

Breaking! Nvidia and AMD may cut off supply of high-end GPUs, and China's AI computing may need to start over

*Note: These are Jeffrey Ding's informal and unofficial translations -- all credit for the original goes to the authors and the original text linked below. These are informal translations and all credit for the original work goes to the authors. Others are welcome to share **excerpts** from these translations as long as my original translation is cited. Commenters should be aware that the Google Doc is also publicly shareable by link. These translations are part of the ChinAI newsletter - weekly-updated library of translations from Chinese thinkers on AI-related issues: <https://chinai.substack.com/>*

Source: xinzhiyuan [AI New Era]

Date: August 31, 2022

Original Mandarin: <https://mp.weixin.qq.com/s/r46GyMScFozhRsNsSx-lsg>

Recently, AMD and Nvidia received news that the US government asked them to cut off the supply of high-end GPUs to China.

On the evening of August 31, ijiwei.com suddenly announced that AMD and Nvidia in China have successively received notices from the headquarters to cut off the supply of top computing chips for artificial intelligence and data centers to customers in China.

AMD side:

1. Suspend all MI100 and MI200 shipments to China;
2. Collect statistics of previous MI100 shipments in China;
3. Collect statistics of the customer list and delivery details in China for the MI200.

On the Nvidia side:

1. Suspend A100 and H100 shipments to all customers and all agents in China; other GPU cards will not be affected;
2. The A100s in stock of each server OEM can continue to be delivered to their respective industry customers, and NVIDIA China has not issued any letters to OEMs at present.

At present, the NVIDIA H100 has just been released at this year's GTC (Nvidia's GPU Technology Conference) and has not yet shipped. But the A100, which has been out for 2 years, leads the latest MLPerf with 16 performance records.

It can be seen that just the cut-off of these two high-end GPUs by NVIDIA directly hits the training of intelligent computing and large AI models.

By press time, AMD and Nvidia declined to comment.

America intensifies pressure once again!

Unfortunately, this news does not seem to be groundless.

The U.S. government has issued a notice to Nvidia on Aug. 26 proposing new export control licenses for the existing A100 and upcoming H100, effective immediately, according to a Nvidia filing with the U.S. Securities and Exchange Commission.

In addition, DGX or any other system using the A100 or H100 and A100X, and any future capable of peak performance and I/O performance at or above the A100, are covered by the new license requirements.

The documents show that the ban targets China, Hong Kong and Russia. However, Nvidia has no paying customers in Russia.

Nvidia said the move could affect H100 development, as well as its ability to support existing A100 customers, and could further require the company to move some business out of China.

Currently, Nvidia is approaching the U.S. government in an attempt to seek exemptions for the company's internal development and support activities.

An Nvidia spokesperson also said, "We are discussing countermeasures with Chinese customers to meet their planned or future needs to purchase alternative products. If the alternative products for customers are insufficient, we will seek exemptions or licenses for customers, but cannot guarantee the US government will approve or respond in a timely manner."

Affected by the ban, Nvidia shares fell nearly 6.6% after the opening bell.

On August 24, Nvidia had expected sales of about \$5.9 billion in the third fiscal quarter. However, as soon as the ban came out, Nvidia's potential sales in China in the third quarter may drop \$400 million. If customers do not want to buy replacement (old) products and the U.S. government refuses to or fails to promptly issue licenses to key customers.

In addition, according to an AMD spokesman, the U.S. government has also notified them that it will stop exporting top-of-the-line artificial intelligence chips to China.

AMD said the new licensing requirements will prevent MI250 chips from being shipped to China, but MI100 chips should not be affected.

AMD believes that the new regulations will not have a material impact on its business. But shares still fell after the opening bell.

How big is the impact on China?

In recent years, the United States has not only continuously tightened export restrictions on chips against China, but also tried to bring manufacturing back to its homeland.

Biden has officially signed the "CHIPS and Science Act", the U.S. Department of Commerce has cut off the supply of equipment to China that can be used for advanced process chip manufacturing below 14nm, implemented export controls on EDA (electronic design automation) software tools, and the U.S. is promoting alliances with other countries .

In addition to hindering Nvidia's business of up to US\$400 million in China, the United States' imposition of a chip export ban on Nvidia and AMD will also affect Chinese research in the field of AI.

However, at present, the impact on the consumer electronics field is not serious. In addition to these two chips, there are also chips available from companies such as Qualcomm, MediaTek, and Samsung.

Intelligent computing

However, most domestic servers are inseparable from these two chips. Obviously, this cutoff by the United States is aimed at computing power, and the purpose is to prevent China from leading the world in the field of artificial intelligence.

Building a powerful intelligent computing center requires high-end GPUs. Without intelligent computing, many fields such as smart transportation, smart cities, and industrial interconnection will be affected.

Therefore, the interruption of chip supply will directly affect the development of the Chinese cloud computing industry and artificial intelligence industry.

According to the latest global supercomputing TOP500 list released in June, the top ten supercomputers basically all use processors or technologies from AMD, NVIDIA, and Intel.

Among them, China's Sunway TaihuLight supercomputer uses a self-developed Sunway 26010 many-core processor. The Tianhe-2 supercomputer uses the Xeon Phi 31S1P coprocessor based on Intel's integrated core architecture.

Therefore, the AMD and Nvidia cut-off has not affected the two most influential supercomputers in China.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	LUMI - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	214.35	2,942
4	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
5	Sierra - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438



However, for domestic companies that provide servers, they are not so lucky. Servers are mostly used by enterprises, such as in data centers and cloud computing.

At present, domestic companies such as Alibaba are starting to develop their own cloud-native processors, which is worthy of recognition.

Metaverse

You know, the GPU is the base for the core computing resources of the Metaverse. In the future, the market size of the Metaverse may exceed \$470 billion.

There are a large number of programs in the metaverse that need to be calculated, and the virtual content, blockchain network, and artificial intelligence technology that constitute the metaverse are all inseparable from the support of computing power.

Without strong computing power, the Metaverse is like a castle in the air.

Therefore, more realistic modeling and interaction in the metaverse requires stronger computing power as a premise, which further illustrates the importance of self-developed GPUs.

Large AI models

In addition, without chips from companies such as Nvidia and AMD, Chinese undertakings in image recognition, speech recognition, and other machine learning domains will also be severely affected.

In 2020, the parameters of the GPT-3 model, a huge development, reached 175 billion, and the size of its training data set will also exceed 500GB.

Training such a large AI model will inevitably consume a lot of computing power. GPT-3 alone consumes 10,000 GPUs and takes 30 days to complete.

So, how will the cut-off of high-end GPUs affect the training of domestic AI models?

Taking the NVIDIA A100 as an example, its deep learning performance can reach 3.5 times that of the V100 in the actual measurement in 2021. In the latest AI chip benchmark results, the A100 broke 16 performance records.

Compared to its predecessor, the A100, which "only" had 54 billion transistors, Nvidia packed 80 billion transistors into the H100 and used a custom TSMC 4nm process.

In terms of computing power, the performance of H100's FP16, TF32 and FP64 are 3 times that of A100, which are 2000 TFLOPS, 1000 TFLOPS and 60 TFLOPS respectively. In addition, H100 also adds support for FP8, with a computing power of up to 4000 TFLOPS, which is 6 times faster than A100.

In AI training, if H100 is used to train GPT-3, the speed can be increased by 6.3 times, and if combined with new precision, chip interconnection technology and software, it can be increased by 9 times.



In addition, with the new Hopper architecture, training of large models can be completed in days or even hours.

Among them, the Transformer Engine can package and process FP8 data at twice the speed of FP16, so the data that can be processed by FP8 at each layer of the model can be doubled in speed.

If the US really cuts off supply of high-end GPUs, are domestic GPUs a viable alternative?

Is the domestic GPU ready?

"芯智讯 [icsmart.cn]" stated in a more detailed summary that, from an overall point of view, Chinese GPUs are still in an early stage of development.

In the field of graphics cards, only Jingjia Micro has made some achievements, but there is still a big gap with NVIDIA and AMD.

As for the high-performance computing field affected by the supply cut-off, the article analyzed that only Iluvatar Corex, Biren Technology and Innosilicon have launched corresponding products.

Among them, according to the official data released by Biren Technology, the BR100 has surpassed NVIDIA's A100 in terms of AI computing power. However, since the product has just been launched, it still needs to be tested by the market.

The United States has been suppressing China for a long time, and domestic GPU manufacturers are also accelerating their own research and development, just to achieve self-sufficiency at a critical juncture.

Now, this self-research road still has a long way to go.



自己都能研发，要相信本国的科技实力。。。。最棒的👍

19 分钟前 · IP 属地北京

新智元



我们自产芯片已经在路上了，老美又帮了我们一把

7 分钟前 · IP 属地贵州

新智元

References:

https://mp.weixin.qq.com/s/V7X2gRVZchHVg9WTrI_P-g

<https://laoyaoba.com/n/831077>

<https://www.reuters.com/technology/nvidia-says-us-has-imposed-new-license-requirement-future-exports-china-2022-08-31/>

<https://www.reuters.com/technology/amd-says-us-told-it-stop-shipping-top-ai-chip-china-2022-08-31/>

Comments on the article



Currently Sugon's GPU (Sugon calls it a DCU) is about 60% of the performance of A100. It's just that the (surrounding) ecosystem is not ideal. Even though it's not that easy to use, it can still do high-performance computing. If the U.S. overestimates itself like this, then this will cause China to perfect its own ecosystem. In fact, this is a good thing. Currently, Sugon's DCU has a price of about 3090. Once Sugon's DCU gets more complete, then Nvidia will be no big deal. (132 likes)

Maybe the next step will be banning deep learning frameworks torch and tensorflow, this will reveal the importance of Baidu's self-developed PaddlePaddle framework. (29 likes)