Location Verification is not Enough: A Dual Framework to Combat Large-Scale AI Chip Smuggling

Seth Lifland [*1] **Shubhrangshu Debsarkar** [*1]

Abstract

Chip smuggling, the illegal practice of moving chips across borders to restricted countries, is a major weakness for western export controls. This paper presents a hardware verification framework to curtail this increasingly important issue. In contrast to the location verification required in the proposed Chip Security Act, we accomplish this with a combination of location verification and offline licensing, two methodologies that are insufficient on their own but together make smuggling economically infeasible. Neither method would present a prohibitive implementation challenge for leading AI chip designers and manufacturers.

1. Introduction

1.1. Chip Smuggling

In an attempt to maintain and extend its lead in the AI race, the U.S. has implemented heavy export controls restricting advanced AI chips and chip-making machines from being sold to China (Sutter, 2025). By forcing Chinese firms to use less powerful chips, the U.S. prevents Chinese developers from easily training state-of-the-art AI models.

In addition to investing in an indigenous semiconductor supply chain and focusing on algorithmic efficiency, China has turned to large-scale chip smuggling; multiple recent cases showcase the extent, including a co-founder of Super Micro Computer being charged in a $2.5 billion USD case (Scannell, 2026). Two other $100+ million USD chip smuggling cases have surfaced since December 2025 (U.S. Department of Justice, Office of Public Affairs, 2026; 2025). Smuggling fundamentally undermines existing export controls, allowing China to gain access to advanced chips. Smuggled chips are also implicitly diverted from U.S. companies, as NVIDIA faces a growing backlog and American companies are rapidly scaling their AI compute clusters (Vena, 2026). Smuggled chips also present barriers to the verifiability of any future international agreement. For these reasons, we believe that chip smuggling is a key issue with current export controls that needs to be patched for future export controls to be effective.

1.2. Technology

1.2.1. Location Verification

One particularly promising smuggling detection tool is hardware-enabled location verification. AI chips are queried via latency-bounded cryptographic challenges from a network of landmark servers, enabling coarse geographic inference (O'Gara et al., 2025; Brass & Aarne, 2024). It is important to note that the network paths are not geodesic, and are instead governed by inter-network policies (Subramanian et al., 2002). This makes delay-based verification methods probabilistic, imprecise, and dependent on landmark server distribution.

The Chip Security Act, which passed through the U.S. House Foreign Affairs Committee on March 26, 2026, explicitly requires location verification for AI accelerators (U.S. Congress, House of Representatives, 2025). We believe that this legislation is a step in the right direction, but there are still several key considerations. In addition to aforementioned imprecision, the primary drawback of this methodology is the lack of enforceability. Upon detecting that a chip is in a non-approved country, there is no mechanism for preventing it from continuing operation.

1.2.2. Offline Licensing

Offline licensing is a mechanism through which AI accelerators can only access full capabilities when in possession of a cryptographically signed license. These systems have hardware-based roots through their cryptographic IDs and secure boot programs (Phala Network, 2025). The license encodes a validity window and usage constraints, such as permitted operations or a compute budget. During execution, the chip enforces these limits internally so features such as high-throughput tensor core usage are gated offline.

Major adversarial security concerns worth addressing are

[1]Department of Engineering, University of Virginia, Charlottesville, Virginia, United States of America. Correspondence to: Seth Lifland <qyb9kg@virginia.edu>, Shubhrangshu Debsarkar <fvc9ch@virginia.edu>.

Preprint. April 30, 2026.

firmware rollback and key extraction (Wang et al., 2021). Firmware rollback involves installing an older, signed firmware version that predates licensing checks or has weaker enforcement; without anti-rollback protections, attackers can simply boot into a legacy state. Key extraction targets the device's hardware-based root secrets and uses them to forge licenses. Ensuring accurate metering of workloads can also be a challenge, and it is important to choose the right metric to measure when licensing.

2. Our Solution

We propose extending the Chip Security Act to require both location verification and offline licensing to curb chip smuggling. The changes we propose would be largely firmware-based, minimizing the need for hardware redesign. Whereas prior work has catalogued these mechanisms as separate options within the broader space of hardware-enabled AI governance (O'Gara et al., 2025), our contribution is a specific composition, in which re-licensing functions as the enforcement gate for location verification, framed as a concrete amendment to pending U.S. legislation.

2.1. Implementation & Policy

The offline licensing we propose would require advanced AI chips to keep a tamper-resistant monotonic counter on each chip, which periodically checkpoints into persistent storage with anti-rollback protection. After a certain number of cycles, the chip's license would expire and the chip's throttling mechanism would limit chip functionality until the chip's license is restored. This would repeat indefinitely for as long as the chip is active.

In order to re-license, a chip would need to temporarily go online and receive a license from remote servers. As part of re-licensing, the server would evaluate location verification signals to determine whether the chip appears to be operating in a restricted jurisdiction. Chips that fail location verification, present invalid attestation evidence, run unauthorized firmware, or report suspicious counter states would not receive renewed licenses.

To resist replay attacks and fake licensing servers, each re-licensing session uses a fresh server-generated nonce and mutual authentication. The chip signs an attestation report containing the nonce, chip identity, monotonic counter state, and license epoch using a hardware-rooted device key. The server verifies this evidence before issuing a signed license token. The token is bound to the specific chip, firmware version, and expiration window, and is signed under a manufacturer-controlled licensing key whose root is embedded in chip firmware. The chip accepts only tokens whose signature chain validates and whose counter is strictly greater than any previous license.

2.2. Rationale

We model chip smuggling as an economic decision. Let the expected utility of a smuggling attempt be:

$$U = P_{\text{success}} \cdot V_{\text{compute}} - C_{\text{smuggle}} - C_{\text{attack}} - P_{\text{caught}} \cdot C_{\text{enforcement}} \tag{1}$$

where P_{success} is the probability that smuggled chips remain usable, V_{compute} is the value derived from their use, C_{smuggle} is the cost of transporting chips across borders, C_{attack} is the cost of bypassing technical protections, P_{caught} is the probability of getting caught by U.S. authorities, and $C_{enforcement}$ is the cost of getting caught. Our objective is to drive $U < 0$ for most actors.

Our policy reduces U along multiple dimensions. Location verification decreases P_{success} by increasing the likelihood of detection or denial during re-licensing. It similarly increases P_{caught} by better detecting smuggling which can be traced back to perpetrators. Offline licensing reduces V_{compute} by enforcing periodic authorization, causing degraded performance or shutdown if licenses are not renewed. At the same time, bypassing both mechanisms requires sophisticated hardware and firmware attacks, increasing C_{attack}.

2.3. Weaknesses

Sophisticated hardware attacks are the main weakness with our proposal. Examples include extracting the device key from the chip or glitching firmware. With respect to location verification, adversaries can relay challenges through low-latency proxies in approved regions, undermining location guarantees (Brass & Aarne, 2024). All of these methods require resources, expertise, and investment, and are not guaranteed to succeed. The higher the uncertainty of an attack method, and the higher the investment and expertise required, the less appealing chip smuggling becomes.

3. Conclusion

Chip smuggling is emerging as a central failure mode of existing export controls. We propose a framework that combines hardware-backed location verification with offline licensing to detect bad actors and condition continued access to high-performance compute on successful re-licensing. This enforcement mechanism increases the cost and uncertainty of chip smuggling, limiting the scale of illicit compute. Future work should focus on strengthening resistance to hardware-level attacks and establishing international standards for licensing and verification. Incorporating these mechanisms into export control regimes would represent a meaningful step toward making AI governance technically enforceable.

Impact Statement

This paper proposes a policy for mitigating chip smuggling, which would help enforce existing and future export controls relating to advanced AI chips. Cracking down on chip smuggling could lead to enhanced race dynamics or increase the likelihood of desperate actions from the CCP. That said, we think that the inherently illegal nature of smuggling makes cracking down on it unlikely to provoke conflict.

References

Brass, A. and Aarne, O. Location verification for ai chips, April 2024. URL https://www.iaps.ai/research/location-verification-for-ai-chips.

Dhanuskodi, G., Guha, S., Krishnan, V., Manjunatha, A., Nertney, R., O'Connor, M., and Rogers, P. Creating the first confidential gpus. *Communications of the ACM*, 67 (1):60–67, 2024. doi: 10.1145/3626827. URL https://cacm.acm.org/practice/creating-the-first-confidential-gpus/.

Mann, T. Three charged in singapore with alleged link to illicit shipments of nvidia gpus to china. *The Register*, February 2025. URL https://www.theregister.com/2025/02/28/singapore_arrest_server_fraud/.

O'Gara, A., Kulp, G., Hodgkins, W., Petrie, J., Immler, V., Aysu, A., Basu, K., Bhasin, S., Picek, S., and Srivastava, A. Hardware-enabled mechanisms for verifying responsible AI development. In *ICML Workshop on Technical AI Governance (TAIG)*, 2025. URL https://openreview.net/forum?id=JgxkV2jzzk.

Phala Network. Gpu tee deep dive: Securing ai at the hardware layer. *Phala Blog*, June 2025. URL https://phala.com/posts/Phala-GPU-TEE-Deep-Dive.

Reuters. Nvidia working on chip tracking or location controls to comply with u.s. export rules. *Reuters*, 2024. URL https://www.reuters.com/technology/nvidia-working-chip-tracking-location-controls-comply-with-us-export-rules-2024-05-03/. Accessed: 2026-04-24.

Scannell, K. Co-founder of tech company charged with diverting $2.5 billion in nvidia ai chips to china in violation of export laws. *CNN*, March 2026. URL https://www.cnn.com/2026/03/19/politics/super-micro-computer-founder-charged-ai-chips-china.

Subramanian, L., Agarwal, S., Rexford, J., and Katz, R. H. Geographic properties of internet routing. In *USENIX Annual Technical Conference*, pp. 243–259. USENIX Association, 2002. URL https://www.usenix.org/events/usenix02/full_papers/subramanian/subramanian.pdf.

Sutter, K. M. U.s. export controls and china: Advanced semiconductors, September 2025. URL https://www.congress.gov/crs-product/R48642.

U.S. Congress, House of Representatives. Chip security act. H.R. 3447, 119th Congress, 1st Session, 2025. URL https://www.congress.gov/bill/119th-congress/house-bill/3447/text. Introduced May 15, 2025.

U.S. Department of Justice. Two chinese nationals arrested on complaint alleging they illegally shipped to china sensitive microchips used in ai applications, August 2025a. URL https://www.justice.gov/opa/pr/two-chinese-nationals-arrested-complaint-alleging-they-illegally-shipped-china-sensitive.

U.S. Department of Justice. U.s. citizens and chinese nationals arrested for exporting artificial intelligence technology to china, November 2025b. URL https://www.justice.gov/opa/pr/us-citizens-and-chinese-nationals-arrested-exporting-artificial-intelligence-technology.

U.S. Department of Justice, Office of Public Affairs. U.s. authorities shut down major china-linked ai tech smuggling network, December 2025. URL https://www.justice.gov/opa/pr/us-authorities-shut-down-major-china-linked-ai-tech-smuggling-network.

U.S. Department of Justice, Office of Public Affairs. Chinese national and two u.s. citizens charged with conspiring to smuggle artificial intelligence technology to china, March 2026. URL https://www.justice.gov/opa/pr/chinese-national-and-two-us-citizens-charged-conspiring-smuggle-artificial-intelligence.

Vena, D. Nvidia investors just got incredible news from ceo jensen huang: 2 highlights from gtc. *The Motley Fool*, March 2026. URL https://www.fool.com/investing/2026/03/16/nvidia-investors-just-got-incredible-news-from-ceo/.

Wang, Z., Xie, W., Wang, B., Tao, J., and Wang, E. A survey on recent advanced research of cps security. *Applied Sciences*, 11(9), 2021. ISSN 2076-3417. doi: 10.3390/app11093751. URL https://www.mdpi.com/2076-3417/11/9/3751.

A. Recent Chip Smuggling Cases

Here are brief descriptions of recent high-profile chip smuggling cases.

- In February, 2025, three men were charged with fraud in Singapore for lying about the destination of imported servers. 22 locations were raided by the police (Mann, 2025).
- In August 2025, two Chinese nationals were arrested for exporting tens of millions of dollars' worth of AI chips. This involved at least 21 shipments of chips which were explicitly restricted without approval from the Department of Commerce (U.S. Department of Justice, 2025a).
- In November 2025, four individuals were charged with a conspiracy to illegally export NVIDIA chips to China. They had already exported 400 A100 GPUs by January 2025, and were interrupted before planned exports which would have sent ten Hewlett Packard supercomputers containing A100 GPUs, as well as 50 separate H200 GPUs (U.S. Department of Justice, 2025b).
- In December 2025, over $50 million USD worth of advanced chips were seized from a smuggler, who had previously exported or attempted to export at least $160 million USD worth of H100 and H200 GPUs (U.S. Department of Justice, Office of Public Affairs, 2025).
- In March 2026, three individuals were charged after ordering 750 servers worth $170 million USD to illegally export to China (U.S. Department of Justice, Office of Public Affairs, 2026).
- In March 2026, Super Micro Co-Founder Yih-Shyan Liaw was charged for diverting $2.5 billion USD worth of servers with NVIDIA's AI chips to China, violating U.S. export controls (Scannell, 2026).

B. Location Verification

B.1. Speed-of-Light and RTT Bounding

Let a verifier V send a challenge to a chip C and measure the round-trip time (RTT) Δt. Assuming negligible processing delay at C, the physical distance d between V and C is bounded by:

$$d \leq \frac{c \cdot \Delta t}{2}, \tag{2}$$

where c is the speed of light in the transmission medium (approximately 3×10^8 m/s in vacuum, lower in fiber).

In practice, propagation occurs over network infrastructure with additional latency sources (routing, switching, queuing). Therefore, the measured RTT is decomposed as:

$$\Delta t = \frac{2d}{c} + \delta_{\text{net}} + \delta_{\text{proc}}, \tag{3}$$

where δ_{net} captures network-induced delays and δ_{proc} accounts for processing overhead. Conservative upper bounds on δ_{net} must be estimated empirically or via calibration.

To tighten bounds, systems employ:

- Multiple geographically distributed verifiers (multilateration)
- Statistical filtering over repeated measurements
- Hardware-timestamped packets to reduce jitter

B.2. Multilateration and Landmark Servers

A single RTT bound yields a distance constraint (a sphere). With multiple verifiers $\{V_i\}$, the chip location is constrained to the intersection of these regions. In practice:

- Each verifier measures Δt_i and computes $d_i^{\max}$
- The feasible region is $\bigcap_i \{x : \|x - V_i\| \leq d_i^{\max}\}$
- Inconsistent constraints indicate spoofing or relay attacks
- Relay attacks can still be somewhat dangerous if they are positioned correctly
- It is important to note that all time-delay based estimations are generally imprecise, which is an advantage when looking through the lens of privacy and secure computing.

Landmark servers must satisfy:

- Known, fixed geographic coordinates
- Low-latency, high-bandwidth network connectivity
- Hardware support for precise timestamping (e.g., NIC-level)
- Resistance to compromise (hardened OS, remote attestation)

Strategic placement (e.g., across continents) improves localization precision and increases the difficulty of adversarial relay.

C. NVIDIA Hardware + Software Ecosystem

Modern NVIDIA AI accelerators (e.g., A100 and H100) incorporate a hardware–software stack designed to support secure execution, isolation, and attestation. These capabilities are primarily exposed through NVIDIA's confidential computing framework, which provides a Trusted Execution Environment (TEE) on the GPU.

At a high level, the security model is rooted in a hardware root of trust embedded within the device. During boot, a secure boot chain verifies the integrity and authenticity of firmware components before execution. This ensures that only signed and trusted code can run on the GPU. Once initialized, the system can establish a protected execution environment in which workloads are isolated from the host system and other processes (Dhanuskodi et al., 2024).

Within this TEE, both code and data are protected during execution. Memory associated with secure workloads is encrypted, and access is restricted to authorized components. This is particularly relevant in multi-tenant or cloud settings, where the GPU may be shared across users and where the host operating system itself is not fully trusted. The isolation guarantees provided by the TEE ensure that sensitive computations remain confidential even in such environments.

A central component of this ecosystem is remote attestation. NVIDIA GPUs support attestation mechanisms that allow a remote party to verify the state of the device before interacting with it. In a typical attestation flow, the device produces a signed report describing its firmware, configuration, and execution environment. These reports are rooted in device-specific cryptographic keys provisioned during manufacturing, enabling a verifier to confirm that the response originates from genuine hardware operating in a trusted state. Recent reporting indicates that NVIDIA has also developed software capabilities related to enforcing geographic usage constraints on its chips, though such mechanisms are not broadly enforced by default (Reuters, 2024).

These capabilities are exposed through NVIDIA's software stack and are not universally enforced by default. Confidential computing modes, secure execution, and attestation workflows must be explicitly enabled by the system operator or cloud provider. As a result, the security guarantees provided by the hardware depend on deployment configuration and policy decisions at higher layers of the system.

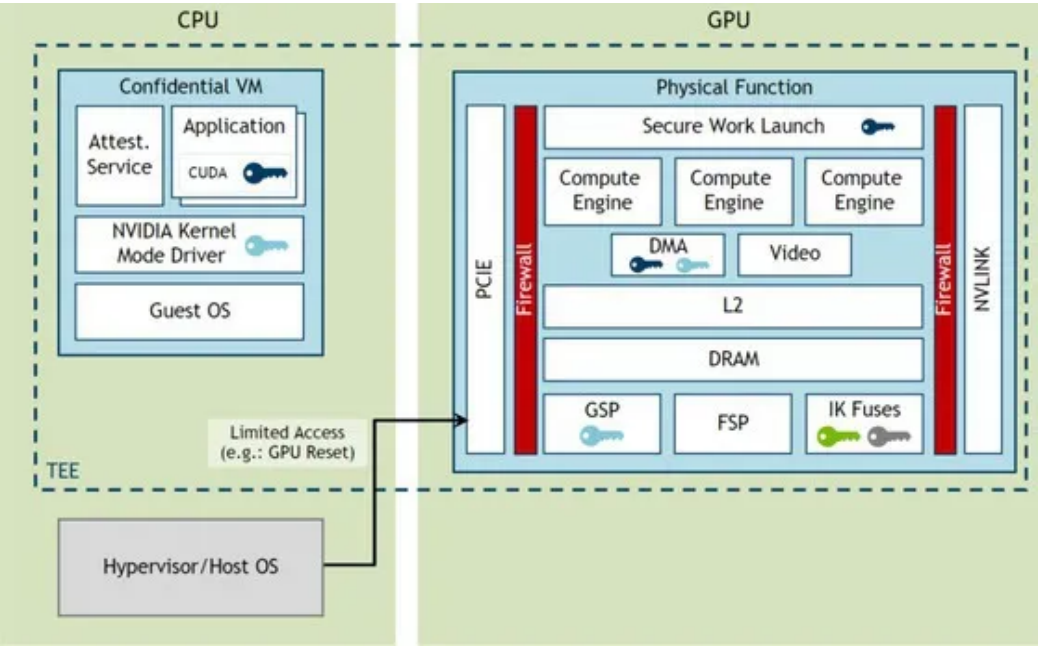

Figure 1. Simplified architecture of an NVIDIA AI accelerator showing secure boot, the embedded security processor, TEE, and attestation interface.

D. Re-Licensing Protocol

The proposed licensing mechanism requires each accelerator to periodically renew its authorization before retaining access to full computational capability. At a high level, re-licensing is a remote attestation protocol combined with a location-verification check. The goal is not only to confirm that the chip is genuine, but also to confirm that it is running approved firmware, has not rolled back its local license state, and appears to be operating in an approved jurisdiction.

During re-licensing, the chip first establishes a secure connection with a manufacturer-approved licensing server. The server responds with a fresh random nonce, which prevents replay of older attestation messages. The chip then produces a signed attestation report using a hardware-rooted device key. This report includes the nonce, a unique chip identifier, the current firmware measurement, the monotonic counter state, and the current license epoch. The inclusion of the nonce binds the report to the current session, while the counter and epoch values prevent reuse of older valid licenses.

The licensing server verifies the attestation report before issuing a new license. This verification checks that the report was signed by genuine hardware, that the firmware version is authorized, and that the reported counter state is consistent with the server's previous records. In parallel, the server evaluates location-verification signals, such as latency-bounded responses from geographically distributed landmark servers. If the chip fails attestation, reports a suspicious counter state, runs unauthorized firmware, or appears to be operating in a restricted jurisdiction, the server refuses to issue a renewed license.

If verification succeeds, the server issues a signed license token. A simplified token structure is:

$$\mathsf{LicenseToken} = \mathsf{Sign}_{\mathsf{MFG}}(\mathsf{chip\_id}, \mathsf{firmware\_hash}, \mathsf{license\_epoch}, \mathsf{counter\_value}, \mathsf{expiration\_window}, \mathsf{allowed\_capabilities}).$$

The token is bound to the specific chip, firmware version, counter state, and validity window. After receiving the token, the chip verifies the manufacturer's signature chain and checks that the new epoch and counter value are strictly newer than the previously accepted license. If these checks pass, the chip updates its local license state and continues operating normally. Otherwise, the chip rejects the token and remains in a degraded or restricted mode until a valid license is obtained.

This protocol provides several security benefits. Fresh nonces prevent replay attacks, manufacturer signatures prevent fake licensing servers from issuing valid tokens, firmware measurements prevent unauthorized software states, and monotonic

counters limit rollback attacks. Location verification alone can only detect suspicious geography; coupling it with periodic re-licensing creates an enforcement mechanism because chips that cannot obtain a valid renewed license lose access to high-performance operation.