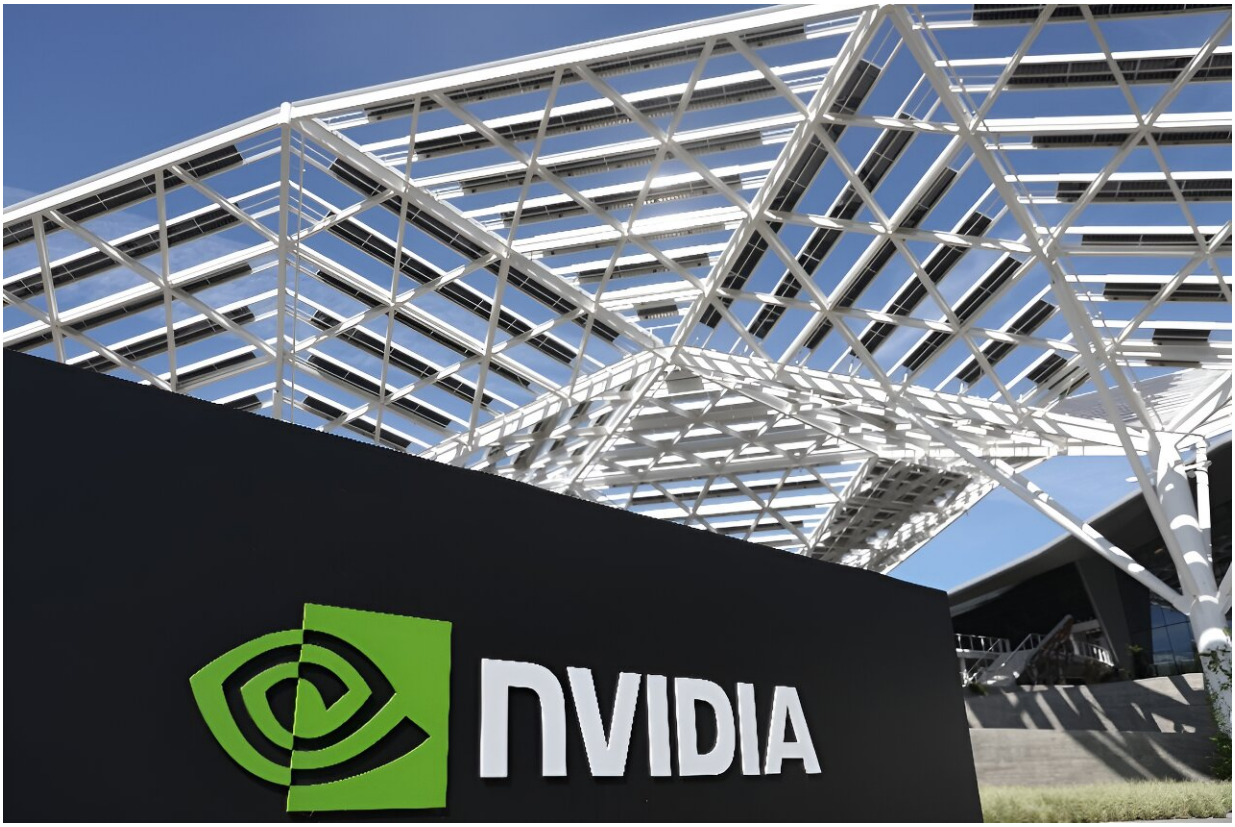
# AI chip crunch: Startups vie for Nvidia's vital component

September 28 2023, by Alex PIGMAN with Julie JAMMOT in San Francisco




Nvidia's CEO and founder Jensen Huang made a wild bet years ago that the world would soon clamor for a powerful chip usually used for making video games, but that could build AI as well.

The artificial intelligence revolution is fully underway, but soaring

demand for its most crucial component has startups scratching their heads on how they can deliver on AI's promise.

Generative AI's lifeblood is a book-sized semiconductor known as the graphics processing unit (GPU)—built by one company, Nvidia.

Nvidia's CEO and founder Jensen Huang made a wild bet years ago that the world would soon clamor for a powerful chip usually used for making video games, but that could build AI as well.

No company working with the generative AI models that fuel today's frenzy can get off the ground without Nvidia's singular product: the latest model is the H100 and its accompanying software.

That painful reality is one that Amazon, Intel, AMD and others are scrambling to fix with their own alternatives, but those attempts could take years.

## 'Not a lot of GPUs'

And with the biggest tech companies throwing all their financial might into generative AI, the smaller fish must go on the hunt to secure Nvidia's holy grail.

"Around the world, it is becoming very hard to get thousands of GPUs because all these big companies are putting in billions of dollars, stockpiling GPUs," said Fangbo Tao, co-founder of Mindverse.AI, a Singapore-based startup.

"There's not a lot of GPUs around," he said.

Nvidia's CEO Jensen Huang made a wild bet years ago that the world would soon clamor for a powerful chip usually used for making video games, but that could build AI as well.

Tao spoke to AFP at the TechCrunch Disrupt conference in San Francisco, where AI startups jostled to make their pitches to Silicon Valley's venture capitalists (VC).

ChatGPT took the world by storm just as Silicon Valley was caught in a nasty hangover from the pandemic when investors threw money at startups, convinced that life had gone irreversibly online.

That turned out to be far-fetched, and the US tech scene entered a downturn with rounds of layoffs and VC money dried up.

Thanks to AI, some of the old mojo is back, and anyone with those two letters on their resume will likely see a red carpet rolled out on the legendary Sand Hill Road, home to Silicon Valley's most storied investors.

But as the startups walk away with their VC cash, the money in their pockets will be quickly forked out to Nvidia for GPUs either directly or through providers to bring their AI dreams to execution.

"We call on a lot of the big cloud providers (Microsoft, AWS and Google) ), and they all tell us even they are having trouble getting supplies," said Laurent Daudet, CEO of AI startup LightOn.

The problem is most acute for companies involved in training generative AI models, which requires that power hungry GPUs work at peak capacity to process troves of data ingested from the internet.

The computing needs are so massive that only a few companies can stump up the cash to build one of these state-of-the-art large language models.


Companies on the AI frontlines point out that Nvidia's primordial role makes it the de-facto kingmaker on where the technology is going.

## 'Sucking the oxygen'

The ten billion dollars investment by Microsoft into OpenAI is widely understood to be paid out as credits to access purpose-built data centers humming with Nvidia GPUs.

Google has built its own models and now Amazon on Monday said it was pumping four billion dollars into Anthropic AI, another company that trains AI.

Training on that mountain of data "is sucking out almost all the oxygen

from the GPU market right now," said Said Ouissal, CEO of Zededa, a company that works on making AI less power hungry.

"You're looking at mid-next year, maybe late next year before you're actually going to get delivery on new orders. The shortage doesn't seem to be letting up," added Wes Cummins, CEO Applied Digital, a company that supplies AI infrastructure.

Companies on the AI frontlines also point out that Nvidia's primordial role makes it the de-facto kingmaker on where the technology is going.

The market is "almost entirely driven by the big players—Googles, Amazons, Metas" that have the "enormous amounts of data and enormous amounts of capital," former Nvidia engineer Jacopo Pantaleoni told The Information.

"This was not the world I wanted to help build," he said.

Some veterans of Silicon Valley said that the frenzied days of Nvidia GPUs will not last forever and that other options will inevitably emerge.

Or the cost of entry will prove too high, even for the giants, bringing the current boom down to earth.



Citation: AI chip crunch: Startups vie for Nvidia's vital component (2023, September 28) retrieved 2 May 2024 from https://techxplore.com/news/2023-09-ai-chip-crunch-startups-vie.html