How Nvidia Built a Competitive Moat Around A.I. Chips

The most visible winner of the artificial intelligence boom achieved its dominance by becoming a one-stop shop for A.I. development, from chips to software to other services.

Jensen Huang, Nvidia’s chief executive, at a conference in August. “The thing that we understood is that this is a reinvention of how computing is done,” he said in an interview.Philip Cheung for The New York Times

Naveen Rao, a neuroscientist turned tech entrepreneur, once tried to compete with Nvidia, the world's leading maker of chips tailored for artificial intelligence.

At a start-up that the semiconductor giant Intel later bought, Mr. Rao worked on chips intended to replace Nvidia's graphics processing units, which are components adapted for A.I. tasks like machine learning. But while Intel moved slowly, Nvidia swiftly upgraded its products with new A.I. features that countered what he was developing, Mr. Rao said.

After leaving Intel and leading a software start-up, MosaicML, Mr. Rao used Nvidia's chips and evaluated them against those from rivals. He found that Nvidia had differentiated itself beyond the chips by creating a large community of A.I. programmers who consistently invent using the company's technology.

"Everybody builds on Nvidia first," Mr. Rao said. "If you come out with a new piece of hardware, you're racing to catch up."

Over more than 10 years, Nvidia has built a nearly impregnable lead in producing chips that can perform complex A.I. tasks like image, facial and speech recognition, as well as generating text for chatbots like ChatGPT. The onetime industry upstart achieved that dominance by recognizing the A.I. trend early, tailoring its chips to those tasks and then developing key pieces of software that aid in A.I. development.

Jensen Huang, Nvidia’s co-founder and chief executive, has since kept raising the bar. To maintain its leading position, his company has also offered customers access to specialized computers, computing services and other tools of their emerging trade. That has turned Nvidia, for all intents and purposes, into a one-stop shop for A.I. development.

While Google, Amazon, Meta, IBM and others have also produced A.I. chips, Nvidia today accounts for more than 70 percent of A.I. chip sales and holds an even bigger position in training generative A.I. models, according to the research firm Omdia.

In May, the company’s status as the most visible winner of the A.I. revolution became clear when it projected a 64 percent leap in quarterly revenue, far more than Wall Street had expected. On Wednesday, Nvidia — which has surged past $1 trillion in market capitalization to become the world’s most valuable chip maker — is expected to confirm those record results and provide more signals about booming A.I. demand.

“Customers will wait 18 months to buy an Nvidia system rather than buy an available, off-the-shelf chip from either a start-up or another competitor," said Daniel Newman, an analyst at Futurum Group. "It's incredible."

Mr. Huang helped found Nvidia in 1993.Philip Cheung for The New York Times

Mr. Huang, 60, who is known for a trademark black leather jacket, talked up A.I. for years before becoming one of the movement's best-known faces. He has publicly said computing is going through its biggest shift since IBM defined how most systems and software operate 60 years ago. Now, he said, GPUs and other special-purpose chips are replacing standard microprocessors, and A.I. chatbots are replacing complex software coding.

"The thing that we understood is that this is a reinvention of how computing is done," Mr. Huang said in an interview. "And we built everything from the ground up, from the processor all the way up to the end."

Mr. Huang helped start Nvidia in 1993 to make chips that render images in video games. While standard microprocessors excel at performing complex calculations sequentially, the company's GPUs do many simple tasks at once.

In 2006, Mr. Huang took that further. He announced software technology called CUDA, which helped program the GPUs for new tasks, turning them from single-purpose chips to more general-purpose ones that could take on other jobs in fields like physics and chemical simulations.

A big breakthrough came in 2012 when researchers used GPUs to achieve humanlike accuracy in tasks such as recognizing a cat in an image — a precursor to recent developments like generating images from text prompts.

Nvidia responded by turning “every aspect of our company to advance this new field,” Mr. Huang recently said in a commencement speech at National Taiwan University.

The effort, which the company estimated has cost more than $30 billion over a decade, made Nvidia more than a component supplier. Besides collaborating with leading scientists and start-ups, the company built a team that directly participates in A.I. activities like creating and training language models.

Advance warning about what A.I. practitioners need led Nvidia to develop many layers of key software beyond CUDA. Those included hundreds of prebuilt pieces of code, called libraries, that save labor for programmers.

In hardware, Nvidia gained a reputation for consistently delivering faster chips every couple of years. In 2017, it started tweaking GPUs to handle specific A.I. calculations.

That same year, Nvidia, which typically sold chips or circuit boards for other companies' systems, also began selling complete computers to carry out A.I. tasks more efficiently. Some of its systems are now the size of supercomputers, which it assembles and operates using proprietary networking technology and thousands of GPUs. Such hardware may run weeks to train the latest A.I. models.

"This type of computing doesn't allow for you to just build a chip and customers use it," Mr. Huang said in the interview. "You've got to build the whole data center."

Last September, Nvidia announced the production of new chips named H100, which it enhanced to handle so-called transformer operations. Such calculations turned out to be the foundation for services like ChatGPT, which have prompted what Mr. Huang calls the "iPhone moment" of generative A.I.

To further extend its influence, Nvidia has also recently forged partnerships with big tech companies and invested in high-profile A.I. start-ups that use its chips. One was Inflection AI, which in June announced $1.3 billion in funding from Nvidia and others. The money was used to help finance the purchase of 22,000 H100 chips.

Mustafa Suleyman, Inflection's chief executive, said that there was no obligation to use Nvidia's products but that competitors offered no viable alternative. "None of them come close," he said.

Nvidia has also directed cash and scarce H100s lately to upstart cloud services, such as CoreWeave, that allow companies to rent time on computers rather than buying their own. CoreWeave, which will operate Inflection's hardware and owns more than 45,000 Nvidia chips, raised $2.3 billion in debt this month to help buy more.

Given the demand for its chips, Nvidia must decide who gets how many of them. That power makes some tech executives uneasy.

"It's really important that hardware doesn't become a bottleneck for A.I. or gatekeeper for A.I.," said Clément Delangue, chief executive of Hugging Face, an online repository for language models that collaborates with Nvidia and its competitors.

Some rivals said it was tough to compete with a company that sold computers, software, cloud services and trained A.I. models, as well as processors.

"Unlike any other chip company, they have been willing to openly compete with their customers," said Andrew Feldman, chief executive of Cerebras, a start-up that develops A.I. chips.

But few customers are complaining, at least publicly. Even

Google, which began creating competing A.I. chips more than a decade ago, relies on Nvidia’s GPUs for some of its work.

Demand for Google’s own chips is “tremendous,” said Amin Vahdat, a Google vice president and general manager of compute infrastructure. But, he added, “we work really closely with Nvidia.”

Nvidia doesn’t discuss prices or chip allocation policies, but industry executives and analysts said each H100 costs $15,000 to more than $40,000, depending on packaging and other factors — roughly two to three times more than the predecessor A100 chip.

Pricing “is one place where Nvidia has left a lot of room for other folks to compete,” said David Brown, a vice president at Amazon’s cloud unit, arguing that its own A.I. chips are a bargain compared with the Nvidia chips it also uses.

Mr. Huang said his chips’ greater performance saved customers money. “If you can reduce the time of training to half on a $5 billion data center, the savings is more than the cost of all of the chips,” he said. “We are the lowest-cost solution in the world.”

He has also started promoting a new product, Grace Hopper, which combines GPUs with internally developed microprocessors, countering chips that rivals say use much less energy for running A.I. services.

Still, more competition seems inevitable. One of the most promising entrants in the race is a GPU sold by Advanced Micro Devices, said Mr. Rao, whose start-up was recently purchased by the data and A.I. company DataBricks.

“No matter how anybody wants to say it’s all done, it’s not all done,” Lisa Su, AMD’s chief executive, said.

Cade Metz contributed reporting.

Audio produced by Tally Abecassis.