Should the US Sell Blackwell Chips to China?

Assessing the impacts of exporting the B30A AI chip

Georgia Adamson, Saif Khan, Tao Burga, Tim Fist | October 25, 2025

Should the US Sell Blackwell Chips to China?

Assessing the impacts of exporting the B30A AI chip

Georgia Adamson, Saif Khan, Tao Burga, Tim Fist | October 25, 2025

Executive summary

The United States is reportedly considering whether to permit sales of NVIDIA's forthcoming B30A chip to China, following lobbying efforts by NVIDIA.[1] The B30A's reported specifications suggest it will enable roughly equivalent capabilities as NVIDIA's flagship B300 by delivering half the B300's performance at half the price.[2] If meaningful volumes of chips with these specifications are permitted to be exported to China, it would have four major implications:

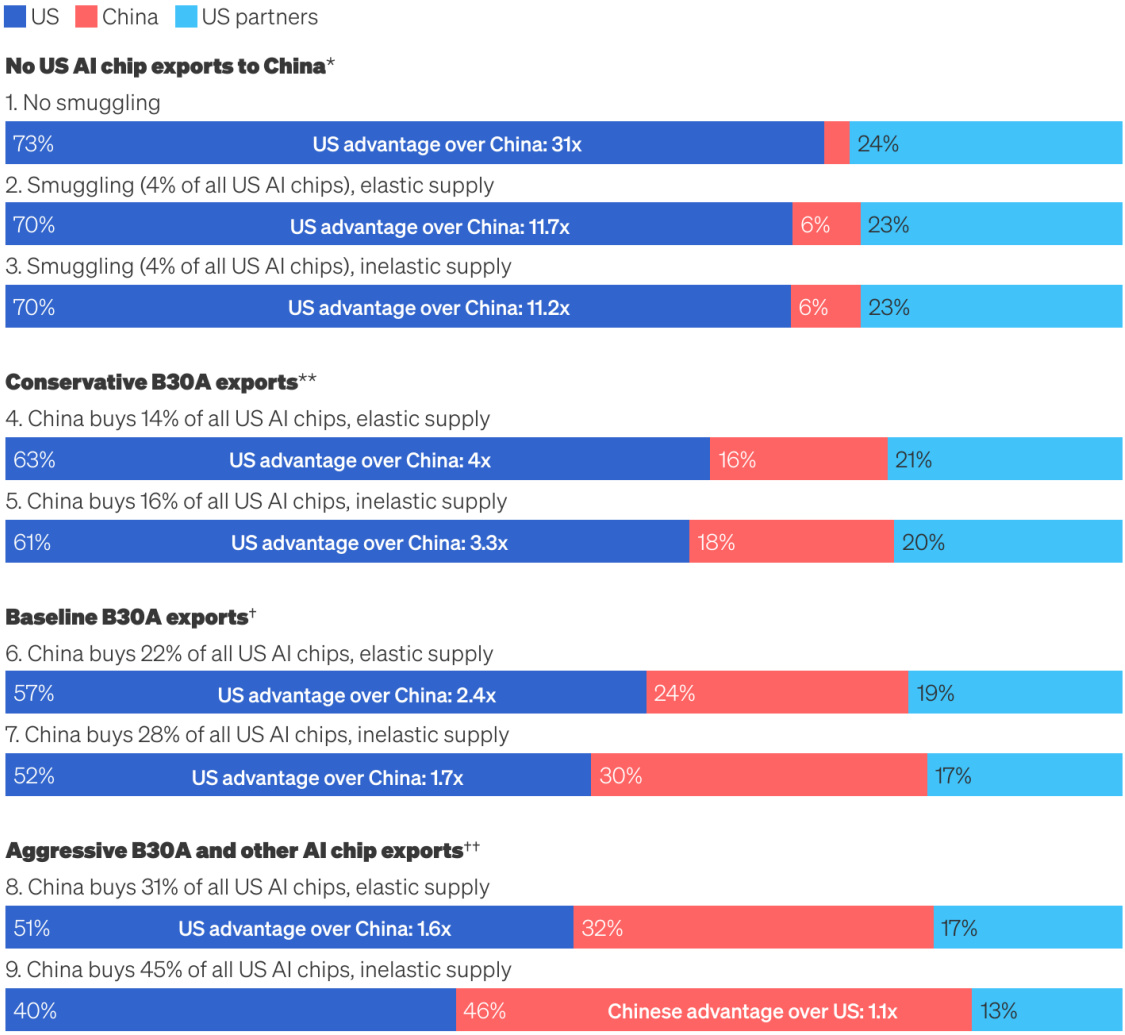
1. **The decision would be a substantial departure from the Trump administration's current export control strategy**, which seeks to deny powerful AI compute to strategic rivals. The B30A would be more than 12 times as powerful as the H20 — a chip which requires a license for export to China and has approved exports in only limited quantities. It would also exceed the United States' current export control performance thresholds by more than 18 times.
2. **Chinese AI labs would have access to AI supercomputers as powerful as those available to US AI labs, at a similar cost.** We calculate that a B30A training cluster would cost around 20% more than a training cluster with equivalent peak processing performance and memory bandwidth based on NVIDIA's cutting-edge B300.[3] This additional cost could be easily covered by state subsidies.

[1] The "B" in B300 and B30A refer to NVIDIA's cutting-edge Blackwell design architecture used in both chips. This piece focuses on the speculated B30A chip specifications as reported by media outlets as of the time of writing. The actual chip name, architecture, or performance characteristics may be different than those suggested by public reporting.
[2] Performance measured in both peak FLOP/s and memory bandwidth. See Appendix 3 for details.
[3] See Appendix 2.

3. **In conditions of supply inelasticity, fewer AI chips would be sold to customers in the United States and the rest of the world.** This could occur if global demand for AI chips exceeded manufacturing capacity, or if Chinese companies using B30As stole market share from US companies selling cloud compute, either in China or elsewhere.
4. **The United States' total AI compute advantage over China would shrink dramatically.** As shown in Figure 1, we estimate that if all US AI chip exports to China are banned in 2026, with no smuggling, the United States would obtain 31x more AI computing power than China. If B30As are instead approved for export to China, this advantage shrinks to less than 4x. In the most aggressive export scenarios involving sales of the B30A chip and comparable AI chips from all other US AI chip companies, this advantage would flip, with China gaining a 1.1x advantage over the United States.

Figure 1: Assessing the impacts of B30A exports by scenario

Impacts measured in terms of percentage of total available 2026-produced AI compute owned by different actors, measured in B300-equivalents. Combines estimates of AI chip production by both US and Chinese firms, as well as estimates of AI chip smuggling.



* Assumes no export of US data center AI chips to China, including the H20
** Assumes NVIDIA has the same share of its revenue from China as it had pre-2022 chip export controls, and that NVIDIA is the dominant exporter
† Assumes Chinese companies pursue a strategy consistent with stockpiling behavior already seen for SME, and that NVIDIA is the dominant exporter
†† Assumes Chinese companies pursue the same stockpiling strategy as described above, but AI chips from all US AI chip companies

Chart: Authors • Source: IFP analysis

IFP

A key argument for allowing exports of the B30A is that sales would satisfy Chinese demand for AI compute that Huawei and other Chinese chip companies would otherwise fill. By cutting off their market, B30A sales could slow China's indigenization efforts and its ability to compete with the US chip industry in global

markets. However, this argument is flawed, for both supply and demand-side reasons:

- Huawei cannot meet demand in either domestic or global markets because it cannot produce AI chips at scale. This is due to domestic chip manufacturing bottlenecks created by extensive US and allied export controls on semiconductor manufacturing equipment (SME).
- China is likely to create artificial demand for Huawei and other Chinese chip companies, such as by maintaining restrictions on US AI chip imports to critical infrastructure. Therefore, US sales will have minimal effect on Huawei's market expansion opportunities.

The most effective way for the United States and its allies to further limit Huawei's capabilities is to halt China's domestic expansion of AI chipmaking. They can do this by tightening country-wide restrictions on SME, especially by barring all exports of deep ultraviolet (DUV) immersion lithography tools needed to make advanced AI chips. The US government can also tighten enforcement of controls on high-bandwidth memory, a critical component that China needs to make AI chips.

Introduction

On January 20, 2025, President Trump issued an America First Trade Policy, directing the Departments of State and Commerce to "eliminate loopholes in existing export controls" to "maintain, obtain, and enhance our Nation's technological edge" over strategic rivals. Building on this strategy, the Administration's AI Action Plan, released on July 10, 2025, pledged to strengthen enforcement of AI compute export controls, noting:

> *Advanced AI compute is essential to the AI era, enabling both economic dynamism and novel military capabilities. Denying our foreign adversaries access to this resource, then, is a matter of both geostrategic competition and national security.*

The Administration's commitment to tightening AI export controls rests on America's main advantage in the AI competition with the People's Republic of China (henceforth "China"): access to large quantities of advanced AI chips. These chips are the basis for AI computing power.

China, meanwhile, excels in many other dimensions of AI production: it produces more STEM PhD graduates than the United States; top Chinese AI company DeepSeek's AI systems demonstrated it could match the efficiency of US competitors earlier this year; China has multiple times the United States' electricity generation capacity; and Chinese companies have the same access as US developers to public internet data used to develop AI systems — as well as data US companies lack, such as that collected by their surveillance state and technology companies.

Despite these advantages, US and allied export controls have successfully constrained China's ability to train and deploy frontier AI models at scale. Since the first Trump Administration's restrictions on advanced AI chipmaking tools, US export controls have aimed to maintain the largest lead possible in the United States' aggregate computing power. These measures have succeeded: today, the United States maintains an estimated fivefold advantage in AI supercomputing capacity over China. Chinese tech firms repeatedly cite a shortage of AI chips as the greatest bottleneck to their AI development and deployment.

Recent statements from President Trump, however, suggest the second Trump Administration may change course on its export control policy objectives. On August 12, 2025, Trump stated that he would not strike a deal with China over NVIDIA's flagship Blackwell chips, which he claimed no other country would have for the next five years. Still, he added, "it's possible I'd make a deal [with a] somewhat enhanced — in a negative way — Blackwell," meaning "take 30% to 50% off of it."

The United States has already approved limited exports of NVIDIA's less powerful H20 chip.[4] But NVIDIA's forthcoming B30A, a downgraded Blackwell model designed for China, is expected to significantly surpass the H20 in performance. If Trump's suggestion of a 30-50% downgraded Blackwell materializes — as media outlets report[5] — China would have access to a chip with roughly 12 to 17 times the computing power of the H20.

In recent weeks, Chinese regulators have banned purchases of the H20 and RTX Pro 6000D — another downgraded NVIDIA chip — ostensibly to promote adoption of domestic alternatives. In response, US policymakers may see an opportunity to

[4] Secretary of Commerce Howard Lutnick has previously claimed that the H20 is only NVIDIA's "fourth-best" chip. This is true for the algorithms used to train AI models, however for efficiently inferencing those models at scale, the H20 is best in class (see Figure 2 of this paper).
[5] Based on reporting that the B30A would have half the processing power as the B300, NVIDIA's most cutting-edge chip to date.

encourage sales of more powerful US chips like the B30A to China in order to deny market opportunities for Huawei and slow China's indigenization efforts. Allowing limited sales could, in theory, help preserve US market share in China while constraining the expansion of Chinese competitors like Huawei, Alibaba, or Tencent from competing with US companies in global markets.

Alternatively, Beijing's restrictions may be intended as leverage for securing B30A chips in upcoming negotiations with Washington. In this scenario, allowing such exports would concede to China's demands while eroding the United States' principal advantage in compute. Sales would accelerate China's frontier AI development, facilitate diffusion across its national security enterprise, and arm Chinese cloud providers with US chips to compete globally with American counterparts. At the same time, such exports may do little to alter China's long-term stated objective of achieving self-sufficiency in advanced AI chips, including through measures designed to stimulate artificial demand for domestically produced chips.[6]

The US government thus faces an inflection point in its export control strategy; policymakers must decide whether to export the B30A or other downgraded versions of cutting-edge Blackwell chips to China. The outcome of this decision will define US export control policy and the future of US-China AI competition. This report assesses the implications of exporting a downgraded Blackwell chip to China — the B30A — based on its reported specifications and the limited public data available.

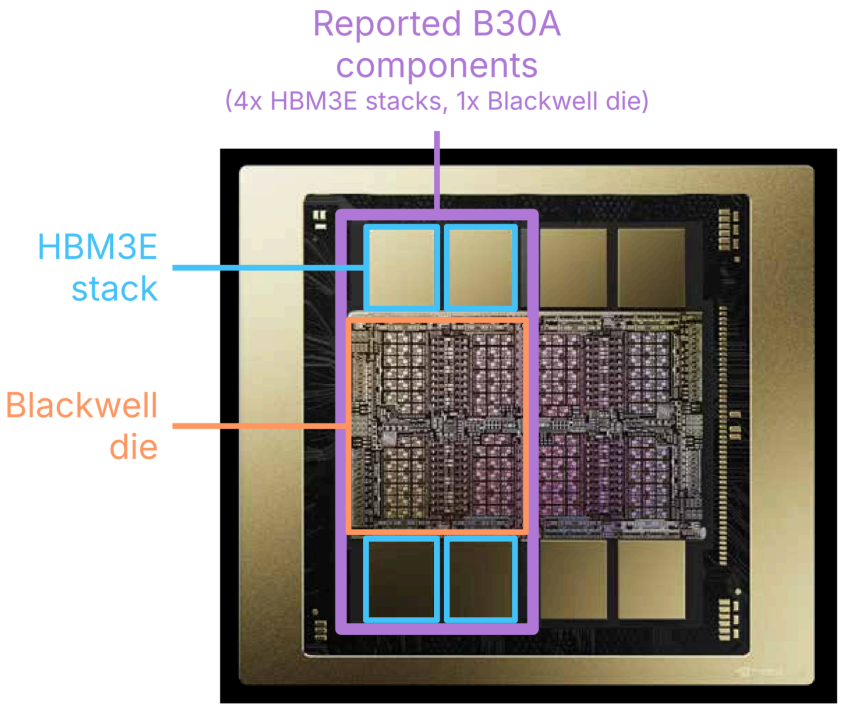
The B30A has similar price-performance to the B300

NVIDIA's flagship B300 contains two pieces of silicon called AI processor dies and eight stacks of high-bandwidth memory (HBM). These dies and HBM stacks perform calculations and high-speed data transfers critical to advanced AI capabilities.[7] By contrast, NVIDIA's downgraded B30A chip would reportedly

[6] For instance, Beijing may look to absorb its own indigenous chips where necessary, such as by banning American AI chips from critical infrastructure as in the case of Micron in May 2023.
[7] A semiconductor die is a small piece of silicon that contains the core electronic circuits of a chip. In the case of AI hardware, an AI processor die carries out the calculations needed to train or run AI models. HBM is a type of advanced memory placed right next to the processor die to enable very fast data transfer. HBM holds the AI model itself, the input data, and the results of the model's computations.

consist of just one die and four HBM stacks, and thus likely attain half the performance of the B300.[8]

Figure 2: B30A components overlaid on B300 system architecture[9]



To compensate for the reduced performance, NVIDIA is expected to price the B30A at about half the cost of a B300 — a move Chinese tech companies have privately called a "good deal." This is because the most relevant measure of AI computing capabilities is not the performance of individual chips, but the performance of large clusters of those chips networked together in data centers. By networking B30As together, China can access computing capabilities similar to those of US AI labs at a similar price. We estimate that to match the capabilities of US labs using a B300-based AI training cluster, Chinese labs would need to spend approximately 20% more when using B30As.[10]

[8] Measured in both FLOP/s and memory bandwidth in TB/s. This configuration for the B30A makes sense, as NVIDIA does not need to substantially alter its existing chip designs, and can flexibly reallocate production capacity between B30As and B300s, depending on demand and whether the B30A is approved for export. NVIDIA may also pursue an alternative configuration, such as by disabling processing cores in the B200 or B300 to reduce performance.
[9] Sources for the Blackwell architecture: NVIDIA GTC25 keynote, NVIDIA Blackwell Ultra system overview. The speculated B30A chip was not mentioned in these sources.
[10] Capabilities as measured in peak FLOP/s and memory bandwidth. See Appendix 2 for details of this calculation.

Table 1: Comparing speculated B30A and B300 performance and cost*

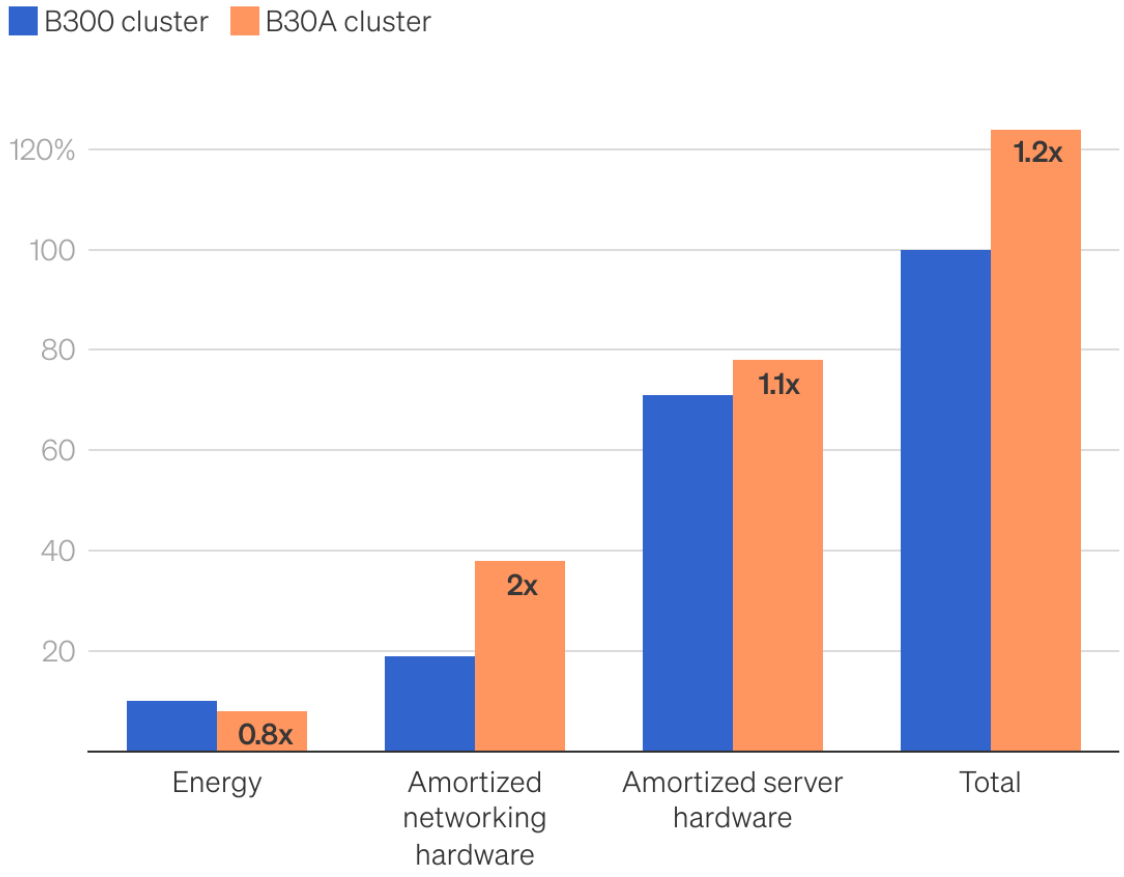
Attribute	B300 chip (flagship)	B30A chip (China version)	B30A compared to B300
Total processing performance (TPP)**	60,000†	30,000††	-50%
Memory bandwidth***	8.0 TB/s	4.0 TB/s‡	-50%
Retail cost	$51–55K	$20–25K	**51–64%** price reduction

* See Appendix 3 for a full table with other chip comparisons.
** Total processing performance (TPP) is the metric used by BIS to assess a chip's overall computing power, as the strongest proxy for AI training performance.
*** Most relevant for AI inference performance.
† Calculated from specs reported in NVIDIA's 2025 roadmap as: 1,100 petaflop/s for a 72-chip system at dense FP4, divided by 72 (to get per-chip performance), multiplied by 4 bits (FP4) to obtain TPP.
†† Based on reporting that the B30A would have half the processing power as the B300; final specifications may differ.
‡ Although we are using the 50% performance degradation estimate, the B300 uses four HBM3E stacks per GPU die, and four stacks can support up to 4.8 TB/s unless memory bandwidth is underclocked. Final specifications may differ.

Table: Authors

Figure 3: Relative cost of a B30A and B300-based AI training cluster

Expressed as % of the costs of a B300-based AI training cluster, for a B30A cluster with equivalent memory bandwidth and raw FLOP/s specifications.



See Appendix 2 for assumptions and calculations

Chart: Authors • Source: NVIDIA, Jarvis Labs, Guosheng Securities, Epoch AI.

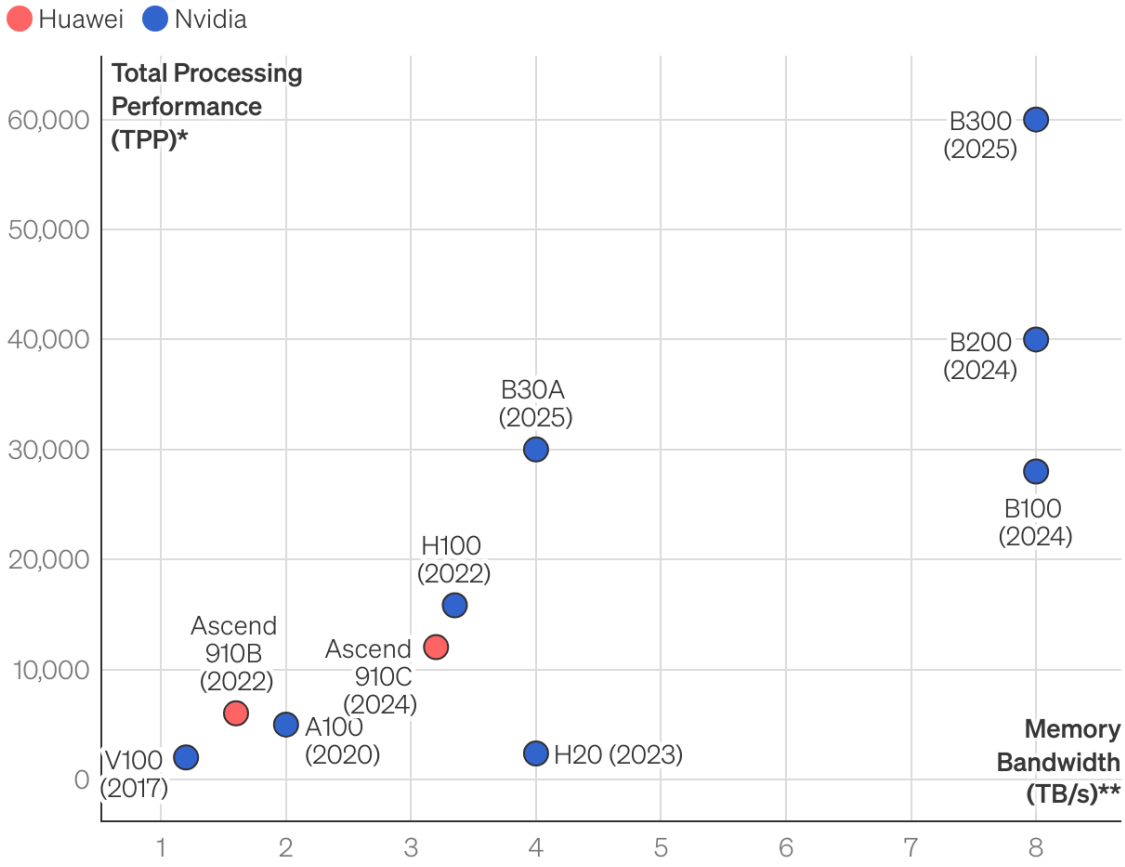
Pricing aside, the B30A is expected to be a much better chip than any models currently available in China. If reported specifications are correct, it would outperform the H20 — currently the top US AI chip available on the Chinese market — by more than 12 times. Moreover, it would exceed the United States' current export control performance thresholds, which determine which US AI chips are permitted to China, by more than 18 times.[11] This means that the B30A can

[11] We calculate this Figure by dividing the B30A's estimated 30,000 TPP by 1,600 TPP export control threshold, which equals 18.75x. We use the 1,600 TPP threshold because chips exceeding a performance density of 5.92, which the B30A does, become controlled if they exceed 1,600 TPP.

achieve almost double the total processing performance (TPP) of NVIDIA's prior flagship chip, the H100, and is 6 times better than its flagship chip before that, the A100.[12] Both these chips are currently banned from export to China.

Figure 4: AI chip performance

The B30A is speculated to have roughly 50% the performance of the B300 at 50% the price. This makes the B30A roughly as price-performant as the best AI chip on the market.



Speculated B30A price and performance are based on public reporting from Reuters and Tom's Hardware. Final price and performance may differ from what is shown here.
* Most relevant for AI training performance
** Most relevant for AI inference performance

Chart: Authors, adapted from Lennart Heim (Aug 2025)

It is possible that NVIDIA will instead choose to ship a downgraded Blackwell-generation chip that differs from the reported specifications of the B30A in name, system design, and/or performance. Still, if a downgraded chip made

[12] NVIDIA's H100 chip and A100 chip have TPPs of approximately 15,800 and 5,000 respectively — by contrast to the B30A's speculated TPP of 30,000 (see Appendix 3).

specifically for the Chinese market has anywhere near the reported capabilities or price-performance (performance obtained per dollar spent) of the B30A, it will be by far the best AI chip to have ever legally entered the Chinese market.

China cannot produce domestic chips that match the B30A

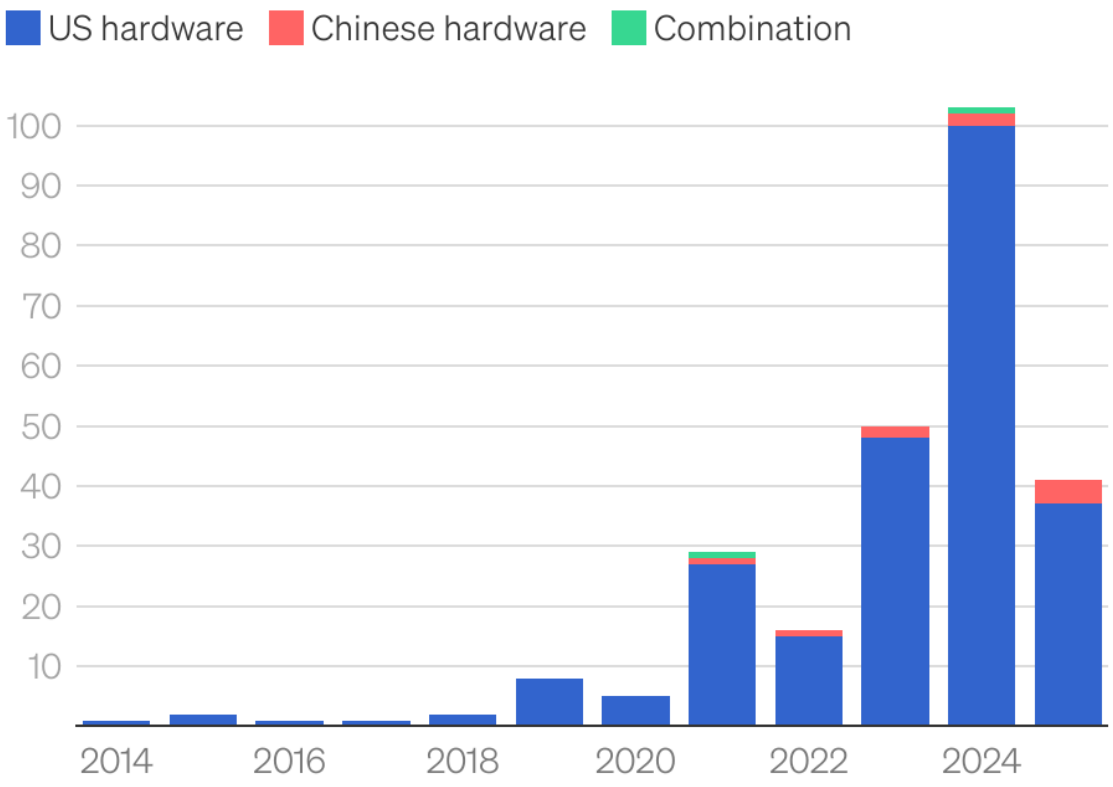
Without competitive domestic alternatives, Chinese companies will be strongly incentivized to bulk-order B30As. Based on technical specifications alone, the B30A would provide AI training and inference capabilities superior to China's own best chip, the Huawei Ascend 910C. The B30A is expected to provide more than double the processing power[13] — the main determinant of AI training performance — than the 910C, as well as more than 12 times greater processing power than the NVIDIA H20.[14]

The B30A would also offer at least 25% more memory bandwidth — needed to inference or "run" AI models — than the 910C.[15] In practice, the performance gap between the B30A and the Ascend 910C will likely be even wider: Huawei's chips are less reliable than US alternatives as they run on software that is prone to malfunction and hard to fix, especially when scaled up into larger clusters. As a consequence, Chinese AI models are trained overwhelmingly on American chips.

[13] Measured in FLOP/s.
[14] Although BIS issued an "is-informed letter" to inform NVIDIA of a license requirement for the export of H20 chips to China, the US government has announced that it would grant these export licenses.
[15] The B30A is speculated to have 2.2x the TPP as the Huawei Ascend 910C. The B30A's speculated 4.0 TB/s memory bandwidth is 25% higher than the Ascend 910C's 3.2 TB/s. Memory bandwidth is the main determinant of AI inference speed.

Figure 5: Chinese AI models by hardware origin

Number of AI models by Chinese developers* created using Chinese or US hardware.



*This includes AI models trained only by Chinese developers, and by Chinese developers in collaboration with external developers. All China+external collaborations in the dataset used US AI hardware.

Chart: Authors • Source: Epoch AI

IFP

Looking beyond individual chips, Huawei also markets its Ascend 910Cs in the form of a compute cluster architecture called the CloudMatrix 384, which contains 384 Ascend 910Cs. If NVIDIA is allowed to sell the B30A to China, it would likely sell it in both standard 8-GPU systems, similar to the DGX B300, and 72-GPU systems, similar to the GB300 NVL72. Since the CloudMatrix 384 contains more individual chips, it may have higher overall processing performance than a

hypothetical B30A server. But this comparison is misleading. GPU servers can easily be networked together into larger clusters. Therefore, the two important considerations for relative US-China AI compute competitiveness are price-performance (i.e., the performance per dollar)[16] and manufacturing scale (i.e., how many chips the company can produce).

On price-performance, there is no reason to think that CloudMatrix systems will be cheaper than equivalent NVIDIA systems on a performance basis. Existing data suggests the opposite: to start, the B30A's performance per purchase price (one aspect of price performance) is 2.9 times higher than the 910C (see Appendix 3). Furthermore, the CloudMatrix's processing performance per watt (another aspect of price performance) is 2.5 times lower than that of top NVIDIA Blackwell servers.

On scale, Huawei is deeply supply-constrained on Ascend 910Cs and, therefore, on the compute cluster architectures that contain them.

China cannot manufacture AI chips at scale

China will likely struggle to manufacture AI chips at scale for the foreseeable future. To produce AI chips domestically, China must manufacture two types of semiconductor dies: AI processor dies and HBM dies. All leading AI chips contain these two components. However, US and allied semiconductor manufacturing equipment (SME) export restrictions have significantly slowed China's indigenous production of both.

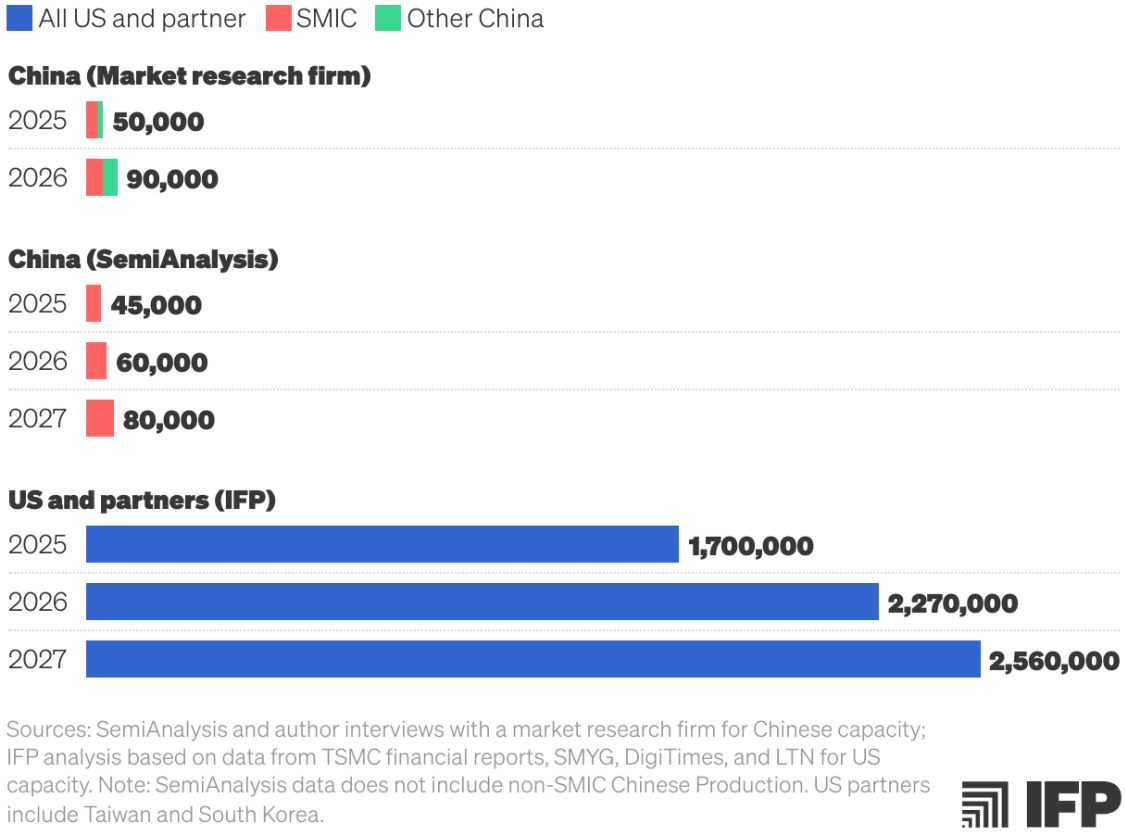
In 2025, US and partner manufacturing capacity for advanced AI processor dies (i.e., "7 nanometer" and more advanced logic wafer fabrication capacity) is about

[16] Price-performance takes into account both purchase price and use costs (due to energy consumption)

35 to 38 times that of China's, after adjusting for quality.[17] The United States is projected to maintain a strong advantage through 2026 and 2027.

Figure 6: US vs. China logic fab capacity at 7 nm and below

In wafers per month (quality-adjusted to SMIC 7 nm)



Sources: SemiAnalysis and author interviews with a market research firm for Chinese capacity; IFP analysis based on data from TSMC financial reports, SMYG, DigiTimes, and LTN for US capacity. Note: SemiAnalysis data does not include non-SMIC Chinese Production. US partners include Taiwan and South Korea.

IFP

In addition, a much higher percentage of China's chips have defects, making them unusable: estimates put China's yields for Huawei Ascend AI chips between 5%

[17] US and partner firms will begin to produce "2 nanometer" chips this year, which are three generations more advanced. For global ≤7 nm wafer capacity for 2025, 2026, and 2027, see analyses from SEMI. For China's ≤7 nm wafer capacity and China's leading chipmaker SMIC 7 nm capacity for 2025, 2026, and 2027, see analysis from SemiAnalysis. For another estimate of China's ≤7 nm wafer capacity, inclusive of SMIC and other Chinese chipmakers, IFP gathered data via private discussions with a market research and intelligence firm. For quality adjustments, we assume: all Chinese capacity remains at 7 nm due to export controls on EUV lithography tools, which are necessary for 5 nm production and beyond at commercially competitive yields; the 2025 percentage of US and partner ≤7 nm capacity at each node maps to TSMC's 2025 end-of-year expected breakdown of 35% 7 nm, 37% 5 nm, 28% 3 nm, and 7% 2 nm; and that 2026 and 2027 US and partner buildout is focused on 2 nm, resulting in a ≤7 nm capacity of 25% 7 nm, 27% 5 nm, 20% 3 nm, and 28% 2 nm. The analysis is derived from wafer capacity, wafer price, and revenues from TSMC financial reports, Toms Hardware, SMYG, DigiTimes, and LTN. We then normalize wafer capacity quality based on transistor count using the following transistor densities: SMIC 7 nm (89 MTr/mm²), TSMC 7 nm (114 MTr/mm²), TSMC 5 nm (141 MTr/mm²), TSMC 3 nm (212 MTr/mm²), and TSMC 2 nm (313 MTr/mm²).

and 20%, while NVIDIA Blackwell chips have achieved normal commercial yields of likely 60 to 80%. If Chinese yields are 12.5% (averaged across the two reports) and US yields are conservatively estimated at 60%, the effective US logic wafer manufacturing capacity advantage is about 170 to 180 times greater.

China is in an even worse position in manufacturing HBM, with virtually no production in 2025 and a 3,090x American advantage. Even in 2026, as China is expected to bring more production online, the United States is projected to make 70 times as much HBM as China; in absolute terms, China's 2026 HBM production can support a mere 275,000 Huawei Ascend 910C chips, equivalent to 55,000 B300-equivalents (B300–eq).[18]

Figure 7: US v. China HBM fab capacity

Estimates normalized to memory storage capacity measured in billions of gigabits

Sources: Trendforce for US and Korea; SemiAnalysis for China. US and partners represent three HBM providers: US-based Micron, and Korea-based Samsung and SK Hynix.

As a result, the United States and partners produce a vastly larger quantity of AI chips than China, as shown in Figure 7. US firms will produce 3.67 million B300-eq in 2025.[19] By comparison, five separate estimates from analysts place the production range of China's AI chip champion Huawei between 40,000 and 146,000 B300-eq in 2025, accounting for only 1 to 4% of total US production.[20] Only Bernstein provides data on non-Huawei Chinese AI chip designers, but they

[18] Trendforce estimates that global HBM shipments in 2025 and 2026 will reach 23.7 billion and 30 billion gigabits, respectively. Meanwhile, SemiAnalysis projects that CXMT is projected to make 40,000 HBM stacks in 2025 and 2.2 million HBM stacks in 2026. CXMT is planning to produce HBM3 in 2026. Assuming both 2025 and 2026-produced stacks are 12-Hi HBM3 stacks, they will each store 24 GB. Therefore, CXMT will produce 2.2 million stacks x 24 GB/stack x 8 bits/byte = 0.42 billion gigabits, while 40,000 stacks store 0.00768 billion gigabits. Therefore, the 2025 US advantage is 23.7 / 0.00768 = 3090x and the 2026 US advantage is therefore (30 - 0.42) / 0.42 = 70x.

[19] See Appendix 1 for the methodology behind this estimate.

[20] The five estimates are from SemiAnalysis, Bernstein, Bloomberg, Financial Times, and the Department of Commerce, each of which report production estimates of the Huawei Ascend 910B and 910C. We convert to B300-eq by normalizing for TPP.

estimate that all such companies combined account for only 50% of Huawei's volume, or only an additional 0.5 to 2% of total 2025 US production.

Figure 8: US and Chinese AI chip production, 2025 and 2026

Quantities are normalized to B300-equivalents. Each row is a separate estimate.

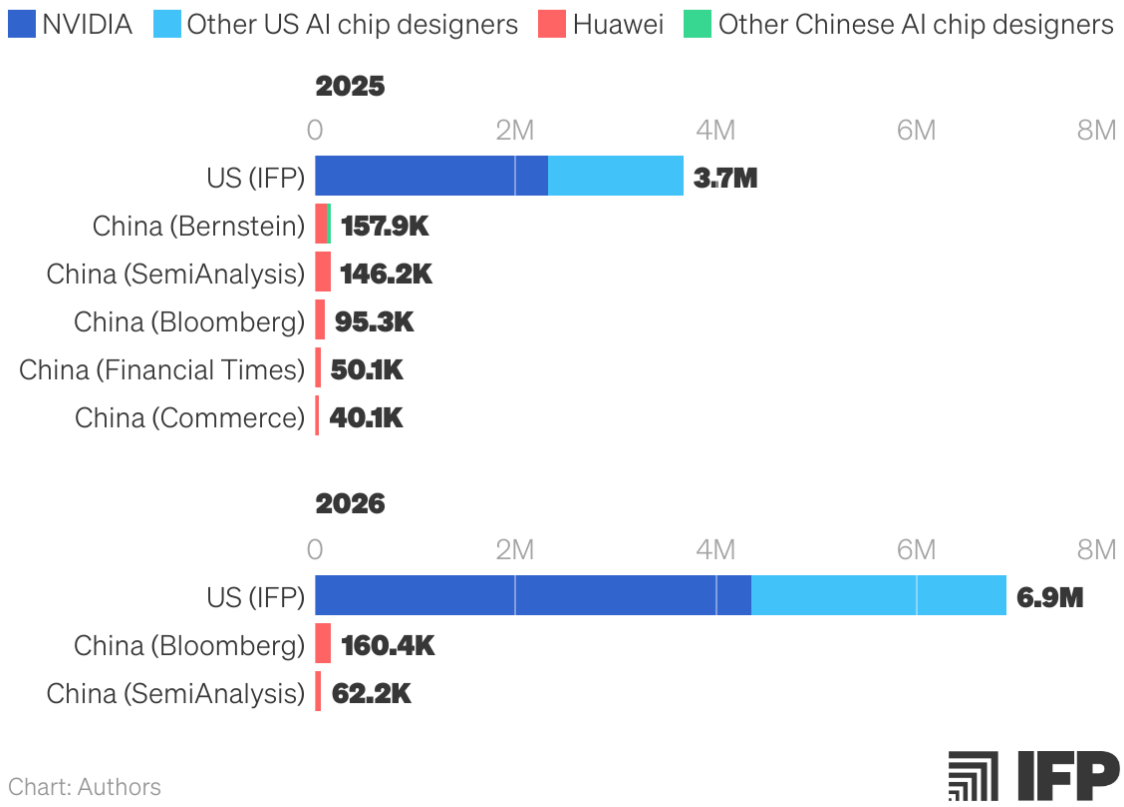


Chart: Authors

IFP

The US advantage will likely increase in 2026. US production will reach 6.89 million B300-eq, while Huawei will remain relatively stagnant, between 62,000 and 160,000 B300-eq, accounting for only 1–2% of total US production. SemiAnalysis, which has the most aggressive 2025 estimate of 146,000 B300-eq, predicts that Huawei's production will decline to 62,000 B300-eq in 2026 due to manufacturing bottlenecks, while Bloomberg expects an increase from 95,000 to 160,000 B300-eq.

Crucially, the Chinese production projections are generally based on manufacturing capacity modeling, not direct evidence that China has ever domestically manufactured Huawei AI chips. All known teardowns of Huawei AI

chips proved that they contain AI processor dies stockpiled from Taiwan Semiconductor Manufacturing Company (TSMC) and HBM from Korean chipmakers Samsung and SK Hynix before the US government tightened restrictions on these components.[21] Therefore, it is possible that China has not yet achieved domestic manufacturing of Huawei AI chips at production volume.

Chinese companies may face regulatory pressure to use domestic chips, but they would likely prefer to also use US AI chips if they are limited by domestic production bottlenecks. As discussed in the introduction to this paper, China's regulators are barring Chinese companies from buying some of NVIDIA's lower-end AI chips in a push to encourage adoption of domestic alternatives. Yet given its limited domestic production, China will not be able to sustain this ban while meeting local demand for AI chips. Going forward, Beijing may use such restrictions to promote the adoption of Huawei's AI chips initially, but then take a more liberal approach once Huawei's supply is exhausted. Alternatively, Beijing may be temporarily restricting NVIDIA chip imports in the hope of building leverage so that the US government will allow the sale of the B30A. In other words, Chinese restrictions on NVIDIA chip imports will likely be short-lived or reduced in scope if Beijing values not falling even further behind in the US-China AI competition.

Selling B30As to China would undermine the Trump administration's policy goals

Since the first Trump administration, American export control policy has sought to hold back China's frontier AI industry and military capabilities by restricting access to advanced AI chips and chipmaking equipment. As such, chips available in China perform well below those sold to US companies, securing the United States' total AI compute advantage[22] and thus its ability to train and widely deploy the most advanced AI models.

The second Trump administration has continued this strategy with its AI Action Plan. A White House AI advisor noted in October 2025 that the Trump

[21] Through 2024, Huawei illicitly manufactured 2.9 million Ascend 910B AI processor dies at TSMC, enough to power 290,000 B300-eq. The US foundry due diligence rule created heightened due diligence requirements on chipmakers like TSMC before servicing customers, making it more difficult for Huawei to repeat this feat. Huawei also stockpiled Korean HBM in 2024 in advance of the US government's China-wide HBM restrictions instituted in December 2024.
[22] Meaning that it owns a much larger total amount of computing power, measured in total quantity of quality-adjusted AI chips.

Administration plans to only approve exports of less advanced AI chips, such as the H20, in quantities that preserve the US AI compute advantage. Michael Kratsios, the Director of the White House Office of Science and Technology Policy (OSTP), also stated the Trump Administration has "kept in place" limits on China's access to the "most high-end chips."

If B30A sales went completely unrestricted (an extreme case), Chinese entities could match US AI computing capabilities at approximately the same cost and scale as their US counterparts. This would severely undermine existing chip export restrictions, rendering them functionally irrelevant. But even approved sales of fewer B30As would still run counter to the administration's expressed policy goals, both by eroding America's compute advantage and by providing the "highest-end" US chip technology in price-performance terms — just in different packaging.

In scenarios where supply is inelastic, or even other scenarios with partial inelasticity, allowing chip exports to China could reduce the global supply of AI compute available to US and allied customers if US and partner chipmakers — such as TSMC or HBM-makers Micron, Samsung, and SK Hynix — were to face future capacity constraints. This is because key inputs for advanced chips, such as HBM, advanced wafer capacity, and packaging, are rivalrous: producing less-advanced chip models like the H20 or B30A consumes the same production resources needed to manufacture cutting-edge chips such as the B300. If TSMC's production lines became supply-constrained, they would be forced to allocate limited resources between these competing products.

The same logic holds for supply constraints in the American and partner HBM industry. In these scenarios, fulfilling HBM orders for downgraded chips bound for China could compete with the production of more powerful chips for US and allied markets. This would not only weaken US customers' access to compute but also the global competitiveness of US cloud providers serving customers abroad. Moreover, permitting B30A exports would likely accelerate the adoption of these chips by Chinese cloud providers, eroding US cloud companies' market share among Chinese customers.

Figure 8 shows nine possible scenarios of B30A chip sales to China in 2026. The scenarios toggle between two assumptions: i) what percentage of US chips are sold or smuggled to China, and ii) the elasticity of supply of US chips. All scenarios assume that the US and US partners get a 75/25 split of US chips that are not sold

to China, in line with current trends.[23] They assume that in 2026, American chip production is 6.89 million B300-eq (see Figure 7 and Appendix 1) and Chinese production is 167,000.[24] The scenarios are grouped into four categories, depending on the administration's decisions on AI chip export controls.

Banned AI chip exports to China

1. **Banned exports with no smuggling:** US export controls apply strictly to all US chips, including the B30A. Since exports are banned, we assume China attempts to smuggle US chips. But China fails due to strong US enforcement mechanisms, such as those proposed in the proposed Chip Security Act. Supply elasticity plays no role since China acquires no US chips. In this scenario, US AI companies acquire **31x** as much AI compute as Chinese firms.
2. **Banned exports with smuggling, elastic supply:** US export controls apply strictly to all US chips. Due to poor export control enforcement, we assume China smuggles 4% of chips US chipmakers were planning to produce.[25] Supply of US chips is **elastic**, under the assumptions that i) there are no US and partner chip manufacturing bottlenecks and ii) Chinese companies do not access US cloud, therefore US cloud providers have no business to lose if Chinese customers now acquire US chips and use them locally. Because of this elasticity, smuggling accounts for **3.8% of US AI chip revenue.** The resulting US advantage in this scenario is **11.7x**.
3. **Banned exports with smuggling, inelastic supply:** US export controls apply strictly to all US chips and we again assume China **smuggles 4% of US AI chips**. Supply of US chips is **inelastic**, either because (i) US and partner chipmaking cannot scale up to satisfy increased Chinese demand without

[23] If US chip supply is inelastic and B30As are sold to China, then both US and partner purchases of AI chips go down proportionate to chips sold or smuggled to China. For example, if the US produces eight chips, and China buys none, then the US would get six and US partners would get two. But if China buys four chips, then the US would get three and US partners would get two. If US chip supply is elastic, then both US and partner purchases of AI chips remain the same regardless of chips sold or smuggled to China.

[24] 167,000 is the average of two 2026 estimates in Figure 7, multiplied by 1.5 to account for Bernstein's assessment that in 2025, non-Huawei Chinese AI chip designers contributed 50% as much compute as Huawei.

[25] We arrive at this estimate as follows: CNAS estimates that China smuggled 140,000 AI chips in 2024. Assuming these were H100s, these chips would represent 37,000 B300-eq. We estimate 2024 NVIDIA AI chip production as 554,000 B300-eq, based on data from Epoch. Given NVIDIA's 60% share in 2024 of TSMC's CoWoS, which roughly represents AI chip market share (see Appendix 1 for discussion), we find 2024 US AI chip production to be 923,000 B300-eq. Finally, if we assume China smuggles the same percentage of US AI chips in 2026 as it did in 2024, then China's 2026 smuggled volume would be 276,000 B300-eq, which is 4% of 2026 US chip production.

depriving US customers or (ii) Chinese customers migrate from US cloud to obtain US chips for themselves. The resulting US advantage is **11.2x**.

Conservative B30A exports

4. **Conservative exports, elastic supply:** China buys 26% of the chips NVIDIA was planning to produce before B30A approvals, consistent with NVIDIA's revenue share from China before US export controls on AI chips were instituted in 2022. Other US AI chip companies — such as Google and Amazon — generally only rent their chips to customers via the cloud, rather than selling directly to third-party customers, so this scenario assumes that only NVIDIA sells to China.[26] However, supply of US chips is **elastic,** so NVIDIA's actual revenue from China is 22% (**14% of total 2026 US AI chip revenue**). The US advantage is **4.0x**.
5. **Conservative exports, inelastic supply:** 26% of NVIDIA's revenue comes from China (**16% of US AI chip revenue**), as supply of US chips is **inelastic**. The US advantage is **3.3x**.

Baseline B30A exports

6. **Baseline exports, elastic supply:** China buys 45% of the chips NVIDIA was planning to produce before B30 approvals. Chinese firms aggressively stockpile chips, potentially subsidized by the Chinese government. This scenario is similar to the recent Chinese subsidized stockpiling of US and allied SME in 2024, when multiple US and allied firms derived well over 40% of their revenue from China despite significant US and allied export restrictions in place. However supply of US chips is **elastic**, so NVIDIA's actual revenue from China is 35% (**22% of US AI chip revenue**)**.** The US advantage is **2.4x**.
7. **Baseline exports, inelastic supply:** 45% of NVIDIA's revenue comes from China (**28% of US AI chip revenue**), as supply is **inelastic**. The US advantage is **1.7x.**

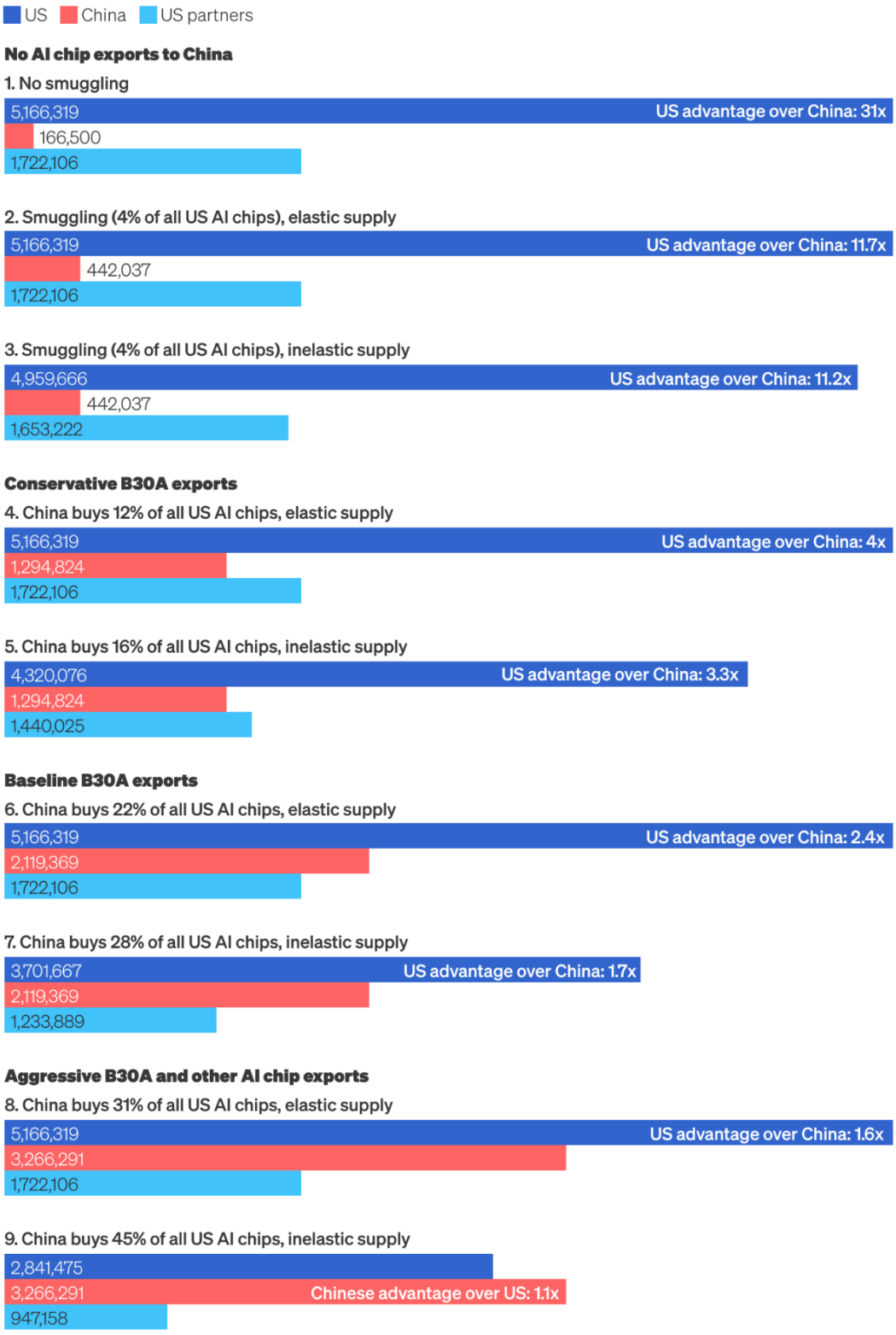
[26] 26% of NVIDIA revenue from China is equivalent to 16% of 2026 revenue from all US AI chipmakers combined (including NVIDIA). AMD shares a business model with NVIDIA in selling chips directly to third-party customers including in China, but their market share is small compared to NVIDIA, Google, and Amazon. Therefore, for simplicity we do not account for their sales to China.

Aggressive B30A and other US AI chip exports

8. **Aggressive exports, elastic supply:** The US AI chip industry as a whole — not just NVIDIA — sells to China 31% of the chips it was planning to produce before exports were approved. Even U.S. hyperscalers with in-house chips, who don't normally sell to third parties, decide to sell to Chinese companies or set up large datacenters in China. However, supply of US chips is **elastic**, so actual **US AI chip revenue from China is 31%.** The US advantage is **1.6x**.
9. **Aggressive exports, inelastic supply:** 45% of US AI chip industry **revenue** comes from China, as supply of US chips is **inelastic**. In this extreme scenario, China now establishes a **1.1x** advantage over the United States.

Figure 9: Assessing the impacts of B30A exports by scenario

Impacts measured in terms of total available AI compute to different actors, measured in B300-equivalents. Combines estimates of AI chip production by both US and Chinese firms, as well as estimates of AI chip smuggling.



Source: IFP analysis

Maintaining a strong AI compute advantage pays several dividends for the United States:

1. The United States can use more compute than China to train powerful AI models, thereby extending its advantage in AI model capabilities. Additionally, the United States would retain a strong advantage in its ability to explore new AI research paradigms, as most compute used for AI model development at frontier labs is likely now used for internal R&D rather than training publicly released models.
2. The United States can host far more frontier AI companies than China, since each requires access to substantial compute.
3. US companies can perform more inference on their models, including answering more AI model queries and, for each query, applying more inference-time compute to enable longer reasoning and more powerful autonomous agents, thereby enhancing AI model capabilities.

Minimizing the export of powerful AI chips to China is the best way to maximize the United States' AI compute advantage in the short term. The best way to maximize this advantage in the long term is to impose tighter controls on both SME and HBM.

Tighter controls on SME and HBM can prevent China from making advanced AI chips

Instead of selling B30As to decrease Chinese demand for their domestic chips, a more effective approach to limiting Huawei's chip supply is to halt China's domestic AI chipmaking capacity expansion. The United States and its allies could tighten country-wide export restrictions on SME, consistent with recommendations issued by the House Select Committee on the Chinese Communist Party. Such restrictions could include tightened bans on lithography, etch, deposition, inspection, and other categories of tools, in partnership with allies.

One of the core recommendations is to bar all exports to China of deep ultraviolet (DUV) immersion lithography tools needed to make advanced AI chips. Currently, exports of most DUV immersion lithography tools to China are restricted, but less advanced tools that are capable of 7 nm processor die production and advanced HBM production, such as the ASML NXT:1980, are still allowed for sale to most buyers in China.

China has no market share in these, or even in less advanced DUV tools. One analyst assesses that China is unlikely to gain meaningful market share or capability in domestic DUV immersion lithography tools in the next 10 years, meaning that export controls on the remaining DUV tools would remain highly effective in the future. Although Chinese firms are reportedly testing a 28 nm DUV immersion prototype, they must still scale two immense hurdles. First, translating a prototype into a commercial-grade tool typically takes years. Second, ASML sold its first commercial DUV immersion tools in 2003, long before it developed much more advanced 7 nm-capable immersion tools in time for TSMC's initial 7 nm production in 2018.

The US government can also pursue stronger measures to prevent the diversion of US and partner HBM, as it is a chief bottleneck to Huawei's AI chip production. One way to do so is to institute stronger due diligence requirements for HBM sales to ensure sales are limited to legitimate, known US and partner companies producing high-end chips that require HBM.

Appendix 1: Estimate of 2025 and 2026 US Production of AI chips

2025 US AI chip production is **3.67 million B300-eq**, or **13.9 million H100-eq**.[27] 2026 production is expected to increase to **6.89 million B300-eq**, or **26.1 million H100-eq**.

Estimating 2025 NVIDIA B300-eq production: NVIDIA's Fiscal Year 2026 (FY26) datacenter compute revenue is projected to be $154.7 billion.[28] NVIDIA's FY26 runs from February 2025 to January 2026 and is a reasonable proxy for revenue derived from chips produced in the 2025 calendar year. Nvidia's 2025 sales are reportedly 80% Blackwell chips, with the 20% remainder mostly H100s. For this estimate, the total processing performance (TPP) and TPP/$100 of each Blackwell chip are assumed to be 42,667 and 93.17, respectively, representing averages of the values for the B100, B200, and B300 in Appendix 3. This suggests an average Blackwell price of $45,796. Recovering the 80/20 Blackwell/H100 volume breakdown requires assuming 2.97 million sales of the theoretical average Blackwell chip and 743,000 H100 sales, with Blackwells accounting for 88% of NVIDIA's datacenter compute revenue. Normalizing to the TPP of the B300, NVIDIA's total 2025 production is 2.31 million B300-eq.

Estimating 2026 NVIDIA B300-eq production: The Installed stock of NVIDIA AI chips at the end of 2024 was 1.02 million B300-eq. Summing 2024 stock with 2025 production results in an end-of-2025 stock of 3.34 million B300-eq.[29] Assuming that NVIDIA stock increases by 2.3x by the end of 2026 to 7.68 million B300-eq, 2026 NVIDIA B300-eq production will be 4.34 million B300-eq.

Estimating 2025 and 2026 US B300-eq production: NVIDIA will use 63% of TSMC's CoWoS packaging capacity in 2025. Virtually all US AI chips use TSMC's CoWoS packaging, and virtually all of TSMC's CoWoS capacity is reserved for US AI chips. Furthermore, non-NVIDIA US AI chips — made primarily by Google,

[27] "US AI chips" refers to US-designed chips no matter where they are manufactured. The vast majority of US AI chips are manufactured by TSMC in Taiwan, but TSMC is now also producing NVIDIA Blackwell chips in Arizona. Production volumes are also distinct from deployment volumes.
[28] Nvidia's FY26 H1 (Feb. to Jul. 2025) revenue was $90.805 billion with datacenter compute revenue of $67.999 billion. Nvidia projects revenue of $54 billion in FY26 Q3. Analysts expect FY26 revenue of $206.6 billion. If datacenter compute remains the same percentage of total revenue in FY26 as it was across FY26 H1, then FY26 datacenter compute revenue would be $154.7 billion.
[29] This calculation does not adjust for some chips reaching end-of-life, as doing so would only change the results by a small percentage.

Amazon, AMD, and Intel — likely have a similar ratio of TPP to CoWoS capacity as NVIDIA chips. Thus, dividing the NVIDIA B300-eq estimate by 0.63 generates the total 3.67 million 2025 US B300-eq estimate. Assuming a similar NVIDIA CoWoS share in 2026, total 2025 US B300-eq production will be 6.89 million.

Table 2: 2025 and 2026 US AI chip production

Chip category	2025 Chip volume	Price	TPP	B300-eq
NVIDIA Blackwell	2,970,000	$45,796 (avg.)	42,667 (avg.)	2,110,000
NVIDIA H100	743,000	25,000	15,824	196,000
NVIDIA 2025 total				2,310,000
NVIDIA 2026 total				4,340,000
US 2025 total				3,670,000
US 2026 total				6,890,000

Table: Authors

IFP

Appendix 2: B30A cluster analysis

Here, we analyze the amortized cost of a cluster of equivalent FLOP/s and memory bandwidth performance using B30As vs NVIDIA's most powerful chip, the B300. Note that these numbers should be taken only as high-level estimates.

We make the following assumptions:

1. The B30A has half the peak FLOP/s and memory bandwidth of the B300 (See Appendix 3).
2. In each case, we are using the 8-GPU server configuration.[30]
3. 8-GPU B300 servers draw 2.4 times as much power as 8-GPU B30A servers.[31]
4. An 8-GPU B30A server will be approximately 55% the price of an 8-GPU B300 server.
    a. Individual B300 cards are reportedly around $53K, while individual B30A cards will reportedly be around $22.5K (see Appendix 3).
    b. While pricing data for 8-GPU configurations of B300s and B30As is not available, we can look at server pricing data for the H200 and its downgraded alternative, the H20. 8-GPU H200 servers are around 10x more expensive than H200 cards,[32] while 8-GPU H20 servers are 13x more expensive than H20 cards.[33] This difference in multiples makes sense, given the relatively fixed cost of other server components relative to price differences between the upgraded and downgraded versions of the cards.
    c. Applying the same multiples yields $530K for 8-GPU B300 servers, and $293K for 8-GPU B30A servers — 55% the price of the B300 version.

[30] NVIDIA offers an 8-GPU server for the B300, known as the NVIDIA DGX B300. It is currently unknown which B30A server configurations will be offered. However, based on historical precedent, an 8-GPU configuration is highly likely — 8-GPU configurations have historically been designed both for flagship NVIDIA AI GPU offerings (H100/200, A100) and for China-specific versions of these chips (H20/H800, A800).
[31] According to a Chinese source 8-GPU H20 servers have a power draw of 4.2kW, whereas 8-GPU H200 servers have a max power draw of 10.2kW – 2.4x the H20 server.
[32] Price data for 8-GPU H200 servers from the same time period (May 2025) puts them at $400K-500K, while individual H200 cards were reportedly $40K-55K.
[33] Chinese financial services company Guosheng Securities estimates that the price of 8-GPU H20 servers as of March 2025 was ¥1.1 million, or around $150K USD. Individual H20 cards are reportedly $10-13K (see Appendix 3).

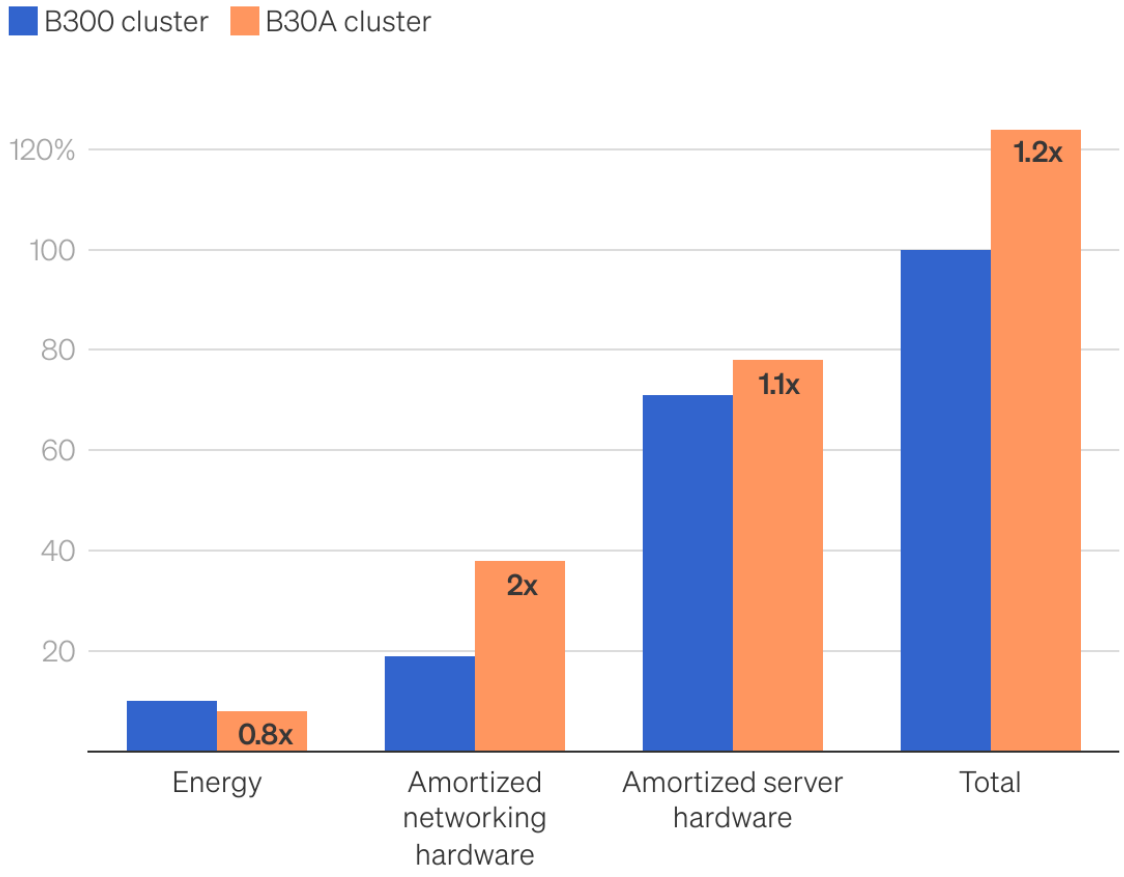
5. In each case, we use a variant of a fat-tree network topology (typical for AI training clusters), where networking costs scale linearly with the number of accelerators.

Taking these assumptions together, a B30A cluster with equivalent FLOP/s and memory bandwidth to a B300 cluster will involve 2x the cost for networking hardware, 0.8x the cost for energy, and 1.1x the cost for server hardware. Based on data from Epoch AI analyzing 41 clusters used to train frontier models, we can derive overall B30A-B300 cost differences by further assuming that 71% of B300 amortized cluster capital expenditures is server hardware costs, 19% is server-to-server networking hardware costs, and 10% is energy costs.[34] This yields an overall amortized cost difference, including servers, networking, and energy, of 1.24x for a B30A cluster relative to a B300 cluster, with equivalent peak FLOP/s and memory bandwidth.

[34] Estimate from Epoch AI, based on a three layer switching network. We assume that these percentages are most relevant to B300s, as the clusters Epoch tracks are those used to train frontier models, and thereby generally used the best available frontier hardware. This estimate also assumes that the amortization period for B30A and B300 servers is the same – i.e. that the hardware has the same operational lifetime in either case.

Figure 3: Relative cost of a B30A and B300-based AI training cluster

Expressed as % of the costs of a B300-based AI training cluster, for a B30A cluster with equivalent memory bandwidth and raw FLOP/s specifications.



See Appendix 2 for assumptions and calculations

Chart: Authors • Source: NVIDIA, Jarvis Labs, Guosheng Securities, Epoch AI.

IFP

A further consideration for cluster performance is *utilization* – how much of the peak FLOP/s and memory bandwidth performance a cluster is actually able to achieve in practice, given the algorithms used to distribute training across AI chips within a cluster. There is little reason to think that utilization for B30A clusters will be substantially better or worse than for B300A clusters, given the ratio of FLOP/s to memory bandwidth is similar for each accelerator (implying a similar parallelization strategy), and the fact that US frontier labs use a variety of different chips for AI training, with peak FLOP/s and memory bandwidth varying to similar degrees as the B30A to B300.

Appendix 3: Full chip performance and price table

Table 3: AI chip performance and prices

	Designer	Model	TPP	Price (USD)	Price (median estimate)	TPP/ $100	Memory bandwidth (TB/s)	TB/s /$10k	Release date	Sources for TPP	Sources for price	Sources for memory bandwidth
	NVIDIA	B30A	30,000	20–25K	22,500	133.3	4	1.78	Dec '25	DCD Reuters	Toms Hardware	Yahoo Finance
	NVIDIA	H20	2,368	10–13K	11,500	20.6	4	3.48	Nov '23	Toms Hardware	Reuters Toms Hardware	SemiAnalysis
	NVIDIA	B300	60,000	51–55K	53,000	113.2	8	1.51	Aug '25	NVIDIA Datasheet	Glen Lockwood (4M/72)	Toms Hardware
	NVIDIA	B200	40,000	45–50K	47,500	84.2	8	1.68	Dec '24	NVIDIA	North Flank	NVIDIA Yahoo Finance
	NVIDIA	B100	28,000	30-35K	32,500	86.2	8	2.46	Dec '24	SemiAnalysis	HSBC	SemiAnalysis
	NVIDIA	H100	15,840	25K	25,000	63.4	3.35	1.34	Sep '22	NVIDIA	Modal	NVIDIA
	NVIDIA	A100	4,992	15–17K	16,500	30.3	2	1.21	Mar '20	SemiAnalysis	North Flank	NVIDIA
	NVIDIA	V100	2,000	9.5K	9,500	21.1	1.2	1.26	Jun '17	NVIDIA	Hyperscalers.com	NVIDIA
	Huawei	Ascend 910C	12,032	25–28K	26,500	45.4	3.2	1.21	Oct '24	Huawei	Caixin EE World	SemiAnalysis
	Huawei	Ascend 910B	6,016	17K	17,000	35.4	1.6	0.94	Oct '22	Huawei	Reuters	CSET

Table: Authors

IFP