FEBRUARY 2025

The AI Diffusion Framework

Securing U.S. AI Leadership While Preempting Strategic Drift

By Barath Harithas

Introduction

In one of its final acts, the Biden administration introduced what may be the most ambitious exercise of technological statecraft in modern history—the **Framework for Artificial Intelligence Diffusion** (AI Diffusion Rule). What was once a simple line drawn around China and a small handful of countries has turned into a sweeping, interventionist framework that dictates not just who can import advanced semiconductors and how many but where they can be deployed, under what conditions they must be secured, and even how AI capabilities can be shared.

Unpacking the U.S. Calculus

In some ways, the framework is a carefully modulated equation, attempting to balance multiple, often competing variables in a single formula. On the one hand, the United States needs to prevent the diversion of advanced graphics processing units (GPUs) to China, increasingly through **third countries**, to ensure that China never reaches escape velocity in the AI race. **DeepSeek** may have tilted the board, but the equation holds—China still faces stubborn compute bottlenecks in **inference scaling** and deployment.

On the other hand, the **exponential growth** in compute demand makes it neither practical nor expedient to limit AI infrastructure to U.S. soil. American and allied firms must expand overseas, but every new deployment introduces an additional vector for diversion. This tension is further complicated by market dynamics. U.S. and allied hyperscalers and neocloud providers operate on economic logic, not national security priorities. They will naturally gravitate toward locations where power is cheap, infrastructure hurdles are minimal, and financial incentives sweeten the deal. But for the United States, these same AI-friendly jurisdictions may not always be reliably aligned with its priorities, nor can they be easily insulated from Beijing's reach.

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

The Three-Tiered Framework: First-Class, Coach, and No Ticket

The AI Diffusion Rule introduces a three-tiered framework for AI access. Tier 1 (T1) countries ride first-class, with near-frictionless access to advanced GPUs. Tier 2 (T2) nations, by contrast, are stuck in coach—allowed on the train but kept at a distance from the front cabin—facing strict caps on GPU access. This is an admission of the **practical limits** of enforcement. Rather than trying to police every node, the rule seeks instead to depressurize the flow of compute globally so there is less surplus floating around that could eventually find its way to China. T2 companies that sever supply chain ties with China and meet stringent compliance standards, however, can import and deploy more GPUs. Together, these controls aim to funnel advanced GPUs into narrow, auditable corridors. The stated objective, however, is clear: to keep T2 countries at least a **generation behind the frontier**, which the rule defines as 10^{26} FLOP— an **order of magnitude** beyond the current capability threshold of 10^{25} FLOP, where models like GPT-4, Gemini 1 Ultra, and Claude 3 Opus operate. In other words, an upgrade to business class is possible, but first-class is off-limits. Meanwhile, Tier 3 (T3) countries like China are left on the platform, locked out entirely, forced to find their own slower track.

A Firewall with No Backdoors: A Full-Stack Approach to Diversion

The United States is also taking a full-stack approach to preventing diversion. The framework goes beyond physical hardware controls to close gaps in **cloud computing access** and **cyber theft of model weights**, which are the final mathematical parameters that encode an AI system's learned capabilities. It is not just about controlling where compute flows, but ensuring that wherever it is deployed, it remains securely barricaded and tethered to U.S. regulatory oversight.

The Collateral Damage on Allies and Partners

The AI Diffusion Rule is a technocratic tour de force, but at what cost? Observers tend to stall on the obvious, first-order questions. Can Huawei and domestic firms backfill for NVIDIA? Not in the short term. Huawei remains at least one to two generations behind both in **GPU performance** and **fabrication capacity**. Will frontline NATO allies like **Poland and the Baltic states** feel slighted? Yes, and bitterly so.

But while the first cracks are easy to spot, the real risk is what follows. How will T2 countries metabolize the rule, adapt to its constraints, and respond in the long run? Countries have now seen the U.S. export control regime grow by the better part of a thousand pages over 30 months and witnessed the once modest and polite fiction of "small yard, high fence" turn into a global regulatory ordinance.

Moreover, where countries once abided the mailed fist when it was directed against China, they may now recalculate, recognizing that the same grip, once reassuringly distant, can just as easily tighten around their own throats. Well-capitalized AI aspirants in T2 will bristle at the fact that their AI ambitions depend on Washington's goodwill and that they are being deliberately kept a generation behind the frontier. The presumption that they will meekly accept dependency or permanent subordination is naive.

The default assumption is that T2 countries will either seek to climb into T1 or drift into China's orbit. They will of course attempt the former. Political opportunism demands nothing less. But this binary framing overlooks a third, and perhaps more disruptive, possibility: A bloc of T2 nations may

collectively chart a "Third Way," not to rival U.S. technological primacy, but to carve out a functional, mid-tier AI stack that provides sufficient strategic autonomy. "De-risking" is often framed as a Western response to China, but T2 nations may now apply the same logic to Washington. The risk for the United States is not open rebellion. It is the slow, deliberate, and accumulative drift away from U.S. control.

Another overlooked fracture in the AI Diffusion Rule is how it may incentivize open-source migration as an escape hatch. If an advanced model is trained on U.S.-controlled hardware, its weights are also now subject to U.S. export controls, unless they are open-sourced and made publicly available. This creates an incentive for T2 nations to lean into open-source AI development. Optically, while the United States builds a walled garden, China, through DeepSeek and other initiatives, positions itself as the patron of open-source development. Ironically, the AI Diffusion Framework, meant to lock in American advantage, may instead midwife the very outcome it sought to prevent—an alternate AI stack and increased open-source development where China, as its most prolific contributor, emerges as the de facto leader.

A common rejoinder is that the AI Diffusion Rule only affects a handful of countries, since most global AI data center capacity is **already concentrated** in the United States and T1 nations. But this grossly understates the fallout for economies like Malaysia, which was on track to become the world's **third-largest data center country by 2026**, as well as India, Indonesia, Saudi Arabia, Singapore, and the United Arab Emirates (UAE)—nations that had staked their AI ambitions on expanded compute access.

This reasoning also rests on an implicit assumption that T2 countries do not need as much compute, because few were ever serious AI contenders. Leading AI research only takes place in a **narrow corridor of countries** (e.g., China, France, the United Kingdom, and the United States). But saying "you were never going to win anyway" is a significantly different proposition from saying "you are not allowed to try." No one tells a kid in the United States that they cannot be president. Empire operates on the same principle. Its benefits must always seem within reach, even to those who will never grasp them. Otherwise, it will breed resentment, not cooperation.

Of course, the United States may believe that the risk of AI proliferation is so severe, and the timeline to artificial general intelligence **so short** (**2–3 years**), that the cost of alienating T2 nations is an acceptable trade-off. But if the United States is trying to consolidate long-term control over the global AI stack, this approach may be too heavy-handed and self-defeating.

Understanding the Tiered Framework

T1: THE INNER CIRCLE

T1 comprises 18 countries and reads like a roll call of long-standing allies:

- **Five Eyes** intelligence partners (Australia, Canada, New Zealand, and the United Kingdom);
- Close Western/**NATO** allies (Belgium, Denmark, Finland, France, Germany, Ireland, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom); or
- Semiconductor heavyweights (Taiwan, the Netherlands, Japan, and South Korea).

But like all alliances, the omissions are as telling as the inclusions. Not all of "Old Europe" or NATO made the cut—Greece and Portugal are missing. More strikingly, **Poland**, **Latvia, Estonia, and Lithuania** are conspicuously absent, despite their vigorous support for U.S. security initiatives, especially in the wake of Russia's invasion of Ukraine.

This partial listing underscores that unofficially, there is a secondary filter. It is not enough to be a stalwart ally on paper; to guarantee lockstep compliance, the United States wants partners whose re-export controls and enforcement frameworks mirror its own. For Washington, even otherwise-dependable allies may fall short if they lack the institutional capacity, enforcement rigor, and willingness to guarantee strict adherence, or if they are viewed as potential diversion risks.

T2: ECLECTIC MIDDLE GROUND

The lion's share of nations fall into T2, a catch-all group that lumps together an eclectic mix of countries with vastly different levels of trust, capacity, and AI ambitions. India, Israel, Singapore, and Switzerland are placed alongside Yemen. There are several reasons why countries may have ended up in T2. Some, like Saudi Arabia and the UAE, have emerged as **key investors in AI** but remain outside the Western intelligence-sharing orbit. Others, such as India, are forging close strategic and trade ties with the United States but have a legacy of nonalignment. Switzerland, with its long-standing tradition of **fierce neutrality**, fits a similar mold. Meanwhile, countries in Southeast Asia and those in Eastern Europe that are officially close U.S. partners have been flagged as **diversion risks**.

T3: THE USUAL SUSPECTS

T3 is a familiar roster of U.S. **arms-embargoed countries**—China, Iran, North Korea, and Russia, joined by the likes of Burma (Myanmar), Syria, and Venezuela.

T1 Access and Restrictions

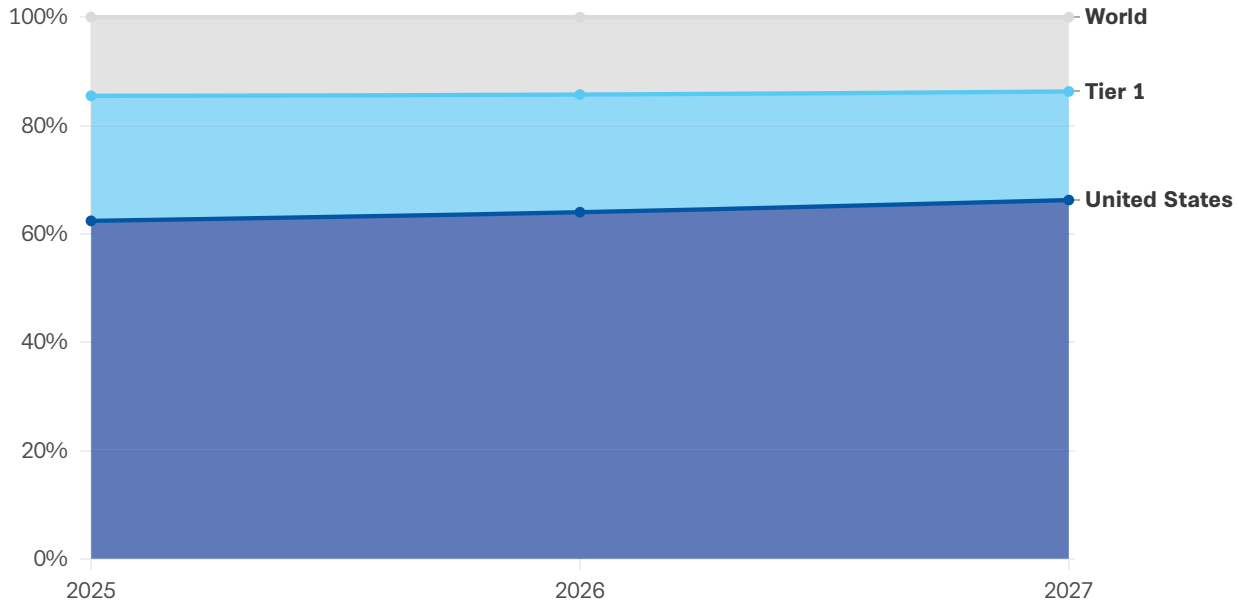
T1 countries will have unrestricted access to advanced GPUs. Companies headquartered or with an ultimate parent in a T1 country can deploy as much computing power as they want within T1 nations. Through a one-time Universal Validated End User (UVEU) authorization, they can also deploy GPUs in T2 countries, albeit within certain limits. Beyond access to hardware, the rule also ring-fences the training of frontier models in T1 countries.

However, even T1 status comes with conditions. U.S.-headquartered companies must keep at least 50 percent of total AI compute in the United States, 75 percent in T1 countries overall, and no more than 7 percent in any single T2 country.

The numbers are striking. But based on data from the AI and semiconductor research firm **SemiAnalysis**, the bulk of AI data center capacity is already concentrated in the United States and T1 countries—over 50 percent in the United States alone and more than 75 percent across T1 nations. SemiAnalysis **further projects** 10+ gigawatt (GW)-scale campuses coming online in the United States over the next two years, reinforcing its lead. Figure 2 shows the stark regional disparities in planned data center capacity through 2027, with North America dominating global capacity—roughly four times larger than either Europe and the Middle East or Asia Pacific, while Latin America trails behind for now.

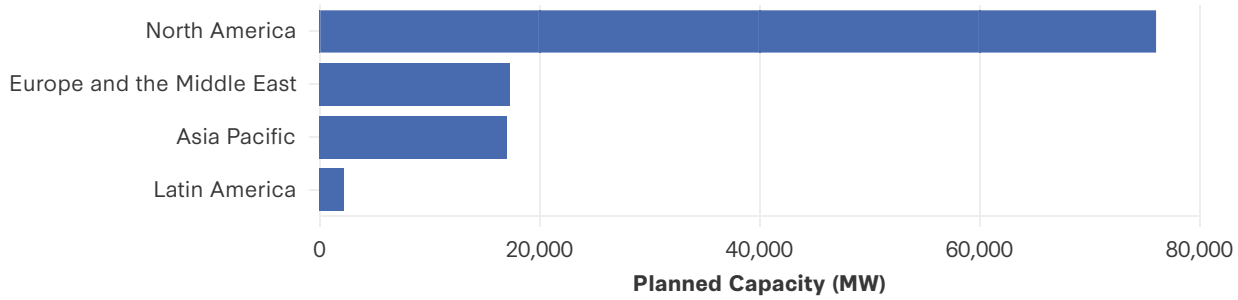
To qualify, however, companies must meet **19 separate certification and policy requirements**, covering everything from GPU installation reporting and transit security to personnel vetting,

Figure 1: Global AI Data Center Capacity: United States vs. T1 Countries vs. World



Source: CSIS analysis; "Datacenter Industry Model," SemiAnalysis, https://semianalysis.com/datacenter-industry-model/.

Figure 2: AI Data Center Capacity Forecast by Region in 2027



Source: CSIS analysis; "Datacenter Industry Model," SemiAnalysis, https://semianalysis.com/datacenter-industry-model/.

monitoring, recordkeeping, and even sanitization and disposal procedures. The strictest requirements revolve around cybersecurity controls for model weights, mandating rate-limiting on data access and restricting interactions to narrowly defined application programming interfaces (APIs) to prevent rapid exfiltration and unauthorized probing of models through interfaces.

In addition, UVEUs must follow "know your customer" (KYC) and red flag guidance, including notifying authorities when training an advanced model. The primary concern is subsidiaries conducting large-scale training runs above the prescribed threshold (i.e., 10^{26} FLOP) and then transferring model weights to China. Collectively, these safeguards are designed to prevent GPU diversion, unauthorized access, and the theft of model weights.

For companies that clear these hurdles, the payoff is substantial: a near-universal hall pass, granting broad authorization to scale AI infrastructure worldwide. UVEUs will also be exempt from the restrictions on model weights—which, under the rules of the Bureau of Industry and Security (BIS), otherwise require a license to export, re-export, or even transfer within a country—provided the destination is a T1 country.

T2 Access and Restrictions

BIS has imposed country-level caps on advanced GPUs for T2 nations, measured by **total processing performance** (TPP), a metric quantifying aggregate compute capacity.

However, TPP is abstract and lacks immediate real-world clarity. As such, a conversion method is often employed to translate TPP into a more tangible benchmark—**NVIDIA's H100** equivalents, the current industry standard for AI workloads.

Each T2 country receives a fixed allocation of 49,901 H100-equivalent GPUs through 2027. This is not an annual quota—once exhausted, whether in 2025 or 2026, no additional GPUs can be imported until after 2027. To accommodate smaller-scale needs, BIS allows T2 entities to acquire up to 1,699 H100-equivalents without a license, though they must notify BIS. These purchases do not count toward the 49,901 H100-equivalent installed base cap, offering flexibility for low-volume imports.

THE SHRINKING REALITY OF 50,000 AND 1,700 GPU CAPS

At first glance, an allocation of ~50,000 H100-equivalent GPUs, which would **cost roughly $1.25 billion**, appears substantial—roughly more than double the **compute used to train GPT-4**, which was trained on roughly **25,000 less-performant A100 chips**. However, the H100 will soon become a previous-generation chip, set to be replaced in 2025 by NVIDIA's newer **B200** and **GB300** GPUs. When the permitted TPP is converted into these next-generation units using **numbers from SemiAnalysis**, the effective number of GPUs shrinks—from 50,000 H100s to 21,987 B200s, and then 13,192 GB300s.

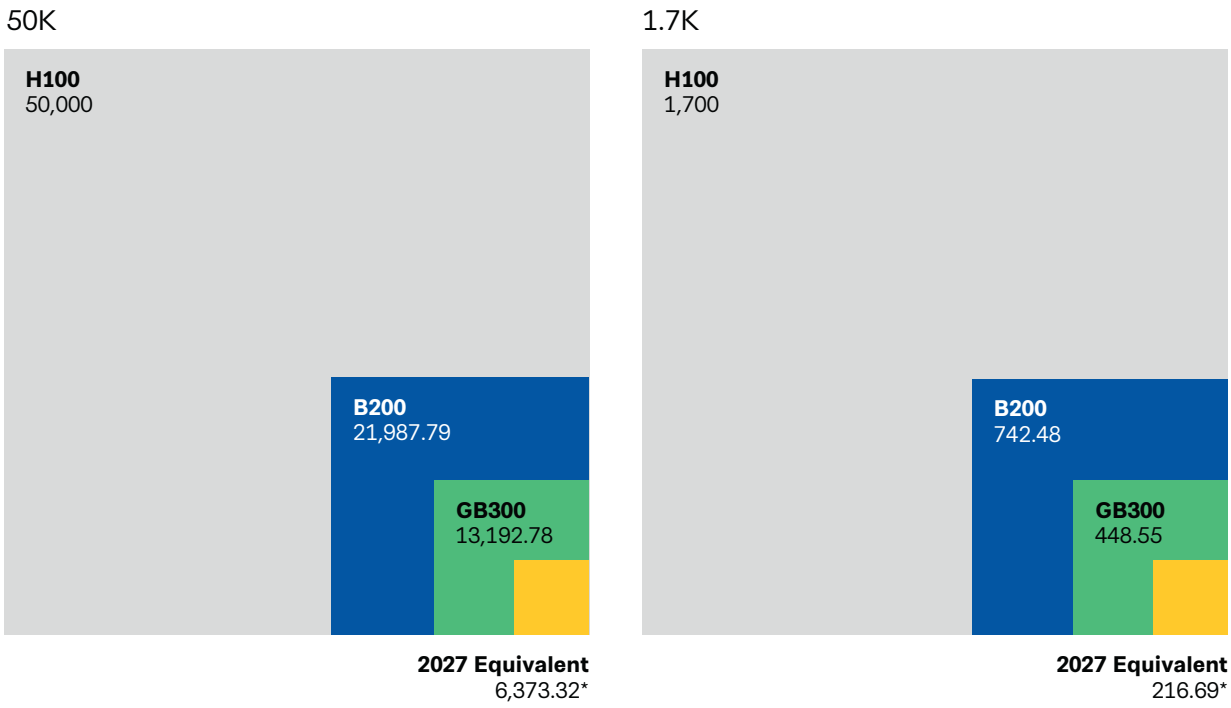
This downward trend accelerates when accounting for hardware improvement rates. According to **Epoch AI**, ML hardware performance has doubled every 1.9 years, meaning that by 2027, a 13,192 GB300 allocation will be effectively reduced by half. Similarly, the seemingly permissive ~1,700 H100 threshold for small purchases translates to **fewer than 500 GB300 GPUs** and an even smaller number of 2027-equivalents. These calculations reveal how quickly nominally generous caps become more restrictive, limiting the compute available to T2 nations in practice. It is worth noting that future hardware performance is uncertain. NVIDIA's generational leaps might yield larger or smaller performance gains than projected and come sooner or later than expected.

HOW T2 COUNTRIES CAN INCREASE THEIR CAPS

Companies from T2 countries looking to deploy a significant number of GPUs can apply for a National Validated End User (NVEU) authorization, provided they meet stringent security requirements.

By so doing, they receive an export license allowing a **phased-in installed base** of up to approximately 100,000 H100 equivalents by the end of 2025, 270,000 by the end of 2026, and 320,000 by the end of 2027. Again, these are not trivial numbers. 100,000 H100s is three times larger than the compute capacity of **xAI's Colossus supercomputer facility** in Memphis, Tennessee, the largest U.S. GPU cluster in 2024.

Figure 3: 50k and 1.7k H100 Equivalents over Time



* Estimate based on author's calculations.

Source: CSIS analysis; Dylan Patel et al., "2025 AI Diffusion Export Controls – Microsoft Regulatory Capture, Oracle Tears, Impacts Quantified, Model Restrictions," SemiAnalysis, January 15, 2025, https://semianalysis.com/2025/01/15/2025-ai-diffusion-export-controls-microsoft-regulatory-capture-oracle-tears/#:~:text=In%20all%20cases%2C%20the%20AI,-models%20of%20this%20size%20with.

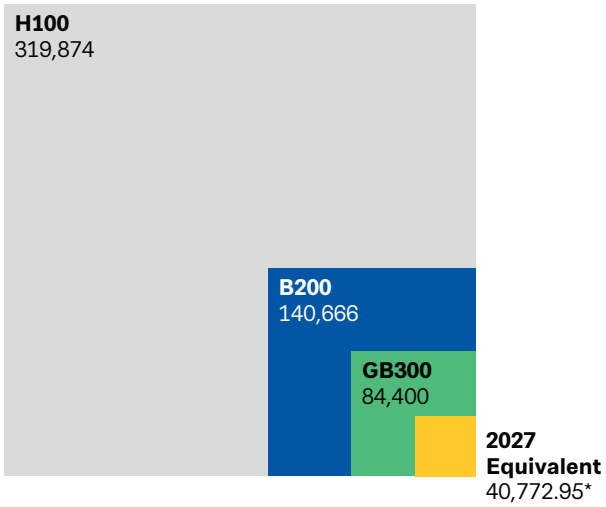
However, as discussed earlier, these caps shrink dramatically when accounting for hardware improvements. When the allocation is converted, the 320,000 H100 equivalents allowed drops to 140,666 B200 equivalents and 84,400 GB300 equivalents. Factoring in hardware improvements—i.e., doubling every 1.9 years—this number shrinks even further.

In order to obtain NVEU status, T2 applicants are "strongly encouraged" to first secure a government-to-government assurance before seeking approval from four federal agencies—the Departments of Commerce, State, Defense, and Energy. While UVEU applicants must also clear this multi-agency review, T2 firms face **structural disadvantages** compared to T1 hyperscalers and neoclouds, which already operate within established compliance frameworks, observability systems, and security protocols, making many of the requirements routine. In addition, T2 NVEU applicants will also face heightened scrutiny in demonstrating efforts to sever supply chain dependencies with China. While not all countries are equally exposed, many have **integrated** Chinese networking equipment and hardware into their infrastructure due to cost advantages, making compliance both technically complex and costly.

T3 Access and Restrictions

There will be a presumption of denial for all T3 countries—in regulatory parlance, that means an automatic "no" for any export license applications involving T3 entities. None of the streamlined "VEU"

Figure 4: 320k H100 Equivalents Over Time



* Estimate based on author's calculations.

Source: CSIS analysis; Dylan Patel et al., "2025 AI Diffusion Export Controls – Microsoft Regulatory Capture, Oracle Tears, Impacts Quantified, Model Restrictions," SemiAnalysis, January 15, 2025, https://semianalysis.com/2025/01/15/2025-ai-diffusion-export-controls-microsoft-regulatory-capture-oracle-tears/#:~:text=In%20all%20cases%2C%20the%20AI,models%20of%20this%20size%20with.

routes (universal or national) apply. In principle, a T3 company cannot just waltz through licensing hoops—it is effectively shut out.

The Uneven Impact of the AI Diffusion Rule

The impact of the rule will, unsurprisingly, be uneven. While it is tempting to cast T1 as the "winners" and T2 as the "losers," this would be an oversimplification. Even within each tier, the outcomes will be far from uniform—some T1 countries will gain disproportionately, while others in T2 will feel the brunt of the restrictions more acutely.

Surprisingly, some T2 countries may benefit under the new rules. For instance, entities in parts of the **Middle East and Central Asia**—where any chip purchase previously required an export license—now have pathways to access controlled chips without a license, provided they stay within the bounds of low-volume purchase pathways (LPPs) or individual country allocations. Countries like the UAE, which spent months negotiating deals such as the Microsoft-G42 partnership and has already secured a **memorandum of understanding** with the United States, will welcome the added clarity and are well-positioned to obtain an NVEU status. Conversely, other T2 nations that previously faced no restrictions will now find their access significantly tightened.

Broadly speaking, the new rule is likely to be a **net negative for T2 nations** with significant AI data center capacity pipelines, particularly if that capacity is not slated for use by U.S. hyperscalers.

Biggest Losers

Based on **SemiAnalysis data**, Malaysia will be the hardest-hit country. Its data center capacity has been surging, rising from just 100 megawatts (MW) in 2023 to a projected ~3.5 GW by 2027, positioning it to become the world's third-largest data center country, behind the United States and China, by 2026.

Nearly half of Malaysia's projected 2027 capacity is optimized for cutting-edge NVIDIA AI accelerators, with facilities capable of supporting power densities of up to **130 kilowatts (kW) per rack**. This has drawn major investments, including NVIDIA's **$4.3 billion partnership** with Malaysian conglomerate YTL to build supercomputing facilities and cloud AI services.

Malaysia is likely to get caught in the crosshairs of the rule as it become a key destination for Chinese colocation and leasing activities. By 2027, ByteDance, the parent company of TikTok, is expected to

lease **628 MW** of its total data center capacity. Similarly, DayOne (formerly GDS) is adding **415 MW** by 2026, much of which is leased to ByteDance.

Oracle finds itself in a bind—hence the escalating blog posts from Executive Vice President Ken Glueck on **December 19**, **January 5**, and **February 4**. Oracle's strategy relied heavily on Malaysia, with **$6.5 billion in planned investments** that will likely exceed the new **7 percent country cap** for T2 nations. While it could theoretically rebalance through aggressive expansion in T1 countries, this would require significant new investment and strategic repositioning.

After Malaysia, India will be the hardest-hit county. However, unlike Malaysia, India's AI infrastructure is more closely tied to U.S. hyperscalers, offering some insulation from the restrictions. Still, with ~3 GW of planned data center capacity, India closely rivals Malaysia in scale, but has even larger ambitions. **Mukesh Ambani**, India's richest person and chairman of Reliance Industries, recently announced plans to build a **3 GW mega data center** campus in Jamnagar, Gujarat, which would be the world's largest, with a projected investment of $20–30 billion. The facility is intended for AI workloads and is **expected to rely on** NVIDIA's leading-edge Blackwell AI processors. The new restrictions threaten to derail these initiatives, potentially thwarting India's aspirations.

Beyond Malaysia, Southeast Asia overall is likely to suffer.

Singapore was an **early mover** in the data center boom of the 2000s, leveraging its connectivity infrastructure. But its energy constraints quickly caught up. In 2019, Singapore imposed a **moratorium** on new data centers after projections showed they **could consume** 12 percent of the nation's electricity by 2030. Since then, Singapore has allocated modest expansions—**80 MW for four new data centers** by 2023 and a **2024 pledge to add 300 MW**, prioritizing green energy options. Given its fundamental constraints, however, Singapore's growth will remain capped, relying on reshuffling its deck to retire legacy facilities and optimize existing capacity.

Indonesia, meanwhile, is less handicapped by similar limitations and has been muscling in on the AI data center space. It recently completed **Phase 1 of the BDx CGK4 campus** in Jatiluhur—a renewable-powered AI data center park, **scalable up to 500 MW**, offering high power density of up to **120 kW per rack**, liquid cooling technologies, and high-speed connectivity to meet the demands of AI workloads.

Indonesia's ample land, energy, and ability to leverage renewable power have also made it a natural magnet for Chinese investment, with Tencent Holdings pledging **$500 million** to develop its third data center in the country by 2030.

Brazil will be the most affected country in Latin America and has positioned itself as a regional AI data center powerhouse. Leading the charge is Scala Data Centers, whose São Paulo campus is set to expand beyond 350 MW, but its real centerpiece is the so-called **AI City**—a proposed 4.75 GW campus that will cost upward of $90 billion. Ordinarily, such a proposal would be laughed out of the room, but the proposed site is **located** next to an idle 3 GW substation and surrounded by untapped wind and hydro power, resources that few other nations can match. With the right financial backing, it could be the largest AI data center in the world. The new U.S. restrictions would, however, significantly undercut these plans.

While the UAE and Saudi Arabia show relatively small confirmed capacities through 2027, this understates their long-term aspirations, with several gigawatt-scale projects that extend well beyond 2027. Most, however, have not yet started earthworks. The Gulf states possess two distinct advantages: huge energy reserves and the ability to deploy state-backed capital with near-limitless patience to build the next generation of AI data centers.

Near-term development is led by G42/Khazna, with **406 MW of planned capacity** across 13 campuses, including a flagship 100 MW Dubai Ajman campus. However, its true ambitions are reflected in its longer-term plans. According to **SemiAnalysis**, G42/Khazna is planning a staggering 5 GW aggregate pipeline across the Middle East, while Google is eyeing a 3 GW pipeline near Saudi Arabia's King Salman Energy Park.

Biggest Winners

While the AI Diffusion Rule does not create outright "winners," apart from the United States, which by design will be the primary beneficiary, T1 countries' AI infrastructure plans will not be at risk and can proceed without regulatory headwinds. In addition, they are likely to siphon off deployments originally earmarked for T2 nations now caught in the rule's constraints. Based on **SemiAnalysis data**, Australia emerges as a standout case among T1 nations, followed by the United Kingdom, Japan, Ireland, Germany, Canada, South Korea, the Netherlands, and Spain. In fact, nine out of eighteen T1 countries have more than 1 GW of planned AI data center capacity, compared to just four countries across all T2 countries, with five T1 countries exceeding 2 GW of planned capacity.

Australia will likely be the largest beneficiary of the rule, particularly in the Asia-Pacific (APAC) region, with its ~3 GW of planned capacity. Australia's energy landscape provides a crucial competitive advantage. Unlike Japan and South Korea, which rely heavily on imported energy, Australia is **one of the world's largest energy exporters**. Australia's consistent **year-round solar irradiation** also translates directly to data center economics, enabling cost-effective power purchase agreements (PPAs) for data center operators.

Moreover, colocation accounts for approximately **75 percent** of Australia's market. While hyperscalers like Microsoft, Google, and AWS can self-build and operate their own facilities, colocation providers offer a compelling alternative: ready-to-use data center facilities that cloud providers can quickly lease, which means lower capital requirements and faster time to market.

The Australian market is anchored by **three mature colocation providers** that rank among the world's best—the "three Goliaths": AirTrunk, NextDC, and Canberra Data Centers (CDC). **NextDC's 550 MW mega campus** demonstrates its capacity for large-scale development, while the ability of AirTrunk, a homegrown success story, to **deploy** direct-to-chip liquid cooling is particularly significant as the industry faces a big transition: NVIDIA's upcoming Blackwell AI chips (GB200) **require** this advanced cooling technology to handle power densities up to 130 kW per rack. Many data center operators globally, including tech giant Meta, have had to completely **redesign facilities** to accommodate these new requirements.

What truly distinguishes the "three Goliaths" is their focus—instead of serving a broad mix of enterprise clients, they primarily build for hyperscale cloud providers—and already well understand their exacting

technical requirements and scaling needs. Combined with its privileged T1 status under the new rules, Australia is ideally positioned to capture displaced AI computing demand from restricted Southeast Asian markets like Malaysia.

The Theory of Success for the United States

The AI Diffusion Rule will funnel T2 countries toward U.S. hyperscalers and allied T1 neoclouds as the default gateway to advanced compute, which the U.S. largely controls, creating a de facto lock-in of AI infrastructure worldwide.

The rule is not just about controlling who gets access to U.S. compute; it is also about forcing countries to choose sides. In order to secure U.S. compute, T2 NVEU applicants will also have to sever supply chain dependencies with China—i.e., they need to declare for Washington and eject Beijing, or risk falling behind. But company-level decoupling will not suffice; securing U.S. approval will require formal government assurances at the national level to get into Washington's good graces.

CREATING ECONOMIC PRESSURE: THE APPROVAL GAP BETWEEN UVEU AND NVEU

While the path to securing NVEU status appears straightforward, as highlighted previously, the process presents significant practical hurdles that could create dangerous delays for countries with ambitious AI data center plans.

Many projects, like those in Malaysia, are purpose-built for top-end AI training, with ultra-dense power racks, advanced liquid cooling, and infrastructure optimized for next-generation chips like NVIDIA's Blackwell line. Yet, even with an NVEU license, TPP limitations and supply constraints may prevent operators from acquiring enough state-of-the-art GPUs to fill those racks. Meanwhile, Tier 1 license holders face a 7 percent limit on how much of their total compute can be deployed in any single country, creating further uncertainty for T2 data centers looking to secure large clients.

Operators could repurpose racks for older GPUs, but these do not require the high-density cooling and power infrastructure already in place, turning specialized, high-capex facilities into underutilized white elephants. A retrofit for lower-performance workloads also undermines the original investment altogether.

Moreover, the economics are unforgiving. Modern data center projects require billions in upfront capital expenditure—from land acquisition and power infrastructure to advanced cooling systems. These investments are typically **highly leveraged**, with financing structures that assume rapid customer deployment to generate cash flow for debt service.

Even short delays awaiting NVEU approval leave these specialized facilities vulnerable. Loan payments may come due without corresponding revenue from AI workloads, but incurring the same fixed costs for maintenance, security, and staff. For operators and investors, the choice will be clear. Unfilled racks mean certain losses. They will likely lease to T1 hyperscalers who can deploy immediately under UVEU status, or risk their facilities becoming stranded assets.

LOCK-IN EFFECTS

The framework effectively drives T2 compute capacity toward U.S. hyperscalers and T1 neoclouds, which gain preferential access to T2 markets. But the implications go beyond just access to compute.

Hyperscalers and neoclouds also offer stable service-level agreements and robust developer ecosystems (e.g., CUDA libraries), with better track records for uptime, enterprise-grade support, and compliance. Over time, the cost and friction of switching to alternative providers like Huawei would become prohibitively high. Migrating datasets, restructuring technical operations, retraining staff, and rebuilding applications would impose significant costs and operational challenges, reinforcing lock-in effects.

Potential Response from China and Allies

THE HARDWARE GAP: CHINA'S COMPUTE GAP

The assumption that China can immediately compensate for U.S. controls with domestic alternatives is not supportable with current evidence. In the near term, China faces significant constraints in both the quality and quantity of chips it can produce.

Quality Gap: Huawei's Ascend series and other Chinese GPU alternatives lag behind NVIDIA by one to two generations. According to **Chris Miller**, Huawei's most advanced AI chip, the Ascend 910B, achieves only 280–400 TeraFLOPS compared to 2,250 TeraFLOPS for NVIDIA's most advanced Blackwell chips. This performance differential of 5.6–8x is reflected in real-world adoption. According to **Epoch AI**, of **263 documented AI models** where hardware was known, only two used Huawei Ascend chips, while 31 Chinese organizations relied on NVIDIA hardware. Even **DeepSeek** trained its models on NVIDIA H800s.

Quantity Gap: **SemiAnalysis projects** that China will produce just 1.8 million Huawei Ascend 910B/910C GPUs by the end of 2025, while U.S. AI labs and hyperscalers are projected to deploy **14.3 million** AI accelerators in the United States, which are significantly more performant, suggesting an even larger compute gap at the aggregate national level. This limited fabrication capacity means that China will likely prioritize domestic needs, limiting their ability to offer a credible alternative to U.S. technology in global markets.

It is worth cautioning that this technology gap **may not be permanent**. In the medium to long term, forced localization and state-backed capital might narrow China's performance gap from one generation behind to on par or slightly behind.

SECOND-TIER STATUS, FIRST-TIER AMBITIONS

T2 countries will, of course, publicly acquiesce to U.S. restrictions. But diplomatic accommodation should not be mistaken for genuine alignment. India, the Gulf states, and other well-capitalized AI aspirants will bristle in private that their AI ambitions depend on Washington's goodwill. The fact that T2 nations are being deliberately kept a generation behind the frontier will also rankle. It is a public, institutionalized reminder that no matter how much they invest, they are not allowed to be first-tier players.

THE FALSE BINARY: THE UNITED STATES VS. CHINA VS. A THIRD WAY

There is a tendency to view AI geopolitics through a Cold War–style binary—that nations must either align with the United States or drift into China's orbit. But this overlooks a third possibility that T2 nations, far from being passive satellites, may seek their own path. T2 nations have no desire to be entirely captive to either the United States or China.

2019 is often cited as China's "9/11 moment," when the **Huawei and ZTE sanctions** forced Beijing to embark on a massive technological self-sufficiency push. The AI Diffusion Rule may trigger a similar reckoning for T2 countries. One may argue that the two are not analogous. After all, Beijing was explicitly cut off, while T2 countries still receive a permissive allocation under the new framework.

But that misses the underlying dynamic. No country makes economic security decisions on the basis of GPUs alone. Increasingly muscular U.S. economic security measures, especially against close allies like **Canada,** will force capitals to rethink long-term dependencies on U.S. technology.

SHORT-TERM ADJUSTMENTS, LONG-TERM REALIGNMENT

Of course, this will not happen overnight. In the near term, T2 nations will maximize their initial allocation of 50,000 H100-equivalent GPUs. Given that access is first come, first served, sovereign AI initiatives will ramp up—governments will ensure national priorities dictate GPU access, rather than individual firms. We may also see the rise of regional compute corridors, with nations pooling resources to overcome individual capacity limits.

In the immediate time frame, these countries will likely procure whatever GPUs they can—via U.S.-validated entities (UVEUs) or the narrower NVEU status. As access to U.S. compute becomes increasingly conditional, however, or simply in anticipation of future friction, well-resourced nations will start to invest in their own high-performance computing infrastructure, aiming not to match the raw performance of top-end GPUs—a feat even China struggles with—but to build a functional, mid-tier alternative capable of supporting AI applications. Rather than attempting to replicate leading-edge GPUs, these collaborations could focus on specialized application-specific integrated circuits (ASICs) designed for narrow but critical industrial and commercial AI workloads.

RISC-V-based AI accelerators, designed collaboratively and manufactured in existing T2 facilities, present another pathway. Unlike proprietary architectures like **x86 or Arm**—both of which are subject to U.S. restrictions—RISC-V's open-source nature allows nations to design their own AI accelerators. That said, RISC-V is not an immediate off-ramp. While **Alibaba** has developed XuanTie RISC-V cores, most RISC-V AI efforts remain in their early stages, requiring significant investment and development. Claiming it as a ready solution for compute sovereignty overstates its current capabilities. But dismissing it entirely would ignore incentives to fast-track progress.

If Washington was already concerned about China's push for **design-out**, the risk now multiplies exponentially. By overplaying its hand, the United States creates a potential alternate compute stack, operating beyond American control.

THE OPEN-SOURCE ESCAPE VALVE

As highlighted earlier, the weights of any model trained using U.S.-controlled compute are subject to U.S. export restrictions if they exceed 10^{26} FLOP—unless they are open-sourced. At first glance, this creates an incentive for T2 nations to lean into open-source frontier AI development to sidestep U.S. regulatory controls.

However, in practice, this escape valve may not be as open as it seems. T2 countries will likely lack the domestic compute capacity to train models beyond 10^{26} FLOP, meaning they would need to rent compute. But this raises a liability risk—how can a T2 entity credibly prove, before training begins,

that it will follow through on open-sourcing the model weights? Without a mechanism to verify intent pre-training, compute providers may simply deny access upfront rather than gamble on post-training compliance. As such, T2 players may find themselves effectively boxed out of both closed- and open-source frontier AI development.

But even if U.S. restrictions block T2 nations from training frontier open-source models, they may trigger a different kind of shift—an acceleration toward compute-efficient, open-source AI development. DeepSeek has already demonstrated that scarcity breeds optimization. What happens when the broader open-source world, united by shared limitations, begins running in the same direction? There is also a curious inversion at play. While the United States moves to lock down and consolidate global compute, China is positioning itself as the provider of last resort, offering high-quality, cost-effective open-source models that others can build on. Instead of needing access to U.S. AI infrastructure, developers could simply build atop China's open-source stack.

T2 states do not need a formal conspiracy with Beijing. But their independent hedging efforts, driven by frustration with U.S. licensing constraints, may naturally dovetail with China's open-source push. Instead of reinforcing a U.S.-led order, Washington's grip could weaken and lead to a world where open-source autonomy, championed by China, becomes the default escape hatch.

THE FRAGILE EQUILIBRIUM AND CONDITIONS FOR UNRAVELING

By tying advanced compute access to strict controls, Washington risks turning T2 countries into reluctant vassals who start looking for side doors at the earliest opportunity.

Regulatory partitions have a half-life. The policy will only hold as long as:

1. The compliance burden remains lower than the cost of switching;
2. Alternate compute stacks remain too inefficient to offer mid-tier alternative U.S. incumbents; and
3. China's domestic AI stack develops more slowly than the United States anticipates.

The AI Diffusion Rule assumes that lock-in will hold because switching costs are high. But if history tells us anything, it is that whenever an access-restricted market grows large enough, the incentive to develop alternatives eventually outweighs the costs of remaining dependent.

Conclusion and Recommendations

It is more likely than not that the AI Diffusion Rule will be embraced and further expanded by the Trump administration. It aligns with a broader escalation of the China containment playbook, fits neatly into an "America First" approach, and serves as a powerful negotiating cudgel to compel T2 nations to align more closely with Washington's AI agenda.

But splintering is not inevitable. There are viable policy pathways to preempt T2 strategic drift and allow the United States to maintain control over the global AI stack.

1. **Establish Clear Graduation Requirements for T2 Countries:** If full T1 status is not feasible, Washington should establish a Tier 2A classification for countries with significant AI infrastructure investments at risk of being stranded, and agree to increase export control enforcement efforts. This tier would receive higher country-level TPP allocations and streamlined NVEU approval processes. There could also be an annual revision of TPP

thresholds, particularly as large-scale U.S. initiatives such as Project Stargate take off, to allay T2 country anxieties that that they are being deliberately outpaced and falling further behind the compute frontier.

2. **Reassert U.S. Leadership in Open-Source AI to Counter China's Inroads:** The AI Diffusion Rule unintentionally incentivizes T2 nations to embrace open-source AI. Rather than cede this space, Washington should preemptively shape the open-source landscape to prevent China from becoming its de facto steward. This could take place by incentivizing international AI research collaborations through U.S. university public compute. This would offer structured, monitored access to AI infrastructure for vetted T2 researchers—allowing the United States to maintain influence over who contributes to the global open-source AI stack. In time, the United States could also develop certification standards for "trusted" open-source models that meet security requirements.
3. **Modernizing BIS and Export Controls Enforcement:** The streamlined licensing burden on BIS creates an opportunity to reallocate resources toward strengthened enforcement and tracking mechanisms that better prevent circumvention. This would dovetail with the growing digitization drive under the Trump administration. BIS could deploy automated TPP tracking systems that provide real-time visibility into GPU deployments, create standardized APIs for reporting and monitoring, and build automated early warning systems for potential diversion.

In some ways, the United States has crossed the Rubicon, and there is no retreat. China will continue its drive for AI self-sufficiency regardless of U.S. actions. But export controls are like self-replicating automata. They tend to expand with each iteration, creating new loopholes, countermeasures, and pressures for escalation. A more adaptive approach, however—one that balances U.S. leadership with credible pathways for allies—could allow Washington to have its cake and eat it too. If there is a better hand to play, now is the time to find it. And if there is anyone that thrives on breaking and remaking the playbook, it is President Trump. ■

Barath Harithas *is a senior fellow in the Economics Program and Scholl Chair in International Business at the Center for Strategic and International Studies in Washington, D.C.* *The author is grateful to Catharine Mouradian, program manager and research associate in the Economics Program and Scholl Chair in International Business, for her valuable assistance on this paper.*

This report is made possible by general support to CSIS. No direct sponsorship contributed to this report.