about a quarter of Nvidia’s sales in the past two full quarters, said Gil Luria, an analyst at the investment bank D.A. Davidson.

Nvidia sells its chips for about $15,000 each, while Google spends an average of just $2,000 to $3,000 on each of its own, according to Mr. Ferragu.

“When they encountered a vendor that held them over a barrel, they reacted very strongly,” Mr. Luria said.

Companies constantly court Mr. Huang, jockeying to be at the front of the line for his chips. He regularly appears on event stages with their chief executives, and the companies are quick to say they remain committed to their partnerships with Nvidia. They all plan to keep offering its chips alongside their own.

While the big tech companies are moving into Nvidia’s business, it is moving into theirs. Last year, Nvidia started its own cloud service where businesses can use its chips, and it is funneling chips into a new wave of cloud providers, such as CoreWeave, that compete with the big three: Amazon, Google and Microsoft.

“The tensions here are a thousand times the usual jockeying between customers and suppliers,” said Charles Fitzgerald, a technology consultant and investor.

Nvidia declined to comment.

The A.I. chip market is projected to more than double by 2027, to roughly $140 billion, according to the research firm Gartner. Venerable chipmakers like AMD and Intel are also building specialized A.I. chips, as are start-ups such as Cerebras and SambaNova. But Amazon and other tech giants can do things that smaller competitors cannot.

“In theory, if they can reach a high enough volume and they can get their costs down, these companies should be able to provide something that is even better than Nvidia," said Naveen Rao, who founded one of the first A.I. chip start-ups and later sold it to Intel.

Nvidia builds what are called graphics processing units, or G.P.U.s, which it originally designed to help render images for video games. But a decade ago, academic researchers realized these chips were also really good at building the systems, called neural networks, that now drive generative A.I.

As this technology took off, Mr. Huang quickly began modifying Nvidia's chips and related software for A.I., and they became the de facto standard. Most software systems used to train A.I. technologies were tailored to work with Nvidia's chips.

"Nvidia's got great chips, and more importantly, they have an incredible ecosystem," said Dave Brown, who runs Amazon's chip efforts. That makes getting customers to use a new kind of A.I. chip "very, very challenging," he said.

Rewriting software code to use a new chip is so difficult and time-consuming, many companies don't even try, said Mike Schroepfer, an adviser and former chief technology officer at Meta. "The problem with technological development is that so much of it dies before it even gets started," he said.

Rani Borkar, who oversees Microsoft's hardware infrastructure, said Microsoft and its peers needed to make it "seamless" for customers to move between chips from different companies.

Amazon, Mr. Brown said, is working to make switching between chips "as simple as it can possibly be."

Some tech giants have found success making their own chips. Apple designs the silicon in iPhones and Macs, and Amazon has deployed more than two million of its own traditional server chips in its cloud computing data centers. But achievements like these take years of hardware and software development.

Google has the biggest head start in developing A.I. chips. In 2017, it introduced its tensor processing unit, or T.P.U., named after a kind of calculation vital to building artificial intelligence. Google used tens of thousands of T.P.U.s to build A.I. products, including its online chatbot, Google Bard. And other companies have used the chip through Google's cloud service to build similar technologies, including the high-profile start-up Cohere.

Amazon is now on the second generation of Trainium, its chip for building A.I. systems, and has a second chip made just for serving up A.I. models to customers. In May, Meta announced plans to work on an A.I. chip tailored to its needs, though it is not yet in use. In November, Microsoft announced its first A.I. chip, Maia, which will focus initially on running Microsoft's own A.I. products.

"If Microsoft builds its own chips, it builds exactly what it needs for the lowest possible cost," Mr. Luria said.

Nvidia's rivals have used their investments in high-profile A.I. start-ups to fuel use of their chips. Microsoft has committed $13 billion to OpenAI, the maker of the ChatGPT chatbot, and its Maia chip will serve OpenAI's technologies to Microsoft's customers. Like Amazon, Google has invested billions in Anthropic, and it is using Google's A.I. chips, too.

Anthropic, which has used chips from both Nvidia and Google, is among a handful of companies working to build A.I. using as many specialized chips as they can get their hands on. Amazon said that if companies like Anthropic used Amazon's chips on an increasingly large scale and even helped design future chips, doing so could reduce the cost and improve the performance of these processors. Anthropic declined to comment.

But none of these companies will overtake Nvidia anytime soon. Its chips may be pricey, but are among the fastest on the market. And the company will continue to improve their speed.

Mr. Rao said his company, Databricks, trained some experimental A.I. systems using Amazon's A.I. chips, but built its largest and most important systems using Nvidia chips because they provided higher performance and played nicely with a wider range of software.

"We have many years of hard innovation ahead of us," Amazon's Mr. Brown said. "Nvidia is not going to be standing still."

Cade Metz writes about artificial intelligence, driverless cars, robotics, virtual reality and other emerging areas of technology. More about Cade Metz

Karen Weise writes about technology and is based in Seattle. Her coverage focuses on Amazon and Microsoft, two of the most powerful companies in America. More about Karen Weise

Explore Our Coverage of Artificial Intelligence

- The F.T.C. opened an inquiry into the multibillion-dollar investments by Microsoft, Amazon and Google in the A.I. start-ups OpenAI and Anthropic, broadening efforts to regulate the power the tech giants can have over the new technology.

- OpenAI said that it was opening an app store for people to share customized versions of its popular chatbot, ChatGPT, as the company works to expand the reach of its flagship technology and turn it into a moneymaker.

- A U.S. congressional committee has asked the Commerce Department to look into whether G42, a giant A.I. company that is controlled by the ruling family of the United Arab Emirates, should be put under trade restrictions because of its ties to China.

- Tools powered by A.I. can create lifelike images of people who do not exist. Can you identify which of these images are real people and which are A.I.-generated?

- In 2024, A.I. is set to advance at a rapid rate, becoming more powerful and spreading into the physical world. Here is what to expect.

- In the hands of anonymous internet users, A.I. tools can create waves of harassing and racist material. It's already happening on the anonymous message board 4chan.

- While they lag behind their U.S. counterparts, South Korean firms' focus on non-English languages could help loosen the American grip on A.I.

- At the Musée D’Orsay in Paris, an A.I. doppelgänger of Vincent van Gogh chats with visitors, offering insights into his own life and death. We asked him some questions.