China keeps buying hobbled Nvidia cards to train its AI models

Enlarge / A press photo of the Nvidia H100 Tensor Core GPU.

reader comments

72 with

The US acted aggressively last year to limit China's ability to develop artificial intelligence for military purposes, blocking the sale there of the most advanced US chips used to train AI systems.

Big advances in the chips used to develop generative AI have meant that the latest US technology on sale in China is more powerful than anything available before. That is despite the fact that the chips have been deliberately hobbled for the Chinese market to limit their capabilities, making them less effective than products available elsewhere in the world.

The result has been soaring Chinese orders for the latest advanced US processors. China's leading Internet companies have placed orders for $5 billion worth of chips from Nvidia, whose graphical processing units have become the workhorse for training large AI models.

The impact of soaring global demand for Nvidia's products is likely to underpin the chipmaker's second-quarter financial results due to be announced on Wednesday.

Besides reflecting demand for improved chips to train the Internet companies' latest large language models, the rush has

also been prompted by worries that the US might tighten its export controls further, making even these limited products unavailable in the future.

However, Bill Dally, Nvidia's chief scientist, suggested that the US export controls would have greater impact in the future.

"As training requirements [for the most advanced AI systems] continue to double every six to 12 months," the gap between chips sold in China and those available in the rest of the world "will grow quickly," he said.

Advertisement

Capping processing speeds

Last year's US export controls on chips were part of a package that included preventing Chinese customers from buying the equipment needed to make advanced chips.

Washington set a cap on the maximum processing speed of chips that could be sold in China, as well as the rate at which the chips can transfer data—a critical factor when it comes to training large AI models, a data-intensive job that requires connecting large numbers of chips together.

Nvidia responded by cutting the data transfer rate on its A100 processors, at the time its top-of-the-line GPUs, creating a new product for China called the A800 that satisfied the export controls.

This year, it has followed with data transfer limits on its H100, a new and far more powerful processor that was specially designed to train large language models, creating a version called the H800 for the Chinese market.

The chipmaker has not disclosed the technical capabilities of the made-for-China processors, but computer makers have been

open about the details. Lenovo, for instance, advertises servers containing H800 chips that it says are identical in every way to H100s sold elsewhere in the world, except that they have a transfer rate of only 400 gigabytes per second.

That is below the 600GB/s limit the US has set for chip exports to China. By comparison, Nvidia has said its H100, which it began shipping to customers earlier this year, has a transfer rate of 900GB/s.

The lower transfer rate in China means that users of the chips there face longer training times for their AI systems than Nvidia's customers elsewhere in the world—an important limitation as the models have grown in size.

The longer training times raise costs since chips will need to consume more power, one of the biggest expenses with large models.

Advertisement

However, even with these limits, the H800 chips on sale in China are more powerful than anything available anywhere else before this year, leading to the huge demand.

The H800 chips are five times faster than the A100 chips that had been Nvidia's most powerful GPUs, according to Patrick Moorhead, a US chip analyst at Moor Insights & Strategy.

That means that Chinese Internet companies that trained their AI models using top-of-the-line chips bought before the US export controls can still expect big improvements by buying the latest semiconductors, he said.

"It appears the US government wants to not shut down China's AI effort, but make it harder," said Moorhead.

Cost-benefit

Many Chinese tech companies are still at the stage of pre-training large language models, which burns a lot of performance from individual GPU chips and demands a high degree of data transfer capability.

Only Nvidia's chips can provide the efficiency needed for pre-training, say Chinese AI engineers. The individual chip performance of the 800 series, despite the weakened transfer speeds, is still ahead of others on the market.

"Nvidia's GPUs may seem expensive but are, in fact, the most cost-effective option," said one AI engineer at a leading Chinese Internet company.

Other GPU vendors quoted lower prices with more timely service, the engineer said, but the company judged that the training and development costs would rack up and that it would have the extra burden of uncertainty.

Nvidia's offering includes the software ecosystem, with its computing platform Compute Unified Device Architecture, or Cuda, that it set up in 2006 and that has become part of the AI infrastructure.

Industry analysts believe that Chinese companies may soon face limitations in the speed of interconnections between the 800-series chips. This could hinder their ability to deal with the increasing amount of data required for AI training, and they will be hampered as they delve deeper into researching and developing large language models.

Charlie Chai, a Shanghai-based analyst at 86Research, compared the situation with building many factories with congested motorways between them. Even companies that can accommodate the weakened chips might face problems within the next two or three years, he added.

© 2023 The Financial Times Ltd. All rights reserved. Please do not copy and paste FT articles and redistribute by email or post to the web.

https://arstechnica.com/?p=1962036