COVER STORY

The AI Lockout

Losing access to Nvidia chips isn't just a hardware problem for China.

BY ELIOT CHEN — MARCH 12, 2023

Illustration by Tim Marrs

The letter almost cost Nvidia $400 million. In late August, the U.S. Commerce Department sent a notice to the American chipmaker ordering it to halt sales of its most important chip, the A100, to Chinese customers. Overnight, hundreds of millions of dollars in potential sales that quarter looked as if they might evaporate. The company's stock price fell 6 percent, and Nvidia warned the drop in sales could even slow down the development of its next generation chip, the H100.

Like almost every other American chipmaker today, Nvidia (pronounced "in-vi-dee-uh") is heavily dependent on the Chinese market. In 2022, more than a fifth of Nvidia's revenue came from China. The company's concern about the letter underscored how U.S. efforts to hamstring China's chip industry could backfire on its own domestic champions. The new requirements, Nvidia warned in its annual report, "may benefit certain of our competitors… and encourage customers in China to pursue alternatives to our products, including semiconductor suppliers based in China, Europe, and Israel."

King Chip

In January, Nvidia overtook Taiwan Semiconductor Manufacturing Company to become the world's largest chipmaker by market capitalization.

— Nvidia — TSMC

Fast forward six months and Nvidia has posted record sales. Not only did the company quickly roll out a modified chip — the A800 — that meets export

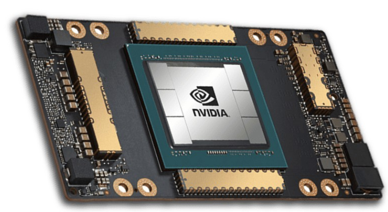
Data: S&P Capital IQ

requirements for China, but it is also cashing in on a broader surge of artificial intelligence (AI) development. Although its China sales did fall, revenue from Nvidia's compute and networking segment — which includes its AI computing business — grew 36 percent year-on-year to $15 billion. And in January, the 30-year-old company overtook Taiwan Semiconductor Manufacturing Company (TSMC) to become the world's most valuable chipmaker by market capitalization. [1]

The real costs of the letter, meanwhile, are being borne by China, which now stands to lose one of the most significant tech races of the century.

Nvidia's chips are seen as essential infrastructure to the rapidly growing field of AI computing. And the U.S. government's ban on selling the A100 to China came at a time when China needs them the most. The debut of OpenAI's ChatGPT in November served as a wake up call to the public about AI's potential, and companies are rushing to develop their own models. AI modeling — the creation, training and deployment of algorithms — is reaching an inflection point, analysts say, moving from pure "training" to "inference," where models can make predictions and produce tangible results. Generative AI models today can write not just prose but software code. Others can produce artwork and compose music. As these AI models get ever more complex, demand for high-performance computing resources is taking off at an exponential rate.

Positioned to supply that demand is Nvidia. Although the Santa Clara, California-based company was once synonymous with video gaming, it has evolved into something of a linchpin for the AI industry, thanks in large part to its proprietary software, called CUDA, which hosts a variety of tools and templates to make programming easier. Developers from Silicon Valley to Shenzhen want Nvidia's chips not only because they are powerful but because they grant access to Nvidia's ecosystem, which is where much of the AI revolution is taking place.



Nvidia's A100 chip. *Credit: Nvidia*

"A lot of things are happening in AI specific to Nvidia," says Hans Mosesmann, a longtime semiconductor analyst and managing director at Rosenblatt Securities. "The CUDA platform has all kinds of drivers, frameworks and acceleration libraries. If you're trying to cure cancer, you can start working on pharmaceutical-related models by using Nvidia-related libraries without having to hire 100-150 engineers to program a custom algorithm. Nvidia is years and years ahead of everybody else because of CUDA."

As a result, losing access to Nvidia chips isn't just a hardware problem for China. Because of its huge investment in CUDA, the company has produced something more akin to a walled garden — with nearly every major player in the AI space residing inside.

The company's proprietary hardware and software is used in supercomputers throughout the world, from privately run systems built by companies like Microsoft and Baidu to national laboratories. In China, where artificial intelligence is a national strategic priority, Nvidia's footprint is especially significant: an estimated 95 percent of the country's large data centers run on Nvidia chips, according to Fubon Securities, a Taiwanese brokerage firm.

An aerial view of Huawei's data center in Gui'an New Area during its opening ceremony on December 20, 2021 in Gui'an, Guizhou, China. *Credit: VCG via Getty Images*

China's AI development thrived within Nvidia's ecosystem, in part because they had ready access to Nvidia's technical support and latest upgrades. In 2010, when China introduced a supercomputer that was able to outperform the U.S.'s top contender for the first time, it used Nvidia chips. The following year, Nvidia began partnering with China's Ministry of Education to offer CUDA programming courses in universities across the country. The company's customer list is a veritable who's who of China's tech sector: Baidu, Alibaba, Tencent, SenseTime, iFlytek and others. Until recently, China's AI sector was considered on par with, or even slightly ahead of, America's.

But having grown up inside Nvidia's walled garden, China's universities, laboratories and tech firms now find themselves locked out. In October, just two months after the U.S. government limited Nvidia's sales in China, the authorities also released a new set of rules that have forced some of Nvidia's R&D and support staff out of the country.[2] Nvidia declined to comment for this story.

"Previously, U.S. trade policy was to sell tech to China that facilitated its advancement in the semiconductor industry, but to ensure our comparative advantage," says Gregory Allen, director of the artificial intelligence governance project at the Center for Strategic and International Studies (CSIS), a Washington D.C. think tank. "Now, we've moved from slowing the pace of Chinese chipmaking to actively degrading it."

> **This AI inflection is happening two to three years earlier than China would like. They're not where they need to be yet.**
>
> — *Hans Mosesmann, a longtime semiconductor analyst and managing director at Rosenblatt Securities*

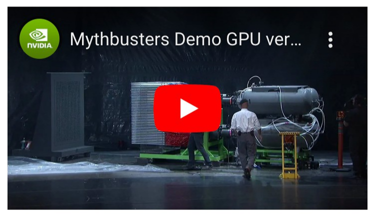
Nvidia quickly offered a lifeline by producing the A800 chip for China, but analysts note that the speed at which the AI sector is now moving could quickly make the A800 obsolete.

Although advances in chipmaking tend to follow Moore's Law — which dictates that the number of transistors on chips doubles every two to three years, unlocking ever more computing power over time — the increase in demand for computing power (or just "compute") by the AI sector is outpacing supply. Some analysts say the amount of compute needed in the largest AI training runs is doubling every three to four months. Already, in October, Nvidia released its next generation AI chip, the H100, which is said to be around three times more powerful than the A100. Orders from Google, Microsoft, Oracle and others have poured in, but it is, of course, banned from sale to China.

"This AI inflection is happening two to three years earlier than China would like. They're not where they need to be yet," Mosesmann says. "If you're China, you've hit an air pocket — you have a problem."

KUDOS TO CUDA

In order to explain the revolutionary power of its chips, Nvidia organized an elaborate stunt in 2008 together with the hosts of Discovery Channel's television program, *Mythbusters.* Using a paintball gun, the company illustrated the difference between its famed graphics processing units (GPUs) and the more traditional central processing units (CPUs), whose production has historically been dominated by companies like Intel.



A video from NVIDIA demonstrating the difference between GPUs and CPUs, made in collaboration with Adam Savage and Jamie Hyneman from the Mythbusters. August, 2008.

The CPU is the heart of a computer. It's capable of complex calculations, but it is mostly forced to perform them one by one. In the *Mythbusters* demonstration, it is represented by a single paintball gun, firing pellets sequentially in a circle to paint a crude smiley face.

What makes a GPU invaluable is that it can do "parallel computing": instead of solving problems sequentially, the GPU breaks a complex problem into thousands or millions of simple tasks and works them all out at once. In *Mythbusters* terms, the GPU is a multi-barreled paintball super-cannon — at the pull of a single trigger, it fires 1,100 pellets to produce a version of the Mona Lisa.

Most personal computers these days contain both a CPU and GPU: the two chips work in tandem, performing complementary tasks. But when Nvidia invented the GPU in the late 1990s, it was seen largely as an accessory for computers, which were limited at the time to displaying relatively primitive graphics, like Microsoft's Clippy.

Nvidia's co-founders — Chris Malachowsky, Curtis Priem and Jen-Hsun Huang, who remains Nvidia's CEO — bet that the next generation of computers would demand more complex graphics, including 3D illustrations. To produce those efficiently, computers would need a new kind of chip.

Computer screens are composed of thousands — if not millions — of individual pixels, each of which displays a specific shade of light. The mosaic of pixels comes together to reveal an image, but rendering that image requires making individual calculations about what color to display in each pixel. These calculations are basic, but the sheer number of them quickly add up to become very costly. Like the paintball super-cannon, Nvidia's GPU makes those calculations all at once.

TSMC's Turning Point

BY GREGOR STUART HUNTER

TSMC has long been viewed as the greatest strength in Taiwan's 'silicon shield.' If China invaded Taiwan, the theory went, the island's outsize role in...

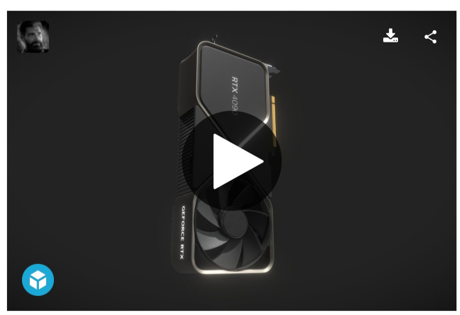
To turn its blueprints into physical chips, Nvidia needed a factory, known as a fabrication plant or fab. For the upstart firm, building its own fab would have been prohibitively expensive and complicated, so Nvidia bounced between several fabs in Europe before landing, in 1998, on TSMC. The two companies have been partners ever since, and today, they are the world's two largest chip companies by market capitalization.

"TSMC in the 1990s was not yet the world's most important firm, but in some ways they [TSMC and Nvidia] have grown up together," says Chris Miller, author of *Chip War: The Fight for the World's Most Critical Technology*. "You could argue Nvidia is the most successful firm to have been founded as a fabless chip designer."

As computer graphics improved, Nvidia found opportunities in the video gaming industry, including installing its chips in Sony's Playstation and Microsoft's Xbox. In China, where the government banned game consoles in 2000 out of fear of their addictive quality, Nvidia sold GPUs for desktop computers, helping to fuel an explosion in PC gaming. Between 2003 and 2010, the share of Nvidia's revenue coming from China tripled from 13 to 39 percent.

It wasn't until the late 2000s that engineers at Nvidia began to realize their chips might have applications beyond video gaming. As it turns out, the core competency of the GPU — taking a complex task and breaking it down into millions of smaller tasks to be worked out at the same time — is well suited to a variety of tasks. Take cryptography, for instance. When a computer tries to decrypt a password, it uses brute force, making consecutive guesses until it gets the right combination — a simple but time consuming form of computation. Intel's i9 chip has at most 16 cores, allowing it to make 16 guesses at a time. But a modern GPU, like Nvidia's RTX 4090, has more than 16,000 cores, allowing it to work orders of magnitude faster.

This, incidentally, is the same principle behind cryptocurrency mining, which essentially involves solving complex math problems for cash. And as interest in cryptocurrencies exploded around 2013, Bitcoin 'miners' bought up huge volumes of GPUs, driving up their price.

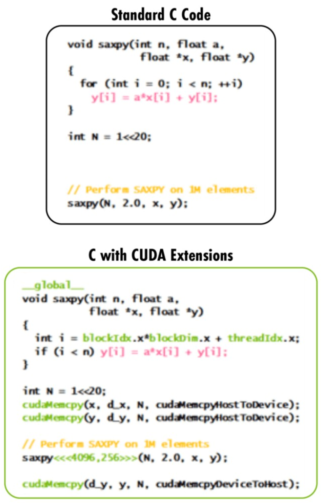


An interactive 3D model of Nvidia's GeForce RTX 4090 GPU.

Nvidia, sensing the creative power of GPUs, put these windfall earnings to good use. Since 2006, the company has spent over $10 billion developing a proprietary platform that allows all of Nvidia's GPUs to be programmed in a standard programming language. For engineers and researchers, the platform, called Compute Unified Device Architecture, or CUDA, was a boon because it made it easy for them to write code that worked across a range of systems — as long as those systems were made by Nvidia.

> **"Although other companies might develop GPUs that have comparable computing power, no other company has a platform with capabilities close to CUDA."**

The benefits of building CUDA weren't obvious to shareholders and analysts at the time. Mosesmann, at Rosenblatt Securities, says that "for the better part of the last 12–15 years, Jensen got a lot of flack" for pouring money into CUDA. "It's something that only in hindsight looks brilliant," he says.



An example showing how a standard C program can be accelerated using CUDA. *Credit: Nvidia*

In retrospect, it's an achievement somewhat akin to what Apple has done with iOS. Many iPhone users didn't intend to get locked into Apple's ecosystem, but once your life makes its way onto iMessage, iPhoto and iCloud, the cost of switching to anything else becomes very high. Likewise, for anyone working with GPUs, Nvidia has become the default choice.

"Nvidia is really widely used right now because of CUDA," says Curtis Hillegas, associate CIO of research computing at Princeton University, where he oversees the procurement and management of the university's supercomputers, which have used Nvidia chips since 2011. "People have their codes developed in CUDA, the GPUs perform well, and Nvidia keeps developing. It's a lot of work to change your code for something other than CUDA."

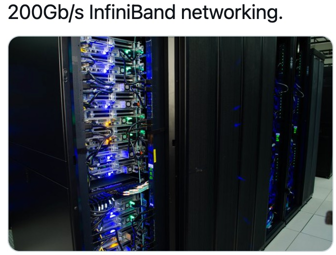
At Princeton, Hillegas says that around 1,500 researchers are using the university's supercomputers at any given time to tackle a wide range of problems, including earthquake seismology, high energy physics and natural language processing. When Princeton's systems can't perform the computations researchers need, they turn to the much more powerful national labs, like the Texas Advanced Computing Center (TACC), a federally funded high-performance computing facility.

CUDA facilitates this coordination as well, further entrenching Nvidia's dominance: Code written in CUDA for Princeton's Nvidia GPUs needs to run seamlessly at TACC — meaning TACC also needs to buy Nvidia GPUs.

"Our systems are designed to run the science problems that can't be run anywhere else," says Tommy Minyard, director of advanced computing systems at TACC. "And Nvidia has made a lot of effort and put forth a lot of software development, not just into CUDA, but all the libraries associated with it. Having those available has been really key."

TACC
@TACC · Follow

TACC's Lonestar6 supercomputer recently added upgraded servers. Side view of 3 new racks of Dell R7525 servers, adding 144 GPUs to original 96 + bringing total GPU count in system to 240 NVIDIA A100 40GB Tensor Core GPUs interconnected by NVIDIA HDR 200Gb/s InfiniBand networking.



11:06 AM · Dec 14, 2022

15 Reply Copy link

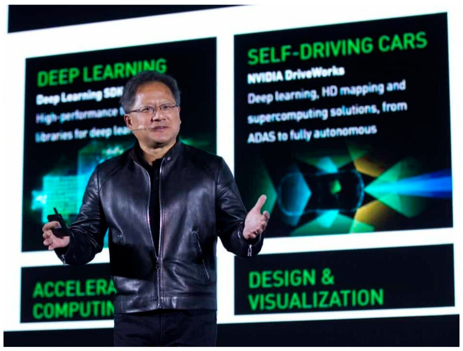
Read more on Twitter

The same dynamic has been at play in China. A recent U.S. study of Chinese large language models (LLMs) — an AI tool that can read, summarize and generate words, like ChatGPT — found that the majority of them were trained using Nvidia chips. It is a sign of how lopsided the competition for the Chinese market has become that, while Nvidia warned of $400 million in losses after the chip restrictions were announced in August, its closest competitor, AMD, stated that the restrictions on its China sales would not have a "material impact on our business."

Given how entrenched Nvidia is in China, Allen at CSIS notes that the country's lack of a domestic software ecosystem like CUDA is "a self-identified source of weakness in China."

But replicating Nvidia's success with CUDA will not be easy, especially as Nvidia continues to invest in and nurture its AI garden. Although other companies might develop GPUs that have comparable computing power, no other company has a platform with capabilities close to CUDA.

"There's a chance that some of these chips could be competitive with Nvidia," says James Sanders, a cloud and infrastructure analyst at CCS Insight, a research and advisory firm. "But the stumbling block is going to be with software."



Jen-Hsun Huang announcing Nvidia AI products at GTC China, September 22, 2016. *Credit: Nvidia*

Clearly aware of Nvidia's position, a bullish Huang described a near-utopian vision for the future of AI computing in an earnings call in late February.

"Over the course of the next 10 years, through new chips, new interconnects, new systems… I believe we're going to accelerate AI by another million X," the chief executive said. Nvidia is no longer "just a chip company," he said, because AI necessitates thinking across the entire product lineup: "all the way from the chip to the data center, across the network to the software."

So far, Nvidia is the only company capable of doing that. Which is precisely why China still needs it.

CHINA'S NVIDIA ENVY

From the outside, the Wan Chai Computer Centre in Hong Kong is a drab building that blends into its surroundings in the city's commercial district. But inside, the maze of corridors lined with fluorescent storefronts buzzes with enthusiasts looking to upgrade their gaming experience. On a recent Saturday afternoon, Nvidia's trademark black-and-green boxes were stacked on dollies; the RTX 4090 — the company's highest end graphics card — was on sale for between $1,800 and $2,100.

Inside the Wan Chai Computer Centre. *Credit: The Wire*

Stacks of Nvidia's 40 series GPUs. *Credit: The Wire*

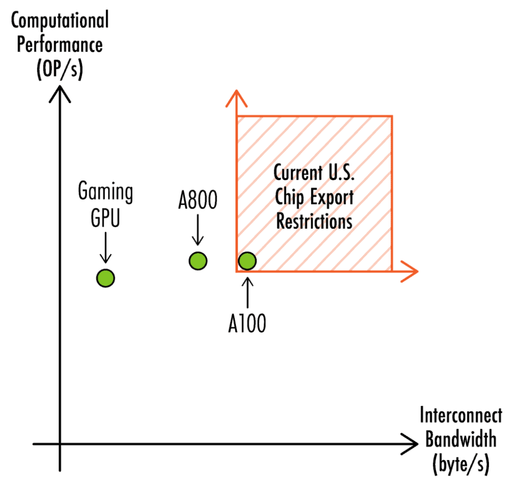
Nothing is stopping Nvidia from selling its RTX gaming chips in China, even though the chips have almost the same computational power as the A100s and have been used by China's top nuclear weapons lab. The Biden administration purposefully left out consumer-grade chips from its export restrictions.

“You only want to ban a certain type of chip,” says Lennart Heim, a research scholar at the Centre for the Governance of AI (GovAI), a research nonprofit. “The U.S. is specifically interested in targeting AI applications for data centers.”

To do this, the U.S. Commerce Department had to largely ignore computational power as a threshold and focus on a second metric — “interconnect bandwidth” — to judge chips. If computational power measures how fast chips can think, interconnect bandwidth measures how fast chips can talk to one another. It is especially important in AI computing, because today’s AI systems don’t fit on a single GPU; instead, many chips are wired together.

The October export ban set the thresholds for computing power and interconnect speed “exactly where the A100 sits,” Heim says.

The current restrictions leave room for companies like Nvidia to eventually sell more powerful chips to China, as long as they throttle the interconnect. This, after all, is what Nvidia did with the A800 — its computational power is roughly the same as the A100, but its interconnect speed is cut by one-third. The same could theoretically be done for the H100 and future generations of Nvidia chips.



A visualization of the U.S. export controls on chips. *Credit: Lennart Heim*. Modifications made to original graphic to adapt to house style.

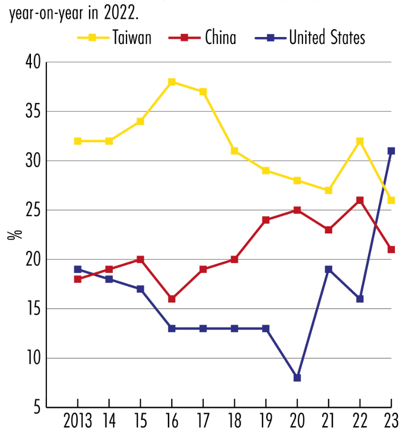
Such a compromise, Heim notes, suggests that the chip restrictions might not be the deathblow to China’s AI industry that some make it out to be.

“Everybody talks about interconnect bandwidth being really important, because it’s important in AI,” he says. “But the history of computation is based on computing performance.”

Jack Dongarra, a professor emeritus of computer science at the University of Tennessee, who has been tracking the world’s fastest supercomputers for decades, agrees, saying the current export controls simply slow down but don’t necessarily stop, China’s AI development. “The computation is not going to happen as quickly but you’re still going to get it done,” he says.

Inflection Point

Demand for Nvidia’s AI chips drove U.S. sales up 91 percent year-on-year in 2022.



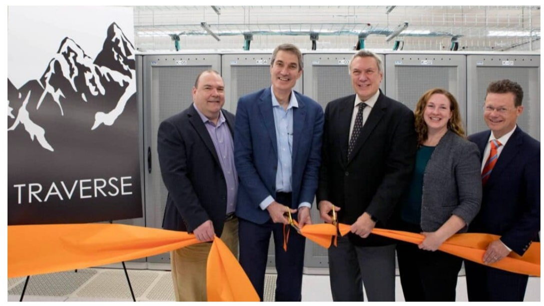
Data: Annual Reports

With the rest of the AI industry barrelling forward at an unprecedented speed, however, any relative delay could be costly for China. The U.S. hopes the export ban has struck the right balance between enabling gamers and thwarting advanced AI weapons systems, but it will take some time before anyone can judge if it is a success. In the immediate future, China’s AI scene is likely OK: companies and local governments were stockpiling the A100 chip before the export ban went into effect, leaving it a reserve of additional computational power. Plus, analysts note, China can try to make up for the slight setback of the A800 versus the A100 with more money and some patience.

“How strong a chip is correlates with money and time. If the chip is better, you can spend less time and money training the model," says Jeffrey Ding, assistant professor of political science at George Washington University, who co-authored the study of Chinese LLMs.

But this strategy will eventually hit a wall, especially once you consider how much electricity advanced AI systems devour. The supercomputers at TACC in Texas, for instance, have an average daily power consumption of six to seven megawatts, equivalent to 5,700 U.S. homes. In his study, Ding found evidence that researchers relying on less efficient Chinese chips were forced to cut short their AI model training runs, potentially due to budgetary constraints.

Despite their high prices (a single A100 retails at around $10,000), Hillegas, at Princeton, says that buying new and more efficient chips every three or so years actually keeps the cost of research down.

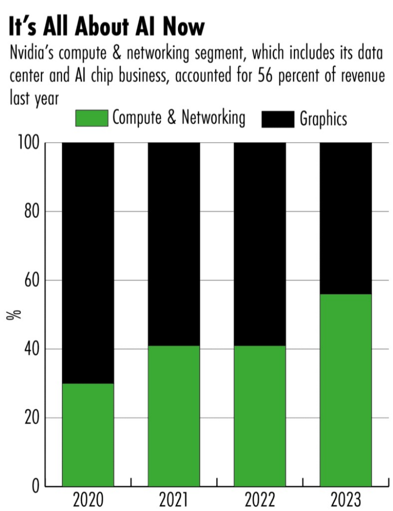


A ribbon cutting ceremony marking the launch of Princeton's newest supercomputer, Traverse. September 30, 2019. *Credit:* *Princeton University*

"What really drives [the need to upgrade systems] is the performance increases on a chip over a generation. If you run a system for three years, one third of the cost is capital cost and two thirds of the cost is running it [electricity and cooling]," he says. "So, it's actually much cheaper to do more capital purchases."

With Chinese companies and researchers likely unable to do large capital purchases of more efficient chips in the coming years, they are facing ballooning costs and considerable delays. Seeing the writing on the wall, there are signs already of how they might try to work around the chip ban.

The *Financial Times* recently reported, for instance, that iFlytek and SenseTime — two AI companies that are on the U.S. entity list — are increasingly relying on cloud computing. By renting time in data centers outside of China, they can still access Nvidia's A100 chips to train their AI models.



Data: Annual Reports

The current U.S. export controls don't prohibit such actions, but China's own data protection laws could make this difficult: New regulations that came into effect last year mandate that companies file for approval from Chinese cybersecurity authorities before sending their data abroad. But if the chip ban pushes Beijing into a corner, it could choose to relax its scrutiny.

Still, analysts note that cloud computing is hardly an ideal solution for China. For starters, it necessitates turning over your data to another company. Even more existentially, it means that your access can be shut off at any moment — giving other countries and companies immense leverage over your operations.

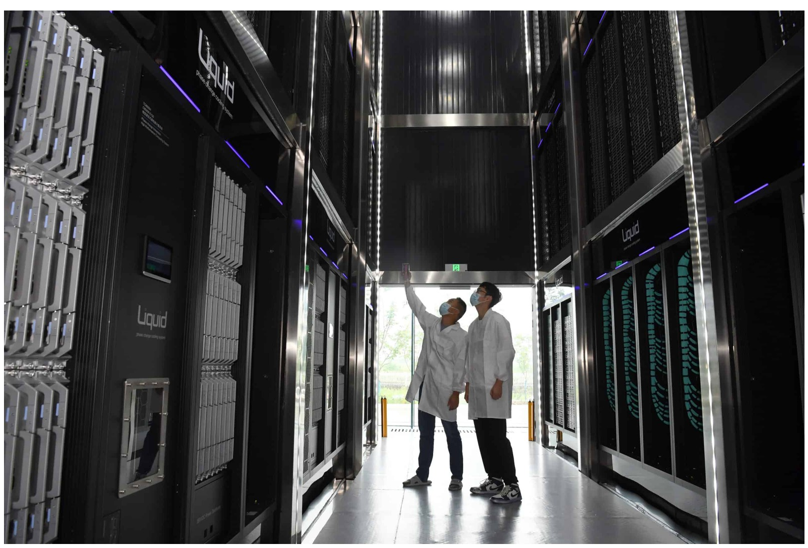
"Cloud computing is favorable from an export control perspective," says Heim, at GovAI. "Once you've sold a chip, it's gone. But you can cut off cloud computing access at any time."

In the longer term, Chinese companies may find reprieve in domestic tech advancements. In the country's most powerful supercomputers, CPUs made by Intel are already being replaced with Chinese alternatives, according to Dongarra, the computer science professor who tracks supercomputers. Homegrown GPU manufacturers are also slowly making progress: Huawei's Ascend 910, a domestically developed AI chip released in 2019, is comparable in performance to Nvidia's V100 chips, the predecessor to the A100, according to a recent study. Another Chinese startup, Biren Technology, is developing an AI chip that it claims can outperform Nvidia's A100 flagship.

> **In the case of China, they're probably going to be willing to spend lots of money on this because of the economic and national security implications.**
>
> — *Tim Fist, an incoming fellow at the Center for A New American Security*

But Huawei and Biren both rely on TSMC to manufacture their designs. And the Taiwanese fab is required to comply with the recent export controls that bar it from working with Chinese firms on advanced chips. Building a domestic fabricator that can rival where TSMC is today, analysts say, could easily take China a decade.

With little hope in catching up, China's best option might be working towards a major disruption to the industry, such as redesigning how AI training takes place. Even in the West, researchers are realizing that the current rhythm is unsustainable: With demand for compute doubling every few months, companies and labs are constantly forced to buy new chips and replace existing systems. Researchers studying "decentralized training" are finding ways to make AI training less communication intensive, allowing AI models to be trained across physically distant GPUs with low interconnect bandwidths, spreading the workload around.



Staff members inspect the phase-change immersion high-performance computing system at Hefei Advanced Computing Center, in Hefei, China, August 13, 2021. *Credit: VCG via Getty Images*

Right now, limits to how efficiently AI training algorithms can communicate with one another make this challenging, but efforts to break through those limits could accelerate, especially in China.

"There is active research happening on how to solve this problem and allow training to be done in a more decentralized way," says Tim Fist, an incoming fellow at the Center for A New American Security, a Washington D.C. think tank. "In the case of China, they're probably going to be willing to spend lots of money on this because of the economic and national security implications."

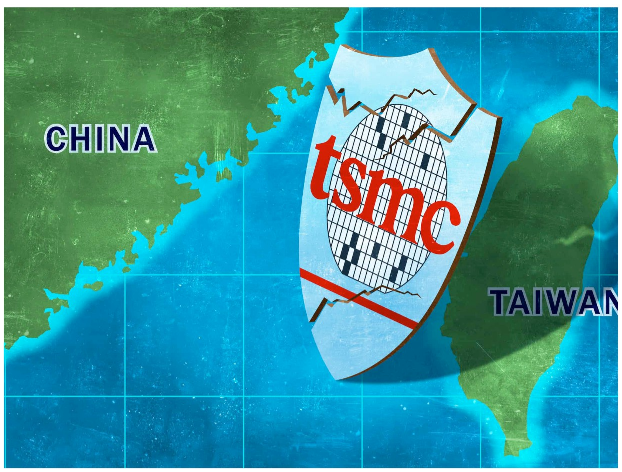
Indeed, given what's at stake, China's only real recourse is to build and tend its own AI garden. This is an enormous task, but it would be foolish to underestimate China. After all, in 2001, China didn't have any supercomputers; today it has more than any other country. And although it's exceedingly difficult to imagine now, perhaps some day none of those computers will have Nvidia chips in them.

Correction: This article has been updated to more accurately describe research efforts into "decentralized training."

Eliot Chen is a Toronto-based staff writer at *The Wire*. Previously, he was a researcher at the Center for Strategic and International Studies' Human Rights Initiative and MacroPolo. @eliotcxchen

---

COVER STORY



TSMC's Turning Point

BY GREGOR STUART HUNTER

TSMC has long been viewed as the greatest strength in Taiwan's 'silicon shield.' If China invaded Taiwan, the theory went, the island's outsize role in the global semiconductor supply chain would create incentives for allied nations to come to its defense. But with TSMC facing significant pressure to expand overseas, and with the U.S. turning the screws on China's own efforts to obtain chips, many in Taiwan fear there are cracks in the shield. Now TSMC must manage its growing pains abroad, while playing defense at home.

THE BIG PICTURE

China Drops the Ball

BY ELLA APOSTOAIE

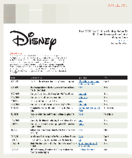
Xi Jinping wants to promote Chinese sports but a series of scandals is clouding their success.

Q & A

Sheena Chestnut Greitens on China's National Security Focus Under Xi Jinping

BY KATRINA NORTHROP

The academic talks about how Xi's idea of national security differs from America's; the recent phenomenon of transnational repression; and why the November protests in China should concern the CCP.



Visit News Products Store

News Products

Our best open-source research on Chinese companies, as well as industry guides to 100 of the most influential people in a China-focused industry.

The Wire China Archives

Read More Articles >