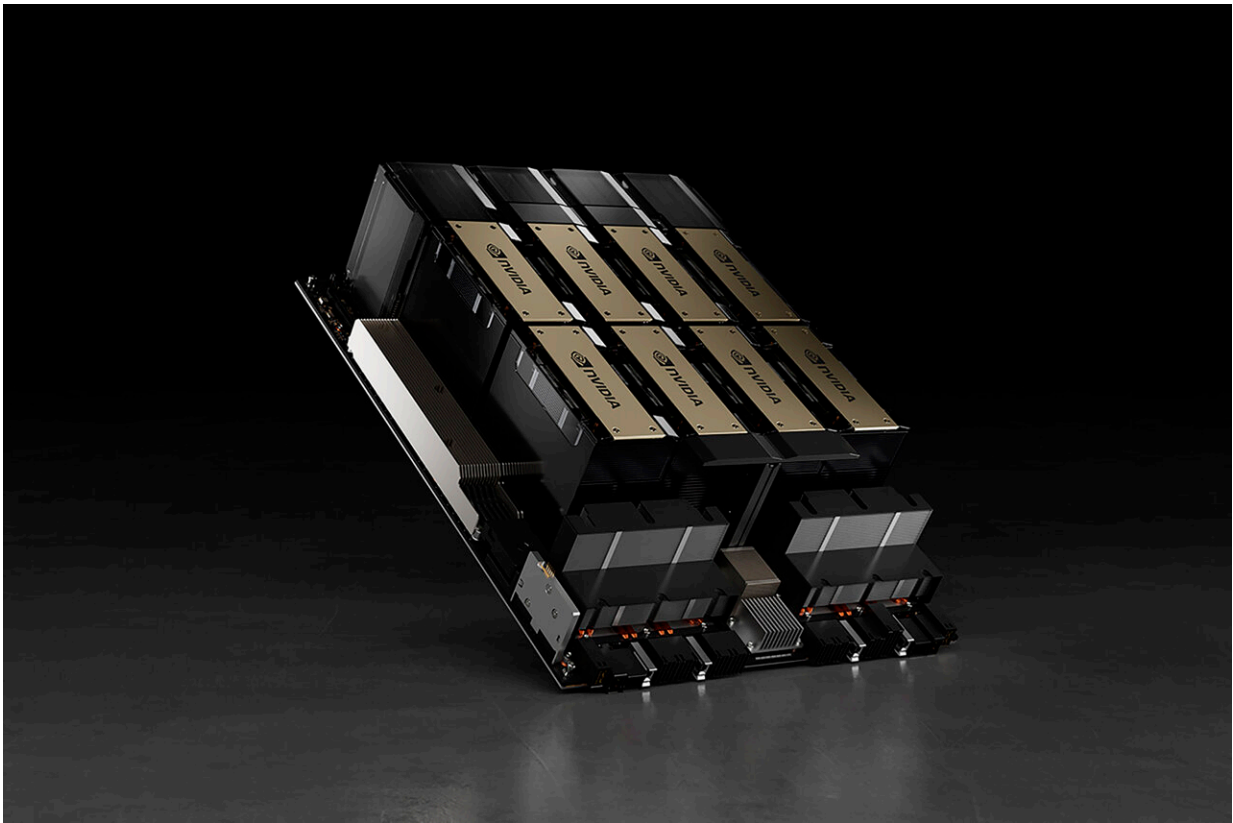


AI chips are hot. Here's what they are, what they're for and why investors see gold

June 1 2023, by David Hamilton



This photo provided by the chipmaker Nvidia shows the company's HGX H100 module, which can use as many as eight AI chips to train artificial intelligence systems and perform other tasks. Such AI chips are tiny slivers of silicon designed to simplify and speed up the development of artificial intelligence systems such as ChatGPT, making them one of the hottest items in technology. Credit: Nvidia via AP

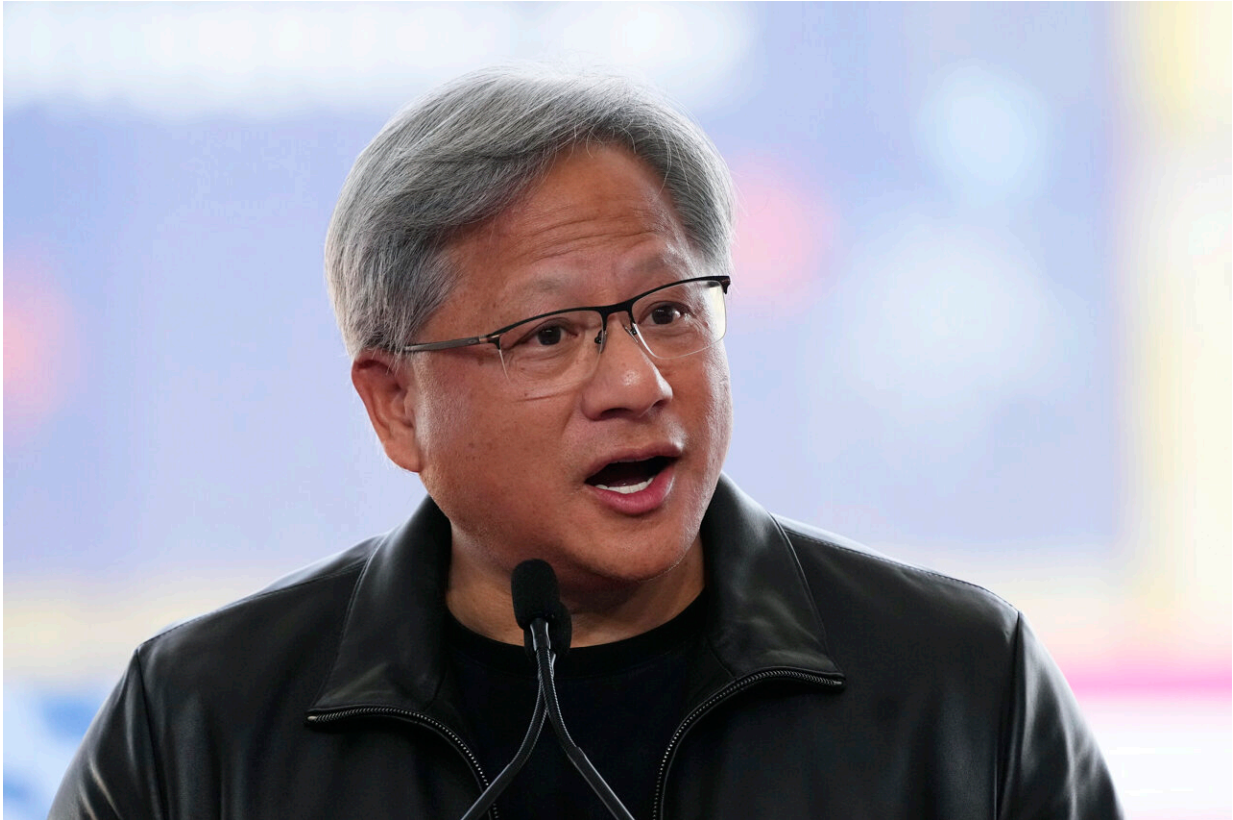
The hottest thing in technology is an unprepossessing sliver of silicon closely related to the chips that power video game graphics. It's an artificial intelligence chip, designed specifically to make building AI systems such as ChatGPT faster and cheaper.

Such chips have suddenly taken center stage in what some experts consider an AI revolution that could reshape the technology sector—and possibly the world along with it. Shares of Nvidia, the leading designer of AI chips, [rocketed up almost 25% last Thursday](#) after the company forecast a huge jump in revenue that analysts said indicated soaring sales of its products. The company was briefly worth more than \$1 trillion on Tuesday.

SO WHAT ARE AI CHIPS, ANYWAY?

That isn't an easy question to answer. "There really isn't a completely agreed upon definition of AI chips," said Hannah Dohmen, a [research analyst](#) with the Center for Security and Emerging Technology.

In general, though, the term encompasses computing hardware that's specialized to handle AI workloads—for instance, by "training" AI systems to tackle difficult problems that can choke conventional computers.



Nvidia Co-founder, President and CEO Jensen Huang speaks at the Taiwan Semiconductor Manufacturing Company facility under construction in Phoenix, Dec. 6, 2022. AI chips and their leading designer, Nvidia, are now at the center of what some experts consider an AI revolution that could reshape the technology sector — and possibly the world along with it. Credit: AP Photo/Ross D. Franklin, File

VIDEO GAME ORIGINS

Three entrepreneurs founded Nvidia in 1993 to push the boundaries of computational graphics. Within a few years, the company had developed a new [chip](#) called a [graphics processing unit](#), or GPU, which dramatically sped up both development and play of video games by performing multiple complex graphics calculations at once.

That technique, known formally as [parallel processing](#), would prove key to the development of both games and AI. Two graduate students at the University of Toronto used a GPU-based [neural network](#) to win a prestigious 2012 AI competition called ImageNet by identifying photo images at much lower error rates than competitors.

The win kick-started interest in AI-related parallel processing, opening a new business opportunity for Nvidia and its rivals while providing researchers powerful tools for exploring the frontiers of AI development.

MODERN AI CHIPS

Eleven years later, Nvidia is the dominant supplier of chips for building and updating AI systems. One of its recent products, the H100 GPU, packs in [80 billion transistors](#)—about 13 million more than Apple's [latest high-end processor](#) for its MacBook Pro laptop. Unsurprisingly, this technology isn't cheap; at one online retailer, [the H100 lists for \\$30,000](#).



A Nvidia Corporation sign is shown in Santa Clara, Calif., Wednesday, May 31, 2023. AI chips and their leading designer, Nvidia, are now at the center of what some experts consider an AI revolution that could reshape the technology sector — and possibly the world along with it. Credit: AP Photo/Jeff Chiu

Nvidia doesn't fabricate these complex GPU chips itself, a task that would require enormous investments in new factories. Instead it relies on Asian chip foundries such as Taiwan Semiconductor Manufacturing Co. and Korea's Samsung Electronics.

Some of the biggest customers for AI chips are cloud-computing services such as those run by Amazon and Microsoft. By renting out their AI computing power, those services make it possible for smaller companies and groups that couldn't afford to build their own AI systems

from scratch to use cloud-based tools to help with tasks that can range from drug discovery to customer management.

OTHER USES AND COMPETITION

Parallel processing has many uses outside of AI. A few years ago, for instance, Nvidia graphics cards were in short supply because cryptocurrency miners, who set up banks of computers to solve thorny mathematical problems for bitcoin rewards, had snapped up most of them. That problem faded as the cryptocurrency market collapsed in early 2022.

Analysts say Nvidia will inevitably face tougher competition. One potential rival is Advanced Micro Devices, which already faces off with Nvidia in the market for computer graphics chips. AMD has recently taken steps to bolster its own lineup of AI chips.

Nvidia is based in Santa Clara, California. Co-founder Jensen Huang remains the company's president and chief executive.

© 2023 The Associated Press. All rights reserved. This material may not be published, broadcast, rewritten or redistributed without permission.

Citation: AI chips are hot. Here's what they are, what they're for and why investors see gold (2023, June 1) retrieved 4 May 2024 from <https://techxplore.com/news/2023-06-ai-chips-hot-theyre-investors.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.
