APR 12
Jacob Bourne

Big cloud server shortage could slow generative AI's breakneck pace

Article

The news: The unanticipated demand to create AI software has outstripped the cloud service capacities of **Amazon Web Services** (AWS), **Microsoft**, **Google Cloud**, and **Oracle**, per The Information.

- Customers are reporting monthslong wait times to rent GPU hardware for AI-building as big cloud limits access over a chip supply crunch.
- **Nvidia**—the go-to GPU chip supplier for advanced AI model training—is **two to three months behind on new order fulfillment for cloud server chips**.
- It's not just chips creating the bottleneck. Generative AI's **steep energy requirements** **are butting up against power supplies in some regions**.

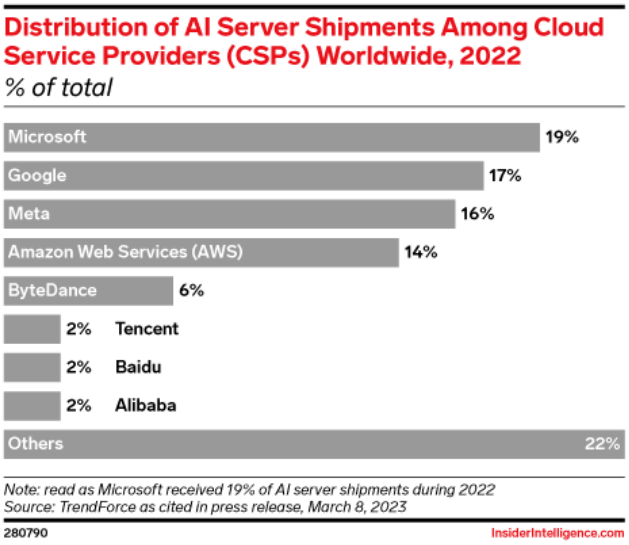
Unequal access: **Bank of America's idea that Nvidia represents "a democratizing opportunity" for generative AI is at odds with tech industry reality.**

- **Twitter** CEO **Elon Musk**, who recently spent tens of millions of dollars on a purchase of 10,000 GPUs for the platform, has pockets deep enough to quickly pivot into the generative AI business.
- Nvidia's newly released **H100** chip could ease the server shortage, but with its price likely significantly higher than the **A100**, it won't level the generative AI playing field.

Startups have to depend on cloud providers who can afford to stockpile Nvidia's pricey A100 chips to deliver the compute power necessary for generative AI. If cloud access were equitable, the dependence wouldn't be a problem, but with some startups like **OpenAI** getting priority status, a generative AI monopoly is likely.

Key takeaways: The cloud crunch could **temporarily slow the generative AI market's breakneck pace, potentially giving the US government a chance to catch** up on regulatory efforts. However, the delays won't amount to the six-month pause that some industry players want.

- Chip supplies will improve, but a global energy crisis should put the enterprise on guard for possible loss of cloud access to AI tools this summer.
- Nvidia's inability to meet near-term chip demand could boost GPU sales for rivals like **Intel** and **AMD**.
- We might see Google, Microsoft, and Amazon devote more resources to in-house AI hardware to reduce reliance on Nvidia and keep cloud revenue streams up and running.



This article originally appeared in Insider Intelligence's Connectivity & Tech Briefing—a daily recap of top stories reshaping the technology industry. Subscribe to have more hard-hitting takeaways delivered to your inbox daily.

- *Are you a client? Click here to subscribe.*
- *Want to learn more about how you can benefit from our expert analysis? Click here.*