

Issue Brief

Tracking AI Chips

Assessing Location
Verification as an Export
Control Enforcement Tool

Authors

Jacob Feldgoise

Kyle Miller

Hanna Dohmen

Executive Summary

It is indisputable that AI chips are being diverted to unauthorized countries, including China. While the exact volume of diversion remains unknown, it nonetheless risks undermining U.S. export controls designed to prevent adversaries from developing and deploying advanced AI capabilities. Location verification—which would allow regulators to determine where AI chips are located—has emerged as a key proposal to strengthen export control enforcement efforts. We assess which types of location verification methods merit consideration, how much they would cost, what their limitations are, and what policy design choices policymakers should weigh. **Ultimately, our analysis suggests that location verification could be deployed for a feasible but non-trivial cost and would enhance U.S. export control enforcement efforts by generating leads and imposing costs on chip smugglers. However, location verification mechanisms have significant shortcomings that limit their effectiveness and ultimate net value, and we believe those shortcomings have received insufficient attention to date. We aim to articulate the limitations clearly to help policymakers make informed judgments about whether and how to pursue such mechanisms.**

A centralized, ping-based location verification (PLV) system—supplemented by physical inspections to investigate leads—would offer the most robust and feasible approach to location verification. We evaluated location verification methods based on four criteria: verifiability, accuracy, security, and repeatability. Based on these criteria, we determine that PLV and physical inspections are the most feasible methods. PLV specifically is more robust than other geolocation mechanisms, inventory management tools, mail-in inspections, and video inspections. PLV, relative to physical inspections, is more cost-effective and scales better in the long term as more chips are deployed.

If policymakers pursue location verification, regulators should initially prescribe a flexible location verification approach, where each responsible entity implements their preferred form of location verification to allow industry to develop, test, and evaluate solutions. After a testing period, the Bureau of Industry and Security (BIS) can mandate the approach(es) it deems most suitable or maintain a flexible approach.

To mitigate the potential for noncompliance, regulators should also centralize responsibility for reporting when a given chip is outside of its intended location; this means a single entity (i.e., BIS itself, the chip designers, or trusted third parties) would be responsible for knowing the intended location of each chip and verifying its location, even if the chip changes hands multiple times. Beyond reporting responsibility, we recommend that a BIS-certified trusted third party operate the infrastructure of the PLV system (i.e., landmark and reporting servers) to avoid conflicts of interest.

Our cost modeling confirms that a PLV system is more cost-effective than physical inspections, and PLV could be deployed for a feasible but non-trivial cost, depending on implementation. The potentially government-borne costs would represent 0.5% to 7.2% of BIS's FY2026 budget if landmark servers are owned or 0.3% to 25.7% if they are rented. Across more than ten million simulated scenarios, physical inspections never achieved higher detections-per-cost than a PLV system, whether the landmark servers are owned or rented. We estimate that a PLV system would cost roughly \$3.1 million to \$28.8 million annually if the landmark servers are owned and \$2.6 million to \$72.4 million annually if they are rented.

Policymakers should have clear expectations of what a PLV system can and cannot achieve, as any PLV system will have significant limitations. Most importantly, a PLV system can only detect diversion after it has already occurred. We identify five additional limitations:

1. **Enabling a PLV system on AI chips in circulation:** For software or firmware implementations of PLV, the system could be enabled on AI chips that have already been manufactured and are in circulation, but doing so would require compliance from each end user who would need to install an update. Hardware implementations of PLV could not be enabled on chips in circulation.
2. **Connectivity between chips and landmark servers:** When a controlled AI chip fails to respond to a PLV system, the system cannot inherently know whether it is due to diversion or legitimate reasons. Chips that remain offline—whether because they are in storage, in transit, or deliberately disconnected by bad actors—produce ambiguity in the system.
3. **Adversarial location obfuscation:** Well-resourced adversaries could trick a PLV system into reporting that a smuggled chip is still in a permitted location. Even a PLV system designed with prudent security measures could be subverted, and as a result some percentage of chips that report being in their intended locations may actually have been diverted.
4. **Approximate geolocation of AI chips:** By positioning one landmark server in every major data center hub—defined as an area containing a large concentration of data centers within a 19-mile radius—outside the United States, China, and Russia, PLV could likely achieve country-level accuracy (an error-circle radius of 116 miles) for the vast majority of PLV-tracked chips. But even in the best case, the accuracy of a PLV system would highly likely be no better than about 20 miles, meaning that the system cannot reliably distinguish

locations between nearby data centers operated by different entities. This may be particularly problematic in areas with a high density of data centers, as well as border regions near restricted locations such as China.

5. **End-user obfuscation:** Because PLV systems verify locations rather than end user identities, the system cannot detect diversion to an unauthorized user where the chip's physical location remains unchanged.

Whether the benefits to U.S. export control enforcement efforts justify the costs will also depend on a series of policy design choices. For each policy design choice, we outline the trade-offs for each question and make a recommendation to aid policymakers in their decision-making. First, we recommend that PLV be implemented with firmware-based signing, as it best balances robustness against adversarial attack, cost, and ease of implementation. Second, we recommend that location verification be required only for data center chips subject to the most stringent licensing requirements to align with current export control objectives. Third, we recommend that BIS-certified trusted third parties operate the landmark and reporting servers in order to eliminate conflicts of interest and leverage private sector expertise. Fourth, we believe the program should aim to detect diversion globally (rather than only to countries of concern) as the additional costs are marginal. Lastly, policymakers also have to consider whether domestically deployed chips are tracked as well, which may require Congress to grant BIS additional authorities.

A U.S. government indictment regarding AI chip smuggling demonstrates the potential benefits and limits of a PLV system. While the PLV system could not have prevented the initial diversion of AI chips to China, it may have alerted U.S. officials to the scheme earlier, potentially helping the government disrupt the operation more quickly. This example highlights that BIS also needs to receive alerts when chips are slow to activate, not just when chips are detected outside of permitted locations.

To fully realize the benefits of a PLV system, BIS must be adequately resourced. Before pursuing any kind of location verification, policymakers should ensure that BIS has sufficient resources and staffing to pursue the leads that such a system would generate. It is well-established that BIS faces resource constraints; therefore, the utility of a PLV system is limited by BIS's capacity to pursue those leads. A recent request to increase BIS's budget to \$450 million for FY2027 would be a positive development, but if BIS enforcement remains seriously resource-constrained, the value of PLV-generated leads will be partially undermined by BIS's capacity challenges.

Table of Contents

Introduction.....	5
Tracking AI Chips	8
Baseline Implementation for Location Verification	11
How a PLV System Works	16
Inherent PLV Limitations	18
Summary	23
PLV Benefits, Costs, and Policy Design Choices	24
Benefits and Costs of a PLV System.....	24
Policy Design Choices	28
Case Study	38
The Smuggling Scheme	38
Failure of Physical Inspections	39
Value of a PLV System	39
Conclusion.....	41
Authors, Acknowledgements, and AI Disclosure Statement	43
Appendix A: Approaches to Chip Location Verification	44
Physical Inspections.....	44
Mail-in Inspections	44
Video Inspections	45
Inventory Management Tools.....	46
Geolocation Mechanisms	47
Ping-Based Location Verification (PLV) System	49
Appendix B: Accuracy of Ping-Based Location Verification	53
Literature Review	53
Trilateration	55
Delay Similarity	74
Appendix C: Centralized vs. Decentralized Location Verification	75
Appendix D: Comparative Cost-Benefit Analysis of Approaches for Location Verification of AI Chips.....	77
Appendix E: Chips Without TEEs	78
TEE-Like Security Features.....	78
Software-Based or Firmware-Based Signing Without a TEE.....	79
Endnotes.....	81

Introduction

Over the last four years, the United States has restricted the export of advanced AI chips to certain destinations, end users, and end uses to hinder adversaries from developing and deploying advanced AI capabilities that are contrary to U.S. foreign policy and national security objectives. The Biden administration implemented the first round of China-wide export controls on AI chips in August 2022 and progressively tightened restrictions in the years that followed. While the Trump administration has not yet issued a comprehensive export control strategy for AI chips, criminal and civil enforcement of existing controls remains a key priority for the administration.¹

Despite increased focus under both the Biden and Trump administrations, enforcement efforts continue to face persistent challenges. While precise figures are disputed and, to a degree, unknown, reporting and enforcement actions suggest that there are substantial volumes of controlled AI chips being diverted to restricted destinations and end users.² Epoch AI estimates that a median of 466,000 controlled AI chips were smuggled into China between 2024 and 2025, representing 663,000 H100-equivalents of compute.³ To put that volume into perspective, in the first quarter of 2026, 663,000 H100-equivalents would have constituted one to two frontier-level clusters' worth of compute, a volume that could substantially reduce compute resource constraints on one or more Chinese AI labs, making it easier for those labs to train and deploy frontier AI models.⁴ If the 663,000 H100-equivalent GPUs were installed into two clusters, each cluster could train GPT-5—which was the most capable AI model when it became available in August 2025—in 8 to 24 days.⁵ A recently uncovered smuggling operation underscores the scale of diversion: in March 2026, a Department of Justice indictment revealed a single scheme from 2024 to 2025 involving AI chips worth \$2.5 billion.⁶ That operation alone could account for up to 145,000 H100-equivalent chips.⁷

The goal of enforcement is not to prevent every single chip from reaching a restricted destination or entity. Some level of diversion is inevitable, and small volumes of diversion likely would not meaningfully augment China's AI capabilities. Where the line is drawn between tolerable levels of diversion and strategically significant levels of diversion depends on judgments about AI capabilities, their national security applications, and the amount of compute necessary to train frontier models and deploy them at scale. That said, if large volumes of cutting-edge GPUs (e.g., over 100,000 AI chips per year) are diverted over an extended period of time, this would likely benefit China's AI development and undermine the objectives of U.S. export controls.

Therefore, policymakers are seeking ways to detect and reduce AI chip smuggling, and a requirement to periodically verify the location of controlled AI chips is one such proposed solution. Such verification could help regulators to confirm that a chip remains at its authorized destination and to flag potential diversion for further investigation and enforcement actions.

Legislators in the U.S. Senate and House of Representatives introduced the Chip Security Act in 2025, which would implement a “chip security mechanism” requirement to verify the locations of controlled AI chips.⁸ On March 26, 2026, the House Foreign Affairs Committee advanced the bill during a full committee markup, and reporting as of this paper’s publication indicates that the bill will likely be included in the manager’s amendment to the Senate National Defense Authorization Act (NDAA) for FY2027, improving its chances to become law.⁹ While legislative hurdles remain, these developments represent the most serious effort yet to advance it through the legislative process. If passed, the Chip Security Act would mandate that the U.S. Department of Commerce require chip designers to outfit covered chips with chip security mechanisms for location verification, which is defined as “a software-, firmware-, or hardware-enabled security mechanism or a physical security mechanism....”¹⁰ When a chip security mechanism fails to confirm that a covered chip is in its intended location, it should alert the exporter of the diversion, and the exporter would subsequently be required to report the diversion to BIS.¹¹

The Act identifies three potential chip security mechanisms: periodic on-site physical inspections, periodic attestation by U.S.-headquartered entities, and ping-based location verification.¹² However, the bill does not prescribe which of the example mechanisms should be used, how they should be technically implemented, or how BIS should design its regulations or processes to implement these mechanisms.¹³

The Chip Security Act has brought the issue into a broader policy debate. For example, the Trump administration recommended further exploration of location verification mechanisms in its AI Action Plan as part of a broader strategy to enhance export control enforcement.¹⁴ Such policies are not without contention, however. In March 2026, the Semiconductor Industry Association (SIA) released a statement formally opposing the Chip Security Act, arguing that “rushing to legislate complex, costly, and unproven security features risks undermining global trust in American semiconductor technologies and efforts to promote the export of the American AI technology stack, eroding our global leadership and competitiveness.”¹⁵ Opposition from SIA, which represents dozens of leading semiconductor manufacturers that would be directly impacted by such policies, highlights the need for pragmatic, unbiased analysis of the benefits and limitations of location verification.

Ultimately, policymakers considering requiring and implementing location verification mechanisms—whether through the Chip Security Act, future legislative proposals, or other regulatory actions—need to understand the trade-offs, costs, limitations, and policy design choices of such an approach, which are influenced by how the chips’ locations are verified technically and how the regulatory implementation is structured.

Some commentators propose going a step further, suggesting operating licenses (effectively kill switches) that would enable AI chips to be turned off remotely where diversion has been detected.¹⁶ These proposals currently appear less politically viable and remain highly controversial in part due to the significant cyber and hardware security vulnerabilities they risk introducing. We therefore focus on location verification mechanisms as the most politically viable solution proposed.

We provide a framework to help policymakers determine whether they should mandate location verification for controlled AI chips, and if so, how to implement it. We first survey the technical approaches available and explain which mechanisms are most reliable and robust against adversarial attacks, based on a set of criteria. We model the relative costs of the two methods that meet these criteria: physical inspections and ping-based location verification (PLV). We then identify the inherent limitations of even the most robust system, estimate what a PLV system would cost to build and operate, and make recommendations for the policy design choices policymakers will have to consider. Finally, we apply this analysis to a real-world smuggling case to illustrate what such a proposed location verification system could and could not have detected in practice to further help policymakers weigh whether the enforcement benefits of location verification justify its costs.

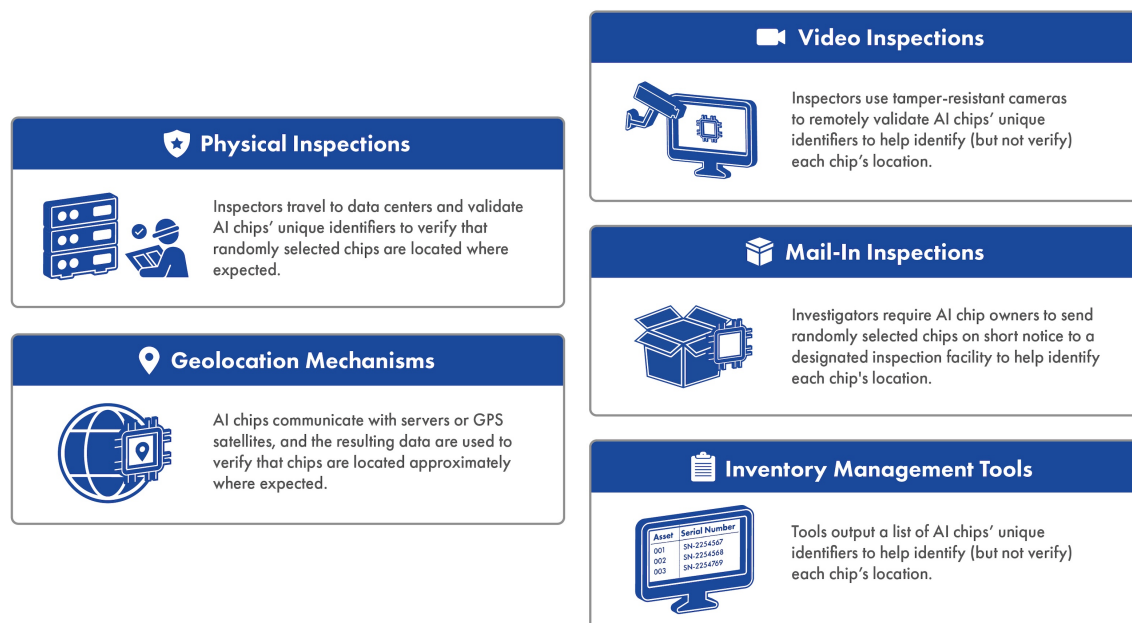
We do not recommend whether or not policymakers should mandate location verification, though we argue that ping-based location verification is the most feasible and cost-effective method when supplemented with physical inspections to investigate leads.

Tracking AI Chips

There are five categories of methods to help identify the location of a chip: physical inspections, mail-in inspections, video inspections, inventory management tools, and geolocation mechanisms.*

Physical inspections involve on-site audits by law enforcement personnel or contracted investigators. **Mail-in inspections** require end users to ship randomly selected chips to law enforcement personnel or contracted investigators for inspection. **Video inspections** involve such personnel or investigators remotely inspecting AI chips using tamper-resistant cameras shipped to the data center. **Inventory management tools**—such as data center infrastructure management (DCIM) tools or other cluster management software—output a list of chips’ unique identifiers (e.g., serial numbers) present at a data center.¹⁷ Lastly, there are a variety of mechanisms that use data generated by chips, servers, and data center infrastructure to remotely and periodically identify and verify their locations (herein referred to as **geolocation mechanisms**).

Figure 1. Five Categories of Methods That Help Identify Chip Locations



Source: CSET illustration.

Physical inspections, mail-in inspections, video inspections, and inventory management tools appear less technically complicated than geolocation methods, but they come with other challenges that limit their feasibility.

* This section overviews these methods, but technical details can be found in Appendix A.

Physically verifying chips on-site is reliable if done properly, but it is also resource-intensive because it requires auditors employed by BIS or a third party to travel to data centers around the world and verify the presence of specific chips.¹⁸ Moreover, physical inspections can be relatively intrusive depending on the number of chips inspected, as proper inspections require deactivating the chips, removing them from server racks, verifying their unique identifiers (e.g., serial numbers), and attesting their authenticity.¹⁹ Less-intrusive physical inspections—for example, only examining the server’s packaging or unpacking the server but not opening the device to view the individual chips—can allow bad actors to trick auditors and are therefore less reliable.²⁰

Like on-site physical inspections, mail-in inspections can be resource-intensive as a sufficient number of inspection locations—likely offices of the U.S. Commercial Service or a U.S. Embassy or Consulate—must be outfitted with the personnel and equipment necessary to inspect chips. Mail-in inspections would also be intrusive for end users as a chip selected for inspection would be out of operation for several days at a time. Furthermore, this method may not verify the information BIS requires: proper inspection at the receiving facility can produce reliable evidence of an AI chip’s identity but does not clearly verify its location, as the method is not accurate and the original location of shipment could be obscured.

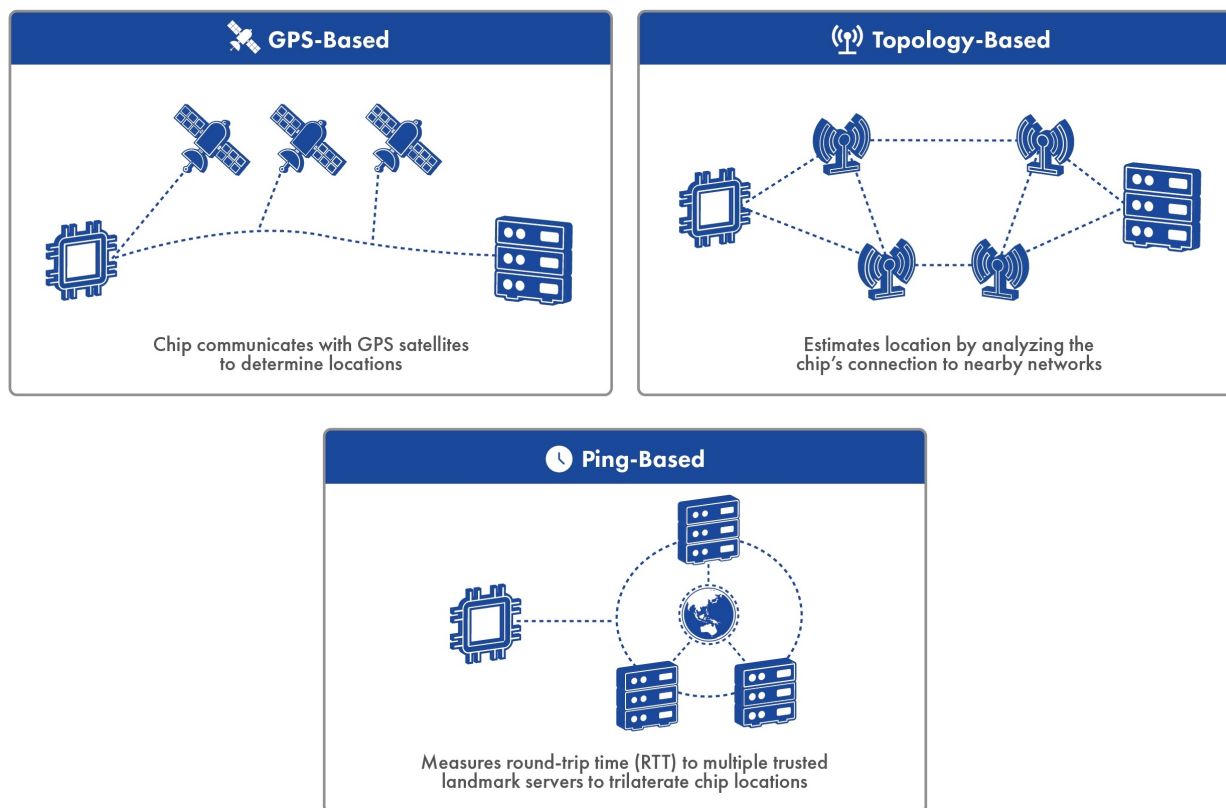
Video inspections may be less resource-intensive than physical inspections as inspectors are not required to travel to a data center, but costs would nonetheless include procuring and shipping tamper-resistant cameras to data centers housing export-controlled chips. Proper video inspections are also intrusive, though somewhat less so than on-site physical inspections and mail-in inspections. Furthermore, this method can attest the authenticity of a chip (if done properly); however, on its own, it cannot directly verify an AI chip’s owner or location.

Inventory management tools, which auditors can access, are commonly used by data center operators to monitor hardware (e.g., chips and server racks). However, they do not directly verify an AI chip’s owner or location, and a sufficiently resourced threat actor can very likely falsify the information they display.* Ultimately, the utility of inventory management tools for auditing chips depends on the security of the tool and how it attests a chip’s identity, and currently such tools are unlikely to be secure enough to enable reliable audits. See Appendix A for more details.

* If a chip’s serial number is listed in an inventory management tool, that on its own does not prove the chip’s location.

Geolocation mechanisms can be based on pre-established infrastructure (e.g., network and telecommunications), installed via software and/or firmware, or embedded in the hardware of the chips themselves, resulting in greater feasibility than physical inspections and inventory management tools allow.

Figure 2. Three Geolocation Mechanism Options



Source: CSET illustration.

There are three general geolocation mechanisms, each of which can be implemented in several ways with varying costs and degrees of robustness and accuracy.²¹ First, GPS-based location verification involves chips communicating with GPS satellites to determine their locations and then sending that information to separate trusted servers. Second, topology-based location verification estimates a chip's location by analyzing its connection to nearby networks.²² Third, ping-based location verification ("PLV") measures the time it takes for a signal to travel from a chip to multiple trusted "landmark" servers and back. The signals' round-trip time (RTT) can be used to "trilaterate" (i.e., estimate) the chip's location.²³ The estimated accuracy of a PLV system depends on the exact PLV geolocation method, the number and location of landmark servers, and the amount of network processing latency when AI chips

communicate with landmark servers.* A PLV system can be supplemented with network topology information to further improve accuracy (see Appendix B).

We evaluate these location verification approaches in the following section.

Baseline Implementation for Location Verification

If BIS chooses, or is required by Congress, to implement location verification for export-controlled AI chips, the system should have two key features. First, BIS should **centralize** reporting responsibility for location verification to a single entity over a chip's lifecycle, which could be BIS, BIS-certified third-party auditors, or chip designers. This entity would receive and monitor location data provided by downstream chip owners as chips change hands, be responsible for verifying the chips' approximate location, and report to BIS if there are discrepancies between intended and approximate locations. Second, BIS should allow **multiple** location verification approaches, so long as they are **verifiable, accurate, secure, and repeatable**. These four criteria are discussed further below.

A **centralized** system would be far less vulnerable to noncompliance than a decentralized system. Under a centralized system, BIS would require that location verification for each chip always be conducted by BIS, a BIS-certified third-party auditor, or the chip designer, even if the chip is resold and changes hands multiple times. This differs from a decentralized regulatory configuration, where location verification would be conducted by the last exporter, which would change with each successive reexport. A decentralized system is more vulnerable to noncompliance; it can be easily undermined if an exporter in the chain is complicit in diversion.[†] For this reason, we advocate for centralized responsibility, regardless of which approaches for location verification are pursued. See Appendix C for a more comprehensive analysis of centralized versus decentralized systems.

We recommend BIS initially pursue a **flexible** location verification system, where each responsible entity can implement their preferred form of location verification so long as it meets four key criteria (discussed below).[‡] While this may result in inconsistent

* See section "How a PLV System Works."

[†] Specifically, the system could be undermined by a last exporter who attests that their customer is complying but does not actually implement location verification.

[‡] As part of this flexible approach, BIS should gather information on the pros and cons of different location verification approaches through requests for information (RFIs). To assess RFI responses, BIS should lean on the technical expertise of the National Institute of Standards and Technology (NIST). This process would help BIS determine which approach(es) are most effective in the long term.

location verification data, we judge that cost is outweighed by the benefit of creating an incentive for industry to innovate new location verification solutions, especially given the fact that location verification for AI chips is nascent and untested. Over time, BIS can determine which approaches are most suitable, and if necessary, require that responsible entities pick from a list of certified location verification options after industry has the opportunity to explore different avenues.

To help guide industry, BIS should prescribe specific criteria that these methods must meet, regardless of which mechanisms industry chooses. Without overprescribing thresholds that would limit industry flexibility, BIS should require that these mechanisms be **verifiable**, **accurate**, **secure**, and **repeatable**.

The method must be **verifiable**, so that the central responsible entity can estimate each chip's physical location without having to fully trust the entity reporting the data. This is a baseline requirement, as reliance on trust makes any method fundamentally vulnerable to chip diversion. For a method to be verifiable, it must (a) produce evidence—preferably involving cryptographic attestation—that establishes whether the AI chip being inspected is the one that investigators intend to inspect and (b) link the chip's verified identity to an estimate of its physical location. If a location verification method is not verifiable, by our definition it can be neither accurate nor secure.

The method must be **accurate**, so that in most cases the locations of chips can be determined at the country level.²⁴ We define country-level accuracy as an error circle small enough to indicate that an AI chip is located either in its expected country or in a neighboring country while ruling out countries farther away.

The method must be **secure**, so that adversaries cannot easily obfuscate the locations of chips. BIS, in consultation with industry, must determine the level of security required for technical approaches, such as geolocation. We recommend BIS not initially prescribe a specific security threshold but instead give industry sufficient time to implement different mechanisms with different degrees of robustness. After that period, BIS should set a minimum security threshold and identify specific mechanisms that meet the threshold. We recommend the threshold should be set such that the technical mechanism cannot be readily compromised at scale (i.e., the vulnerability cannot easily be exploited repeatedly across many chips) by a vulnerability that could be identified and exploited by a moderately sophisticated threat actor.²⁵ See Appendix A for more details on security for technical location verification approaches. For other location verification methods such as physical inspections, security stems from properly vetting auditors and contractors.

Lastly, the method must be **repeatable**, so it can deter potential smugglers and detect chip diversion within shorter timeframes. BIS must consider how often a chip’s location should be verified in order to aid export control enforcement efforts. More frequent inspections increase the likelihood of detecting ongoing diversion operations before the chips leave the reach of U.S. law enforcement. For a location verification method to be repeatable, it would preferably occur no less than twice per year, but optimally, checks would recur weekly or monthly.

Table 1. Comparison of Location Verification Methods

	Ping-Based Geolocation	Physical Inspections	Topology- Based Geolocation	GPS-Based Geolocation	Inventory Management Tools	Mail-In Inspections	Video Inspections
Verifiable	✓ ^B	✓ ^B	✓ ^B	✓ ^B	✗	~	✗
Accurate	✓	✓	✓	✓	✗	✗	✗
Secure	✓ ^B	✓ ^B	✗	✗	✗	✗ ^E	✗
Repeatable	✓	~ ^D	✓	✓	✓	~ ^D	~ ^D
Low Cost	~ ^C	✗ ^C	✓	~ ^C	✓	~	~
Unintrusive ^A	✓	✗	✓	✓	✓	✗	✗

Notes:

(A) Intrusiveness is the degree of interference in the operation of data center servers.

(B) To be verifiable, in addition to estimating an AI chip’s physical location, these methods must also produce evidence—preferably involving cryptographic attestation—that establishes whether the chip being inspected is the one that investigators intend to inspect.

(C) For physical inspections, the cost primarily depends on the number of data centers and chips inspected; for ping-based geolocation, the cost primarily depends on the number of landmark servers; for GPS-based geolocation, the cost primarily depends on the type and amount of GPS hardware.

(D) While these partially repeatable location verification methods meet the “repeatable” criterion, it is infeasible for each method to physically inspect, mail in, or video inspect each chip many times per year.

(E) Mail-in inspections are not secure because the origin of the shipment can be obfuscated. For example, chips could be shipped from export-controlled countries within relatively short time periods using express shipping.

Source: CSET.

Of the location verification options considered in the previous section, two appear to meet each of the key criteria: PLV and physical inspections. PLV systems allow the centralized, responsible entity to verify chip locations, can recurrently locate chips on a daily or weekly basis, are (to varying degrees) more secure than other geolocation mechanisms, and are accurate enough to identify chip locations at the country level. With 500 landmark servers, PLV will likely achieve region-level accuracy (worst case) to city-level accuracy (best case). By positioning one landmark server in every major data center hub—defined as an area containing a large concentration of data centers within a 19-mile radius—outside the United States, China, and Russia, PLV could likely achieve country-level accuracy for the vast majority of PLV-tracked chips. Once all major hubs have been covered, the remaining landmark servers should be distributed among smaller data center hubs to provide baseline region-level accuracy.*

Physical inspections also meet the four key criteria, but the strengths and weaknesses of this method differ from PLV. Proper physical inspections are more accurate, and potentially more secure, than PLV. Detecting diversion with physical inspections has greater value than detecting the same diversion with PLV because physical inspections provide more detailed information and are more likely to directly prove that diversion has occurred. Furthermore, PLV likely has a higher false-positive rate than physical inspections.²⁶ However, while physical inspections can be repeatable (multiple times a year), they cannot be conducted at the same rate as PLV without incurring significant additional costs from auditor salaries and travel expenditures. Poor repeatability and high cost-per-chip are the greatest limitations of physical inspections.

Other location verification methods (i.e., mail-in inspections, video inspections, inventory management, and GPS- and topology-based geolocation mechanisms) are generally inaccurate, unverifiable, or insecure. Mail-in inspections are inaccurate because the mail-in window must be long enough to allow for shipping logistics, making it difficult to ascertain a chip's country of origin with high confidence. On their own, video inspections and inventory management tools are not verifiable; while they may help identify an AI chip's owner, they do not directly verify its owner or location.[†] GPS-based and topology-based mechanisms are generally susceptible to location spoofing, making them insecure. See Appendix A for more details.

* See section "How a PLV System Works."

[†] A video inspection process or inventory management tool must integrate chip attestation to help identify an AI chip's owner. Furthermore, to perform location verification, the tamper-resistant camera used for video inspections and the inventory management tool must also integrate one of the three geolocation mechanisms discussed in this paper: GPS-based, topology-based, or ping-based.

However, in keeping with a flexible location verification system, if industry can improve the accuracy, verifiability, and security of these location verification methods, then they should be allowed to implement them. Furthermore, for highly trusted end users, such as U.S.-headquartered cloud service providers (CSPs), BIS could likely loosen the four criteria, making present-day mail-in inspections, video inspections, inventory reports, and non-PLV geolocation data viable options for these firms.²⁷

Still, to identify the most feasible and cost-effective location verification method, which will be the focus of the rest of the report, we conduct a comparative cost-benefit analysis of PLV and physical inspections.

Cost Modeling: PLV Versus Physical Inspections

To determine which of the two approaches that meet the criteria is more cost-effective—PLV or physical inspections—we modeled their benefits and costs in over ten million simulated scenarios. We hypothesized that PLV would be more cost-effective than physical inspections; therefore, we deliberately chose model parameters favorable to physical inspections to determine whether PLV is more cost-effective even under those conditions. Moreover, we modeled two separate costs for PLV: a PLV system in which landmark servers are owned and operated directly by BIS, chip designers, or a third party, and a PLV system in which landmark servers are rented from CSPs.

We compare the two methods using detections-per-cost: the expected number of diverted chips a method detects divided by the dollar cost (i.e., the higher the detections-per-cost, the more detections of diverted chips per dollar). **We find that PLV is more cost-effective than physical inspections. Physical inspections did not have more detections-per-cost than renting or owning PLV in any of the 10.5 million scenarios we simulated.** In the vast majority of modeled scenarios, both PLV options had strictly higher detections-per-cost; in the remaining scenarios, physical inspections tied one or both PLV options but never outperformed either. We conducted sensitivity checks on all key constants and concluded that the model outcome is not sensitive to any of our choices. For additional details on the cost comparison model, see Appendix D.

Because PLV not only meets the four baseline criteria for an acceptable location verification method but also is more cost-effective than physical inspections (the other acceptable method), the remainder of this report focuses on implementation of a PLV system. We consider the limitations and policy design choices to help policymakers determine whether and how to implement such a mechanism.

How a PLV System Works

Standing up a PLV system would involve configuring controlled AI chips to communicate with landmark servers and follow a location verification protocol when they do so.* Landmark servers then report timing data to a central reporting server, which would aggregate the timing data to estimate each controlled chip's location and compare it against the chip's expected location. As chips are sold to downstream customers, the last exporter would share the intended location of each chip with the entity operating the reporting servers.

With 500 landmark servers, a PLV system will likely achieve region-level accuracy (worst case) to city-level accuracy (best case).²⁸ By positioning one landmark server in every major data center hub—defined as an area containing a large concentration of data centers within a 19-mile radius—outside the United States, China, and Russia, PLV could likely achieve country-level accuracy for the vast majority of PLV-tracked chips. After all major hubs have been covered, the remaining landmark servers should be distributed among smaller data center hubs to provide baseline region-level accuracy. See subsection “Positioning Landmark Servers” in Appendix B.

The accuracy of a PLV system largely depends on the exact PLV geolocation method, the number and location of landmark servers, and the amount of network processing latency when AI chips communicate with landmark servers. The simplest PLV method is trilateration (i.e., geolocating an object by measuring time-of-flight to known landmark servers), which our model for the baseline PLV system relies on. Trilateration accuracy improves as landmark servers are placed closer to AI chips. Therefore, the landmark server operator should place landmark servers first in metropolitan regions with high concentrations of AI chips, and the remaining landmark servers should be distributed among smaller data center hubs to provide baseline region-level accuracy. But even in the best case, PLV will likely be no more accurate than an error circle with a 19-mile radius. See Appendix B for more details.

A global PLV system would likely need between 39 and 500 landmark servers. We find that 500 landmark servers, distributed globally, can very likely guarantee PLV accuracy of 1,000 miles or better, which we consider to be region-level accuracy. A low-effort PLV implementation would likely require at least 39 landmark servers; if distributed globally, these servers can very likely guarantee PLV accuracy of 2,120

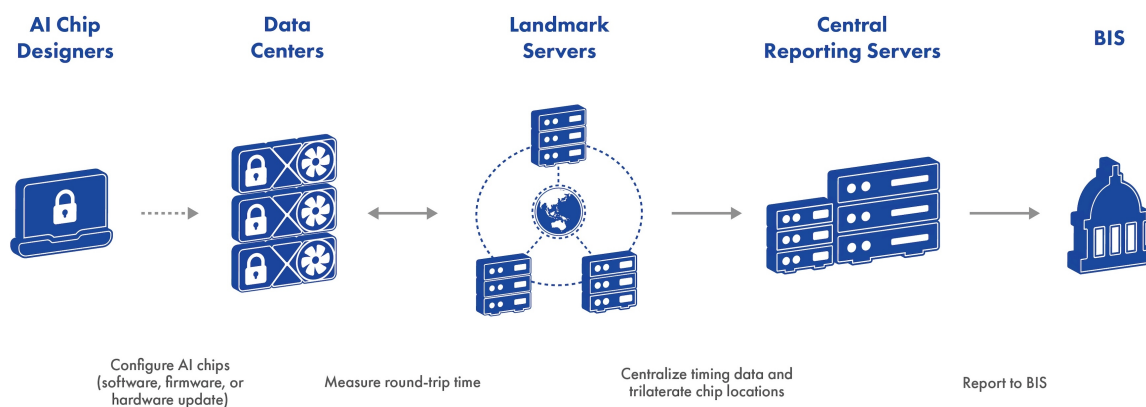
* The protocol should be developed through an international standards-setting organization, allowing relevant stakeholders (e.g., AI chip designers and landmark server operators) to participate and reach consensus.

miles or better, which we consider to be continent-level accuracy. However, our model does not calculate the overlap of multiple landmark servers, which could substantially improve PLV accuracy beyond the worst case (see Appendix B). Still, if BIS judges that trilateration is not accurate enough, delay similarity or a hybrid approach could improve accuracy albeit at the expense of greater complexity and potentially higher cost.

To enable this system, software, firmware, or hardware modifications would need to be made to each controlled AI chip. For chips that have not yet been manufactured, the mechanism could be added during the production process. Chips that have already been produced could be outfitted remotely with software or firmware changes, but this would require each end user to install the update. The challenges associated with enabling a PLV system on a manufactured chip are discussed further in the “Inherent PLV Limitations” section.

To make the PLV system more secure, each controlled AI chip would preferably have and use a trusted execution environment (TEE), which is an isolated area within a processor that securely stores data and runs code.²⁹ A TEE would help ensure each chip has not been tampered with and that the PLV protocol is being performed correctly and securely. However, enabling the TEE may also reduce the performance of AI workloads running on each chip.³⁰ Implementations of the PLV system can proceed without a TEE, but they are less secure. See Appendix E for more details on non-TEE AI chips.³¹

Figure 3. Steps of a PLV System in Practice



Source: CSET illustration.

An important policy question is how secure policymakers want the system to be. There is generally a trade-off between security and costs: the more secure the system is from adversarial attacks, the costlier it is for hardware designers to implement and maintain.

Moreover, while the most secure PLV system is likely to withstand some low-cost attacks, it still comes with limitations and vulnerabilities that sophisticated adversaries can exploit—no system is impervious, especially when attackers have physical access to the chips.³²

Inherent PLV Limitations

While a PLV system may be the most robust option available for verifying the locations of controlled AI chips, the technology has inherent limitations that no implementation can fully eliminate. It is critical that policymakers understand these limitations in order to develop realistic expectations for PLV systems to enhance enforcement efforts. These limitations should therefore inform judgments about whether, and under what conditions, a PLV system's benefits justify its costs. Ultimately, a PLV system will be less effective against diversion tactics that involve changes in end users rather than locations, and it will produce ambiguous signals that require investigative resources to track down.

This section identifies five limitations. First, enabling a PLV system on an already-manufactured AI chip may introduce compliance concerns, as users would be required to update the chip's software or firmware. Second, a PLV system cannot distinguish between a chip that has never been activated or has gone offline for legitimate reasons and one that has lost connectivity due to diversion. Third, resourced and skilled adversaries may be able to spoof location data to make diverted AI chips appear to remain at their intended locations, effectively obfuscating a chip's actual location. Fourth, PLV systems verify approximate, not precise, geographic locations, which means that the mechanism cannot distinguish data centers in close proximity operated by different entities. Lastly, because PLV systems identify locations rather than the identities of users, the system cannot detect unauthorized changes in the end user when the chip's location remains unchanged. Table 2 summarizes each limitation and the implication for enforcement efforts.

Table 2. Summary of Inherent Limitations of PLV Systems

Limitation	Enforcement Implications
Enabling PLV on the Chip	Enabling PLV on already-manufactured AI chips requires end users to update the chip's software or firmware, which introduces compliance issues that can limit coverage.
Lack or Loss of Connectivity	Unresponsive AI chips produce ambiguous signals as enforcement officials cannot distinguish legitimate unresponsiveness from diversion without additional information.
Adversarial Location Obfuscation	Sophisticated actors may spoof location data, causing diverted AI chips to inaccurately appear at their authorized locations.
Approximate Geolocation	PLV systems are highly likely to have an error radius greater than about 20 miles, meaning that the system would be unlikely to be able to distinguish between nearby data centers operated by authorized and unauthorized end users.
End-User Obfuscation	The system cannot detect unauthorized transfers where the chip's physical location does not change, such as colocated data centers or sale of a data center facility to a restricted end user.

Source: CSET.

Enabling PLV on the Chip

To enable a chip to communicate with landmark servers, a software, firmware, or hardware mechanism must be added to each controlled chip. For new (unmanufactured) AI chips, any of the three mechanisms could be incorporated during the production process. However, chips that have already been produced can only be updated remotely with a software or firmware update, which requires each end user to install it, creating a compliance and enforcement challenge. Legitimate operators may take time to comply, and bad actors have an incentive to not install an update that would reveal their chips' locations and to revert the firmware to an older version that does not support PLV. Therefore, policymakers should consider that it could take several years before a majority of controlled chips in operation are PLV-enabled. This gap between policy implementation and real-world effect limits how much value BIS can realistically gain from a PLV system in the early stages.

Lack or Loss of Connectivity

When a controlled chip does not ping landmark servers for an extended period of time, the system cannot determine whether it is due to diversion or legitimate reasons. A chip may be unresponsive for a number of reasons: because it is in storage or in transit to an authorized destination, is disconnected for maintenance purposes, is installed in a legitimate facility that prohibits it from communicating with landmark servers, or has been decommissioned. At the same time, however, it may be unresponsive because it has been diverted to a restricted destination or installed in an unauthorized facility that is deliberately preventing communication with landmark servers. In both cases, it looks the same in the data of the PLV system.

This ambiguity creates a trade-off for U.S. government enforcement efforts. On the one hand, if unresponsive chips are treated as *legitimate*, the system will fail to flag some genuine diversion. Conversely, if unresponsive chips are treated as potentially *illegitimate*, the system will likely generate a large volume of false positives, further straining already under-resourced BIS enforcement efforts.

To address this issue, BIS should require the last exporter to report not only the intended end user and destination, but also the expected timeline for each chip to be activated and whether end users will operate chips in ways that inhibit connections to outside networks (e.g., air-gapped facilities). This would allow BIS to identify chips that have “gone dark” longer than expected and are more likely to have been smuggled. In cases where chips are expected to be turned off or warehoused for long periods of time, physical inspections would likely be the primary way for BIS to confirm that the chips have not been smuggled.* This scenario highlights how physical inspections supplement and address the limitations of PLV and how an effective location verification policy should leverage both methods.

BIS should work with industry to formulate timeframes for when lack or loss of connectivity should trigger further investigation. Collecting and maintaining this data, however, may be difficult in practice as end users may not always know at the time of purchase when a chip will be activated.³³

* Warehoused chips present a practical challenge for physical inspections. To conduct proper physical inspections, auditors must connect chips to workstations to attest their authenticity. But when chips are warehoused, inspectors would need to remove them from their packaging and set them up, which would take more time than if the chips were already in server racks. While we recognize the practical challenges to physically inspecting warehoused chips, we assess it is still feasible with enough resourcing.

Adversarial Location Obfuscation

Bad actors who do not simply disconnect their diverted chips will likely attempt to obfuscate the chips' locations by exploiting vulnerabilities in the PLV system. Sophisticated actors could trick the PLV system into reporting that the diverted chip is still in its intended location, and they could know the intended locations if that information is shared by the smugglers.

Location obfuscation could be conducted in one of three ways: tricking geolocation, attacking landmark or reporting servers, and tricking attestation.³⁴ First, a bad actor who has smuggled an AI chip can easily introduce additional latency into their network to purposefully increase the RTT of a landmark server's ping. This would degrade the system's accuracy for geolocating that chip, making the chip appear to be anywhere within a very wide geographic region.³⁵ Second, a bad actor could directly attack the landmark servers or reporting servers to falsify information. Third—and likely most challengingly—if a bad actor can fool the cryptographic attestation procedure used to authenticate the identity of each controlled AI chip, the bad actor could place fake AI chips in intended locations and divert the real chips.

Individual data center AI chips can cost tens of thousands of dollars, and a rack-scale server containing 72 advanced NVIDIA Rubin GPUs can cost over \$5 million.³⁶ Therefore, at a per-chip value, bad actors can justify spending substantial resources to defeat a PLV system on each diverted chip. Furthermore, if a bad actor can compromise the PLV system at scale (i.e., the vulnerability can easily be exploited repeatedly across many chips), it would undermine the effectiveness of the system.

Even a PLV system designed with prudent security measures could likely be subverted by a sufficiently motivated and resourced actor. As a result, some percentage of chips reported as being in their intended locations may actually have been diverted.

Approximate Geolocation

The PLV system's accuracy is unlikely to be better than 19 miles, meaning PLV cannot reliably distinguish between nearby facilities within that radius (see Appendix B). Data centers are often in very close proximity. For example, there are more than 120 data centers within a ten-mile radius in Paris, over 70 in Singapore, and over 25 in Kuala Lumpur.³⁷ In any of these locations, a chip diverted from one data center to another one within the error radius of the PLV system would very likely go undetected. This issue also applies to data centers located near border regions around restricted countries, where the error radius may encompass both the intended location of the chip and restricted end locations within countries such as China.³⁸

The significance of this limitation depends on the reporting scope of the policy, which is discussed more in the “Policy Design Choices” section. If the PLV system is expected to verify that a chip is at its intended location anywhere in the world (as opposed to only verifying that the chip is not located in China or other countries of concern), imprecise geolocation data significantly degrades the usefulness of the signals the system sends to BIS. For example, it could confirm to BIS that a chip is in Malaysia, its intended location, without being able to determine which data center in Malaysia it is in. This problem is likely most acute in countries with higher levels of diversion—in particular, countries in Southeast Asia.

End-User Obfuscation

PLV systems are intended to verify locations, not end users. Therefore, the PLV system cannot detect an unauthorized change in the end user when the chip’s location does not change. There are two primary scenarios in which end-user obfuscation may occur.

First, a chip is physically transferred between entities that are located in the same data center.³⁹ For example, a license is not required to export a controlled chip to a French-owned cloud company in Paris, but a license is required to transfer that chip to a Chinese-owned cloud company colocated in the same data center (because U.S. export control regulations require licenses for transfers to Chinese-headquartered companies located anywhere in the world). The French company could divert chips to the Chinese company without the PLV system alerting BIS. The chip would briefly go offline during the physical transfer and then reappear in the same approximate location. The temporary shutdown of the chip may not trigger an investigation by U.S. authorities because chips can be temporarily turned off for a variety of legitimate reasons, such as server maintenance. If the chip is not pinged during the period when the chip is temporarily turned off, then this transfer would go entirely undetected.

Second, a chip is never physically moved at all. Instead, a legitimate entity in a permitted location sells its facility to a restricted end user. For example, a French-owned cloud company (or its customers) would sell its colocation data center setup, including the controlled AI chips, to a Chinese company. In this situation, the chips remain continuously online at their authorized location, and the PLV system does not detect any diversion.

In addition, the PLV system would not detect when a legitimate end user provides a Chinese company with remote access to their chips. This remains a loophole in the current U.S. export control system.

These scenarios show that the PLV system is unable to detect chip diversion that involves a change in end user, but not in location. Addressing end-user diversion in either situation would require additional regulatory mechanisms beyond location verification, such as Know Your Customer due diligence requirements or restrictions on remote access of controlled chips.

Summary

Ultimately, a PLV system offers the most useful and technically feasible option for location verification of controlled chips. It is more scalable than physical inspections, more accurate and verifiable than mail-in inspections, video inspections, and inventory management tools, and more robust against adversarial attacks than alternative geolocation mechanisms, such as GPS-based and topology-based mechanisms. Nonetheless, the PLV system has inherent limitations that constrain the enforcement value to regulators, including creation of ambiguous signals when chips go offline, susceptibility to spoofing by well-resourced adversaries, verification only of approximate locations, and inability to verify end-user identities. These limitations do not fully eliminate the value of a PLV system and do not necessarily mean the costs of location verification outweigh the benefits, but they do underscore that the system's usefulness will depend on how it is implemented.

PLV Benefits, Costs, and Policy Design Choices

Even if PLV is the most feasible and cost-effective option for location verification, policymakers and BIS regulators must still determine whether the system's enforcement benefits justify the costs that it would impose on industry, regulators, and the broader global chip ecosystem. To that end, we modeled the costs of a PLV system and assessed a series of regulatory design choices that would shape the system's effectiveness. To provide regulators and policymakers with a framework for evaluating current and future policy proposals, this section begins by assessing the baseline costs and benefits of any PLV system and then considers the key implementation questions that further affect the cost-benefit trade-offs (see Table 4 for a summary of the benefits, costs, and our recommendations).

Benefits and Costs of a PLV System

Policymakers should consider the trade-offs between the benefits that a PLV system provides for export control enforcement and the costs to procure, build, configure, and run all its components.

Benefits

A PLV system would provide two key benefits to export control enforcement. First, it would help BIS generate leads by identifying chips that are outside of their intended locations. This data could improve BIS's understanding of diversion patterns and volumes, which in turn could help the Bureau write more targeted rules to prevent future diversion and smuggling. It is important to note, however, that a PLV system does not directly prevent smuggling, but instead detects it while ongoing, or more likely, after the fact. Therefore, the enforcement value of a PLV system largely depends on whether BIS is sufficiently resourced to investigate the leads (either directly or through a contracted third-party auditor), as well as the number of false positives that result in investigative dead ends and false negatives that stem from the PLV system's inherent limitations.

The second benefit of a PLV system is that it imposes costs on chip smugglers, which may provide a meaningful deterrent effect. Currently, if smugglers trick the last exporter with false documentation and the chips are physically brought into their control, then BIS can only identify them through conventional investigative means (e.g., audits, investigations, tips). A PLV system would be an additional hurdle that smugglers must overcome, as they must either expend resources to undermine the PLV system and/or exploit its inherent limitations (e.g., end-user obfuscation or location obfuscation).

Costs

Our cost modeling suggests a PLV system could be deployed for a feasible but non-trivial cost, depending on implementation. We estimate that a PLV system would cost roughly \$3.1 million to \$28.8 million annually if the landmark servers are owned and \$2.6 million to \$72.4 million annually if they are rented. The potentially government-borne costs would be roughly between 0.5% and 7.2% of BIS's FY2026 budget if landmark servers are owned or 0.3% and 25.7% if they are rented.⁴⁰ If BIS receives the \$450 million it requested for FY2027, a PLV system with owned landmark servers would cost 0.2% to 3.7%, while a PLV system with rented landmark servers would cost 0.1% to 13.4%.

The baseline costs of a PLV system include the cost to (a) stand up and operate landmark server infrastructure; (b) maintain the full PLV system, including the landmark servers; (c) develop a technical standard for the PLV protocol, which would ensure interoperability between chips from different vendors; and (d) design and update the software, firmware, or hardware of chips to communicate with landmark servers. We estimated lower and upper bounds for these four types of cost.

Table 3. Lower- and Upper-Bound Cost Estimate by PLV Cost Type

Cost	Borne by...	Cost per Year	
		Owning	Renting
Landmark Server Purchase/Renting and Operation^A	Industry or U.S. Government	\$94k–\$11.81m	\$493k–\$56.68m
PLV Maintenance		\$1m–\$5m	\$100k–\$3.75m
Follow-on Enforcement	U.S. Government	–	–
Standards Development^B	AI Chip Designers	\$1m–\$2m	
Firmware Development^B		\$1m–\$10m	
Disruption to AI Chip Hardware or Firmware Design		–	–
AI Chip Performance Degradation	AI Chip Users	–	–
Total Annual Cost		\$3.1m–\$28.8m	\$2.6m–\$72.4m

Notes: We do not estimate the cost of follow-on export control enforcement, changes to chip designers' development and security roadmaps, and AI chip performance degradation.

(A) For this cost distribution, we assume that landmark and reporting servers would be operated by BIS or a trusted third-party contractor.

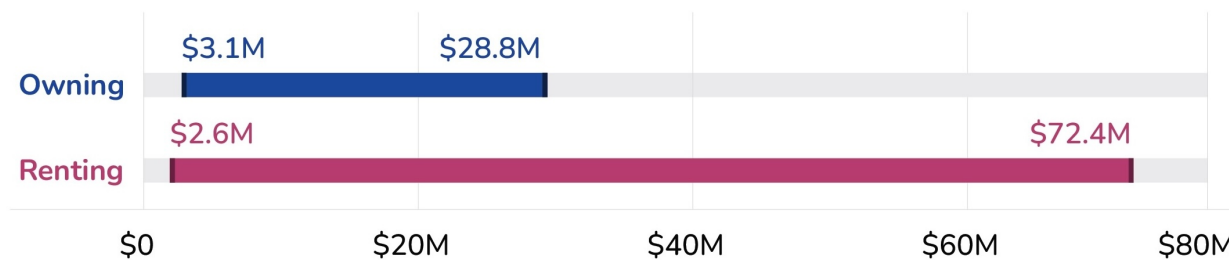
(B) Standards and firmware development are one-time fixed costs amortized over five years. We estimate that standards development costs range from \$5–10 million (\$1–2 million per year) and firmware development costs range from \$5–50 million (\$1–10 million per year).

Source: CSET Location Verification Cost Model (see Appendix D).

The landmark servers in a PLV system can either be rented from CSPs or owned and operated directly by BIS or a third party. We find that the cost ranges of renting versus owning landmark servers largely overlap. The lower-bound cost of renting is slightly below that of owning, but renting has a much higher upper bound than owning, which means that one is not definitively cheaper than the other. We assess that the annual cost of a PLV system, when owning 39–500 landmark servers, likely ranges between \$3.1 million and \$28.8 million, while a system that rents the same number of servers from a CSP likely ranges between \$2.6 million and \$72.4 million.⁴¹ Using our estimate that approximately three million data center AI chips are in circulation that are not (a)

sold to China or (b) owned by U.S.-headquartered CSPs or other U.S. hyperscalers, the costs of owning or renting servers are between \$1.03 and \$9.57 per chip and between \$0.86 and \$24.14 per chip, respectively. To see the full details of our cost modeling, see Appendix D.

Figure 4. Annual Cost Range Estimates by Landmark Server Procurement Model



Source: CSET Location Verification Cost Model (see Appendix D).

We do not model three other indirect costs of PLV: the cost of AI chip performance degradation, changes to chip designers' development and security roadmaps, and follow-on export control enforcement (e.g., follow-on physical inspections after a PLV system detects potential diversion). We omit these indirect costs because they are difficult to quantify in dollars and are idiosyncratic. Nonetheless, it is important for policymakers to consider how these costs influence the value of a PLV system.

First, PLV could potentially degrade the performance of controlled AI chips if PLV processing competes with AI workloads for common computation or communication resources. Frequent communication with landmark servers (e.g., multiple times a day) may interfere with the chips' regular processing. Those designing the system should consult with industry technical experts to identify the proper frequency of landmark server-to-chip communication to ensure performance issues are limited. Overall, we assess that AI chip performance will not be significantly degraded if the PLV protocol is designed properly.

Second, requiring PLV could disrupt ongoing chip development, especially if designers are forced to slow down their chip R&D cycle to make hardware changes for PLV or reallocate limited resources to write firmware for PLV. This cost depends on the degree to which the hardware or firmware components of a PLV system fit, or do not fit, within a given company's pre-established R&D roadmap.

Third, BIS may need additional funding to act on the export enforcement leads that PLV generates. While enforcement capacity is a critical consideration for making productive use of PLV leads, we found it would be too difficult to estimate those costs in our model.

Taken together, a PLV system offers BIS useful information to enhance enforcement efforts at a feasible but non-trivial direct cost. We did not quantify additional indirect costs that are difficult to model without further industry input. How the costs and benefits balance out depends, in part, on the design choices we examine in the following section.

Policy Design Choices

The value of a PLV system depends heavily on a series of policy design choices that shape how the system is built, who operates it, and how broadly it applies. We consider five questions policymakers must answer: (1) the level of security the regulator should prescribe for how chips communicate with landmark servers, (2) which chips should be subject to location verification, (3) who should be responsible for maintaining the landmark and reporting servers, (4) whether the policy should apply worldwide or only to countries of concern, and (5) whether chips operated within the U.S. could and/or should also be tracked as part of an always-on PLV system. The table below summarizes these choices, our recommendations, and the pros and cons for each recommendation.

Table 4. Pros and Cons of Recommendations for Policy Design Choices

Design Choice	Recommendation	Pros	Cons
Security Level (resistance to adversarial location obfuscation)	Firmware-based signing	• Firmware-based signing reduces the attack surface and increases the reliability of location data relative to software-based signing • Firmware-based signing imposes lower implementation costs relative to hardware-based signing • Implementable on chips already in circulation via remote firmware updates	• Firmware-based signing imposes higher implementation costs relative to software-based signing (chip designers must be involved) • Firmware-based signing may have unidentified vulnerabilities that may be exploitable by sophisticated adversaries seeking to obfuscate chip locations
Chips Subject to Location Verification	Data center chips subject to the most stringent licensing requirements	• Aligns location verification policy scope with export control policy scope • Concentrates effort on the chips most likely to be diverted by adversaries	• Leaves lower-performing data center and non-data center chips untracked
Operation of Landmark and Reporting Servers	BIS-certified trusted third party	• Ensures consistent trilateration methods and reporting thresholds • Eliminates conflicts of interest in chip designers self-reporting • Draws on private sector expertise while BIS retains oversight via certification of third-party auditors	• BIS required to stand up and maintain a certification and oversight process of third-party auditors • Server operating costs must be funded (open question whether the seller, buyer, or government bears the costs)
Reporting Scope	Worldwide scope	• Enables detection of diversion to Chinese-owned data centers located outside of China	• Requires full supply-chain tracking of intended locations, increasing compliance costs and the number (and distribution) of landmark servers
Always-On System (tracking of domestic chips)	No recommendation	• Closes a significant vulnerability in the PLV system that would otherwise prevent detection of U.S.-origin chips resold or diverted to controlled end users	• Lacks clear legal authority under current export control law • May require expansion of ICTS authorities to cover transactions without a foreign adversary nexus • Raises privacy concerns and lacks precedent

Source: CSET.

Security Level

What level of security should policymakers prescribe?

The security and robustness of a PLV system (i.e., how resistant the system is to adversaries falsifying a chip's location) depend largely on how AI chips cryptographically sign the round-trip time (RTT) challenges sent by landmark servers. More secure signing mechanisms make it harder for adversaries to trick the system into signing false location data, but they are also costlier to implement. Policymakers face two key considerations in balancing this trade-off: (1) the cost-security balance across various implementation options and (2) other unknown or unremediated vulnerabilities that persist regardless of how it is implemented.

Regarding the cost-security balance, there are three ways this signing process can be implemented on AI chips that have TEEs. From lowest to highest security, the three approaches are software-, firmware-, and hardware-based signing. First, **software-based** signing is where the chip responds to and signs the RTT challenge in software running within the user-accessible part of the chip's TEE. Although this configuration is the lowest effort of the three options because it only requires development and installation of new software, it is also the least secure; the chip's response to RTT challenges is not signed by a key derived from the chip's "hardware root of trust" and therefore could be more easily falsified. Second, **firmware-based** signing is where the chip's response to an RTT challenge is generated and signed in the firmware of the chip's TEE. This would involve hardware designers updating the TEE firmware to generate and sign a response to the RTT challenge. Third, **hardware-based** signing is where the chip designers add dedicated hardware to the chip to underpin signing of RTT challenge responses. For the first two options, the hardware designers would need to remotely update their chip software and/or firmware to sign RTT challenges. But hardware-based signing would require changes to the chips' design and manufacturing process, so it could only be implemented on chips not yet manufactured. See "Step 2: Configure AI chips to Communicate with Landmark Servers" in Appendix A for more details.

A practical complication is that many AI chips lack TEEs, and because TEEs rely on structures physically fabricated in the chip, they cannot be added to previously produced chips. Implementations of the PLV system can proceed without a TEE, but they are overall less secure (see Appendix E for more details).⁴² Policymakers should carefully consider how the chip industry—which produces a range of heterogeneous chips that would all require some sort of software, firmware, or hardware update to sign RTT challenges—would be able to implement a PLV system with an acceptable

level of security. Moreover, if the security provided by TEEs is deemed necessary for the PLV system, policymakers should consider the additional costs incurred by hardware designers whose AI chips currently lack TEEs, such as AMD.

Regardless of which signing mechanism is chosen, unknown or unremediated vulnerabilities will exist. No system is impervious, and even the most secure hardware-based signing mechanism is susceptible to attacks such as TEE.fail (which is exploited to compromise TEE security).⁴³ Policymakers should understand that any system may have vulnerabilities—especially unknown vulnerabilities that can arise when bad actors have physical access to the chips. Any implementation of a PLV system is unlikely to withstand advanced persistent threat actors (i.e., threat actors operated by nation states or state-sponsored groups) that have the resources and expertise to identify and exploit vulnerabilities in the PLV system or to compromise the organizations that maintain the PLV system. Prescribing a higher security level reduces—but does not eliminate—risk and raises the cost to undermine the system, so the security gains of each level should be weighed against the costs.

Weighing these considerations, firmware-based signing offers the most workable approach, balancing security, cost, and ease of implementation. This approach provides a strong middle ground as it avoids signing in the user-accessible software and can nonetheless be deployed on chips already in circulation through remote firmware updates. Because the added security benefits of hardware-based signing are also unlikely to eliminate vulnerabilities, particularly against advanced persistent threat actors, firmware-based signing represents the most secure yet feasible path to a PLV system.

Chips Subject to Location Verification

Which chips should be subject to location verification?

Policymakers will need to determine which export-controlled AI chips should be subject to location verification. The primary export control classification numbers (ECCNs) that define export-controlled chips (3A090, 4A090, and 5A002.z) include two categories of chips that present different considerations for a PLV system.⁴⁴

The first category includes **data center AI chips above a specified performance threshold**.⁴⁵ This category encompasses a wide range of performance levels, from the most powerful NVIDIA GPU on the market—Rubin—down to AI chips with 1.1% of Rubin’s normalized peak performance.⁴⁶ The highest performing chips in this category face the most stringent controls, meaning exports of these chips are subject to license requirements for over 40 countries. The lower performing data center chips in this

category face licensing requirements to U.S.-arms-embargoed countries only.⁴⁷ BIS could require PLV tracking across both categories or set a threshold to focus on only the most capable chips.

The second category includes **high-performance chips that are not designed for data centers** (i.e., powerful gaming and workstation chips).⁴⁸ These chips also face license requirements to U.S.-arms-embargoed countries. Requiring PLV tracking for these chips is considerably more challenging because the end-user base is much more diffuse than for data center chips. This creates three challenges. First, if PLV tracking were implemented in software, it would be difficult to enforce compliance (i.e., installation of the software) across so many actors. Second, it would be more difficult to determine the expected location of non-data center chips than data center chips because the latter are deployed by far fewer actors.⁴⁹ Third, extending PLV to non-data center chips would mean tracking consumer-grade hardware around the world, introducing privacy concerns.

Due to these challenges of tracking non-data center chips, we recommend that the location verification system focus on only data center chips subject to the most stringent licensing requirements (i.e., the highest-performing chips for which export licenses are required for over 40 countries). This would mean the system focuses on the most capable chips, which adversaries have the strongest incentive to divert. The fact that lower-performing data center chips only face licensing requirements to U.S.-arms-embargoed countries suggests that policymakers are less concerned about diversion of such chips. Therefore, extending location verification to lower-performing data center chips would be misaligned with current policies. Similarly, extending it to cover non-data center chips risks increasing the compliance costs without a proportional enforcement value. Wherever BIS (or Congress) draws the line, however, there will be strong demand for chips just below that threshold and therefore an incentive for chip designers to develop chips just beneath it, as long as there is demand for that performance level.

Beyond the thresholds for which types of chips to track, there are three additional considerations: how to identify decommissioned chips, whether to provide special exemptions for certain types of end users, and when verification requirements should be sunsetted for certain chips.

First, BIS will need to find a way to identify chips that are not responding to PLV pings because they have been legitimately decommissioned. AI chips typically are decommissioned five to eight years after installation (although some sources claim shorter lifespans).⁵⁰ For example, consider if NVIDIA A100s (first released in 2020)

were subject to a PLV system today. A portion of the global A100 fleet has likely already been decommissioned. From the perspective of the PLV system, decommissioned A100s would appear non-responsive, and BIS would not know whether any given non-responsive chip was legitimately decommissioned or whether bad actors are purposefully inhibiting the chips' connectivity. This problem also exists for chips manufactured after the instantiation of a PLV system. In either case, assuming that BIS chooses to interpret a non-responsive chip as potential diversion, each decommissioned chip that is not identified as decommissioned becomes a false positive, adding noise to the system.

Second, it may be warranted to exempt chips from the PLV system for certain end users, particularly for U.S. and allied military, intelligence community, and other government end users running AI chips in classified settings. A PLV system that geolocates such chips would almost certainly raise security concerns for those organizations. However, an exemption would likely necessarily involve BIS (or a trusted third party) maintaining a ledger of exempted chips by organization, which itself could present a security concern. The best way to resolve these concerns is unclear.

Lastly, policymakers should also consider when to sunset location verification requirements for certain chips. Because we suggest aligning the scope of location verification with the scope of U.S. export controls on AI chips, one logical approach BIS may take is to sunset the requirement when U.S. export controls on certain chips are revised due to foreign availability assessments; for example, when Chinese-designed chips reach a comparable quality (and are produced in sufficient quantities) that U.S. export controls on that item would be considered ineffective, BIS may remove license requirements for that item and concurrently sunset location verification.⁵¹

Operation of Landmark and Reporting Servers

Which entities should policymakers designate to operate landmark and reporting servers?

Policymakers will need to decide how much ownership BIS will take of operating the PLV system and how much will be delegated to the chip designers or a trusted third party. This choice involves determining (1) who operates the landmark servers that communicate with controlled chips and receive geolocation data, and (2) who operates the reporting servers that trilaterate chip locations. These responsibilities could be assigned to the same entity or split between different entities.⁵² Each arrangement involves trade-offs related to conflicts of interest and data consistency.

On the one hand, industry may prefer that a trusted third party compare the data rather than the government, so that specific location data need not flow directly to BIS unless discrepancies arise. On the other hand, trusting private companies with such data also raises questions about oversight and accountability. For this report, we assume the party operating the reporting servers would also be responsible for comparing the data and reporting to BIS whenever a chip's location fails to match its expected location.

If chip designers operate the reporting servers, they are effectively self-reporting on whether their chips (albeit not in their possession) have been diverted. In that role, the chip designers may face reputational consequences, enforcement scrutiny, or even friction with customers. While firms may be trustworthy, the structural incentives would not be aligned. Ultimately, centralizing the reporting servers under BIS or a trusted third-party contractor would lead to fewer conflicts of interest.

Even if a standardized PLV protocol governs how AI chips exchange pings with landmark servers, it may not specify how a set of RTT measurements are converted into a location estimate. Without separate standards, different reporting-server operators may develop their own trilateration methods and choose their own thresholds for flagging estimates that fall outside the expected geographic area. Differences in either choice could result in inconsistent reporting to BIS. BIS could address this issue by standardizing the location estimation methodology and establishing clear flagging thresholds.

On balance, the strongest configuration is likely one in which a BIS-certified trusted third party operates both the landmark and reporting servers. A trusted third party would eliminate conflicts of interest that arise when chip designers are entrusted with operating the servers. Consolidating operations under a single entity or small number of entities ensures consistent trilateration methods and reporting thresholds. Contracting with a third party also allows BIS to draw on existing private sector expertise.

There are several options for who pays the third-party contractor: the buyer, the seller, or the U.S. government. These options are all feasible, and ultimately this must be decided by policymakers and appropriators. Currently, in the export control context, exporters (i.e., sellers) already bear the costs of conducting due diligence on their customers and maintaining compliance programs.⁵³ Therefore, AI chip exporters, which are sometimes but not always the chip designer, could be required to contract with and pay the accredited third party.⁵⁴ Alternatively, the buyer (or last exporter) could be required to contract with the third party, which has to be demonstrated to the seller before any sale can proceed.

The primary cost driver for a PLV system is that it requires a geographically distributed network of secure landmark servers that must be configured and operated. As shown in Table 3, we estimate that the annual potentially government-borne cost of owning landmark servers for a PLV system likely ranges between \$1.1 million and \$16.8 million. Conversely, the annual potentially government-borne cost of renting the servers from CSPs could range between \$0.6 million and \$60.4 million.⁵⁵

Reporting Scope

Should policymakers seek a worldwide or country-of-concern reporting scope?

The PLV system could be designed to detect and alert BIS to chip diversion anywhere in the world or only when a chip appears in a country to which exports are restricted, such as those subject to U.S. arms embargoes (i.e., Country Group D:5, which includes China).⁵⁶ The difference between the two options is primarily about what triggers a reporting requirement to BIS and what exporters are required to report. The choice has implications for the system's enforcement value and the costs it imposes on industry and BIS.

Under a **worldwide scope**, the PLV system would compare each chip's actual location to its intended location recorded in the reporting server. Any discrepancy between the locations would trigger an alert to BIS. For example, if a chip intended for use in Singapore is detected in Malaysia, BIS would be notified even though Malaysia is not a restricted country. Under a **country-of-concern** scope, the PLV system would notify BIS only when a chip is located in a country to which exports of controlled chips are restricted, such as those in Country Group D:5.⁵⁷

In practice, BIS would likely implement either option using a broad licensing requirement accompanied by a license exception. Instead of requiring approval for individual license requests, BIS can offer license exceptions that let exporters bypass the licensing requirement if they satisfy certain conditions. A worldwide reporting scope would likely condition the license exception on both the implementation of a PLV system for a controlled chip and notification to BIS whenever the chip changes hands. The second condition under a worldwide scope is essential to be able to detect when a chip is not in the location it is intended to be following each resale or transfer of the chip. A country-of-concern scope, by contrast, would be conditioned solely on implementation of a PLV system, and it would not require notifications for downstream sales. Detecting that a chip is in a restricted country requires no knowledge of where the chip was supposed to be.

Based on the relative costs and benefits of a worldwide versus country-of-concern scope, we recommend that policymakers pursue a worldwide policy scope.

A worldwide scope would allow BIS to detect physical diversion of chips anywhere in the world, enabling the agency to better identify trends in smuggling routes and improve its global enforcement efforts. Critically, this would allow BIS to detect diversion of chips to Chinese-owned data centers located outside of China, which could be used by Chinese companies to access advanced chips. This data could help BIS write more targeted rules to prevent future diversion and smuggling and to better facilitate cooperation with foreign enforcement agencies. Additionally, the risk that any chip used in an unauthorized location could trigger BIS scrutiny would likely have a general deterrent effect on smuggling, though the level of deterrence is hard to forecast and measure.

A country-of-concern scope, by contrast, would leave significant knowledge gaps for BIS. While it would identify diversion to the highest priority locations, it would not identify diversion to Chinese-owned data centers in non-restricted countries (e.g., Singapore and France), and it would not help BIS better understand and track broader diversion patterns. This approach may also inadvertently incentivize diversion to Chinese facilities abroad, since smugglers would know that chips owned by Chinese companies but located in non-restricted jurisdictions would not be reported to BIS.

Relative to a country-of-concern scope, a worldwide scope would only impose a marginal increase in compliance costs on end users who would be required to report when a chip is moved to a different facility or resold. Therefore, the benefits of a worldwide scope outweigh the additional compliance costs.

Always-On System

Should policymakers require an always-on system that tracks chips in the United States as well?

Under current export control authorities, BIS cannot legally require location verification for chips wholly designed, produced, and sold within the United States (i.e., “U.S. chips”) as such transactions do not involve an export. Excluding U.S. chips, however, would leave BIS (and the PLV system) blind to the locations of a substantial portion of controlled chips, and if those chips were diverted, the PLV system would not alert BIS.⁵⁸

As a result, an always-on system for U.S. chips is likely the most effective approach, but it is intrusive, introduces privacy concerns, and lacks precedent. An always-on system would likely raise privacy and national security concerns as it could expose the location

of chips operated by the U.S. government and military as well as foreign governments. These concerns could potentially be mitigated by exempting those chips from the PLV system.*

In any case, an always-on requirement cannot be implemented with existing export control authorities, and BIS would need to rely on other authorities, such as Information and Communications Technology and Services (ICTS), to regulate U.S. chips.⁵⁹ The current ICTS program, however, only regulates items and technology with a connection to a foreign adversary, and therefore likely cannot regulate U.S. chips, as defined.⁶⁰ Requiring an always-on system for U.S. chips would likely require modifying the ICTS program. Because that program relies on the International Emergency Economic Powers Act (IEEPA), congressional action may also be required if the authorities granted by IEEPA are deemed insufficient to authorize a version of the ICTS program that applies to transactions without a connection to a foreign adversary.⁶¹

* See section “Chips Subject to Location Verification” for more details.

Case Study

To illustrate both the potential value and limitations of a PLV system, this section considers how such a system would have impacted a real-world chip smuggling operation that was uncovered by the U.S. government.⁶² Specifically, we look at a recent smuggling operation in which the defendants allegedly conspired to divert \$2.5 billion of servers containing AI chips to China across multiple shipments. We find that a PLV system could not have entirely prevented the initial shipment of AI chips from being diverted to China, but it may have alerted U.S. officials to the scheme earlier, potentially helping the U.S. government disrupt the operation more quickly.

The Smuggling Scheme

Between 2024 and 2025, a Supermicro co-founder and vice president, along with two Taiwanese citizens, allegedly conspired to divert \$2.5 billion worth of servers housing export-controlled chips to China. The servers were smuggled in a series of shipments over the course of more than a year, meaning that earlier shipments already reached China while subsequent orders and shipments were still being placed and diverted. Supermicro originally purchased the chips, packaged them into server racks in the United States, exported them to Supermicro facilities in Taiwan, and then shipped them to a “pass-through” company in Southeast Asia that was operated by the smugglers. Once in Southeast Asia, smugglers used a separate company to repackage the servers into unmarked boxes and shipped them to China.

To facilitate the scheme, the smugglers provided false documents to hide the true identity of the pass-through company in Southeast Asia, arranged for favorable Supermicro auditors to inspect the chips and end location, and staged dummy servers to circumvent U.S. government checks. From Supermicro’s perspective, its auditors conducted proper inspections of the servers, and the pass-through company was deemed the legitimate end user of the chips. When the chips arrived at the pass-through company’s warehouse in Southeast Asia, dummy servers were placed into the original server boxes, along with the proper labels and serial numbers. Auditors from both Supermicro and the U.S. Department of Commerce went to the warehouse, inspected the server boxes that now housed the dummy servers, and unboxed them for inspection. Because the proper labels and serial numbers were in place, along with having “favorable” Supermicro auditors, the investigators concluded that the real chips were in the intended end location.

Failure of Physical Inspections

This case illustrates why physical inspections of packaging or serial numbers are insufficient as a verification mechanism. The smugglers successfully deceived auditors who conducted on-site inspections with relatively unsophisticated measures such as transferring labels and serial numbers to dummy servers and repackaging them in authentic boxes. Proper physical inspections should involve attestation of the chips themselves (i.e., not the packaging), although this can be far more time consuming when chips are warehoused because attestation requires unpacking and turning on each unit.

Value of a PLV System

One potential benefit of a PLV system is that it could have a deterrent effect on smugglers, discouraging diversion of AI chips in the first place. The deterrent effect is inherently difficult to quantify, and it is unclear whether it would have prevented this particular scheme.

Assuming the smugglers would not have been deterred, the PLV system alone could not have prevented the diversion of the initial shipment of this particular scheme or helped auditors detect the diversion during inspection. How a PLV system would have impacted this scheme and aided BIS largely depends on when and where the chips were turned on. Because the servers were turned off and remained in their packaging during transport and storage, a PLV system would not have been able to communicate with the chips and trilaterate their location before they were turned on by the end users in China. A PLV system would also not have impacted the earlier phases of the scheme, including the provision of false documentation and the decision by Supermicro to ship to a pass-through company in Southeast Asia.

However, if the initial shipment of chips was activated in China and able to communicate with landmark servers, a PLV system could have generated an enforcement lead for BIS and other enforcement agencies. As the smuggling operation consisted of a series of shipments, some chips smuggled in the first shipment could have been turned on in China before subsequent shipments had been diverted. Once the smuggled chips pinged the landmark servers from a restricted destination, the reporting servers would have flagged the discrepancy to BIS.

BIS could then investigate this lead by tracing the sale: initially, officials would likely only know the unique identifiers for each diverted chip as well as the company that designed them (i.e., NVIDIA). First, officials would reach out to NVIDIA to identify the server manufacturer to which the company sold the chips. After identifying Supermicro

as the server manufacturer, officials would reach out to identify the company's customer, which was the pass-through company in Southeast Asia. Finally, BIS would need to investigate the pass-through company to determine whether it diverted the chips to China or whether it resold to another legitimate company. Ultimately, even with the lead provided by the PLV system, a significant amount of time and effort is required to identify the offending party and prevent the diversion of future shipments.

Importantly, this assumes that the chips had not been tampered with, their communication with landmark servers was unobstructed, and the PLV system was functioning as intended. But if the Chinese end users prevented the chips from pinging the landmark servers or obfuscated their locations, then nothing would have been flagged to investigate unless BIS knew when the chips were supposed to be activated. Bad actors could operate the chips in an air-gapped facility or a local network not connected to the internet, establish firewalls to block any communication with landmark servers, or find and exploit vulnerabilities in the PLV system to obfuscate the chips' locations. From BIS's perspective, the chips would either be turned off in the warehouse of the pass-through company or pinging from data centers in the intended end location.

This scenario highlights the importance of chip activation timelines, as smugglers and illicit end users will likely prevent chips from pinging landmark servers once they have been smuggled. BIS should know roughly when chips are intended to be activated, which could be required as part of an export license, so an investigation can be triggered when chips fail to communicate with landmark servers for long periods of time. To avoid smugglers undermining the system with long activation timelines, BIS should consider capping the activation timeline.

Taken together, the PLV system would not only impose costs on the smugglers if they sought to avoid detection but also generate an enforcement lead whenever chips fail to ping landmark servers, which physical audits or other standard investigative methods could then follow up on.

Conclusion

Location verification for advanced AI chips may enhance the enforcement of U.S. export controls, but policymakers should approach it with clear expectations of what such a system can and cannot achieve and at what cost.

Our analysis suggests that if policymakers choose to pursue location verification, a centralized ping-based location verification (PLV) system—supplemented by physical inspections to investigate leads—offers the most favorable balance of feasibility, cost-effectiveness, and robustness against adversarial attacks.

Across more than ten million simulated scenarios, our cost modeling found that physical inspections do not detect more diverted chips per dollar than PLV in any case. Standing up and running a PLV system is not entirely cost-prohibitive for BIS, but the costs are nonetheless non-trivial and should be taken into consideration. Taking all costs together, a PLV system would cost roughly \$3.1 million to \$28.8 million annually if the landmark servers are owned and \$2.6 million to \$72.4 million annually if they are rented. If the U.S. government bears the costs of landmark servers, owning the servers would represent 0.5% to 7.2% of BIS's FY2026 budget, while renting the servers would represent 0.3% to 25.7%. If BIS receives the \$450 million it requested for FY2027, the maximum government-borne cost would be 13.4% of BIS's budget.

Additional non-quantifiable, indirect costs including follow-on enforcement would require BIS to be well-resourced enough to act on the leads created by the PLV system. Therefore, before implementing a PLV system, policymakers should ensure that BIS has sufficient resources and staffing to be able to effectively use the information that such a system would provide. A recent 23% budget increase to \$235 million represents a positive development, but if BIS remains resource-constrained, the utility of a PLV system that produces additional leads (and potentially a high volume of false leads) will be partially undermined by the capacity challenges BIS already faces. While these constraints alone do not preclude a PLV system, it is important for BIS to have sufficient resources to take full advantage of it.⁶³

Even a cost-effective and cost-feasible PLV system has inherent limitations. A PLV system cannot prevent diversion as it can only detect diversion after the fact and under certain conditions. The system cannot verify the locations of chips that remain offline and cannot detect when chips are transferred within colocated data centers or between data centers that are located within the margin of error of the PLV system. Nor can a PLV system detect when chips are being used by unauthorized end users in authorized locations.

As the Supermicro case study illustrates, a PLV system could potentially have generated an earlier enforcement lead for investigators, though this would have depended on whether and when the chips were used and established connectivity once smuggled into China. In any case, it is important to recognize that a PLV system could not have prevented the initial diversion. Moreover, the usefulness of the system in this case would have been dependent on whether BIS possessed the capacity to investigate the alert and how quickly enforcement officials could have done so.

The system's effectiveness will also depend on a series of policy design choices, such as the prescribed security level, the scope of chips covered, the entities designated to operate the reporting and landmark servers, the reporting scope of the policy, and whether domestically deployed chips are also tracked. Evaluating these choices requires policymakers to carefully weigh the costs to the U.S. government and industry against the enforcement gains.

In sum, location verification is one tool among many for export control enforcement. Its value will depend not only on the strength of the technical and regulatory implementation but also on how the system can be integrated into BIS's overall enforcement efforts.

Authors

Jacob Feldgoise is a senior data research analyst at CSET, where his research focuses on AI chip supply chains, export controls, and emerging technologies talent.

Kyle Miller is a senior research analyst at CSET, where he works on the CyberAI Project focusing on the impact of open models and AI diffusion, dynamics around AI accessibility, and compute.

Hanna Dohmen is a senior research analyst at CSET, where her research focuses on U.S.-China technology competition, AI chip supply chains, and export controls.

Acknowledgements

The authors thank Andrew Lohn, Chris Rohlf, Saif Khan, Onni Aarne, Erich Grunewald, Emanuele Del Sozzo, Kevin Wolf, Steph Batalis, Katherine Quinn, Rebecca Gelles, Matt Mahoney, Catherine Aiken, Owen Daniels, Danny Hague, Susan Moynihan, and Shelton Fitch for their feedback on this report.

AI Disclosure Statement

The authors disclose that they used generative AI tools (i.e., ChatGPT and Claude) to identify typos and errors in the manuscript, to suggest wording improvements, to generate draft code blocks, and to draft figures. All code used in support of this paper was human-reviewed, and all figures used in this report were human-designed. The authors take full responsibility for the integrity of all content.



© 2026 by the Center for Security and Emerging Technology. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc/4.0/>.

Document Identifier: doi: 10.51593/20260002

Appendix A: Approaches to Chip Location Verification

There are five categories of methods to help identify the location of a chip: physical inspections, mail-in inspections, video inspections, inventory management tools, and geolocation mechanisms.

Physical Inspections

Location verification can be performed by physical checks involving recurring on-site audits by law enforcement personnel or contracted investigators. Proper checks would involve removing chips from server racks, comparing their unique identifiers (e.g., serial numbers) against a list of chips expected to be at that location, and plugging them into a workstation to attest that the hardware and firmware are authentic.⁶⁴ This would involve using software toolkits such as nvTrust to attest the chips' authenticity. Merely inspecting serial numbers is insufficient, as bad actors can label dummy chips with the serial numbers of GPUs that are supposed to be in the data center, as demonstrated by the case study.

It would be challenging for a bad actor to defeat physical checks if the chips are properly inspected, but proper inspections come with downsides. While reliable, physical checks are resource-intensive because they require many auditors counting large quantities of chips across geographically dispersed data centers. They are also intrusive and disruptive as they require temporarily deactivating and removing chips from server racks to attest their authenticity. Furthermore, physical checks can be complicated if chips remain in their packaging, as a proper check would require removing chips from their packaging to attest their authenticity.

Overall, we find that physical checks are less cost-effective than PLV for verifying the locations of AI chips, but they may be an effective supplement to PLV (see Appendix D for our cost-benefit analysis).

Mail-in Inspections

Chip locations can be identified (and potentially verified) through mail-in inspections, where law enforcement personnel or contracted investigators require certain AI chip end users to mail a random selection of export-controlled chips to an inspection location.⁶⁵ Inspectors would calibrate the shipping window to the expected shipping time from the chip's expected location to the nearest inspection location (likely an office of the U.S. Commercial Service or a U.S. Embassy or Consulate) so that each end user is given only as much time as shipping legitimately requires; whether the chip arrives within that window would help determine whether the chip was likely in its expected

location. Since physical inspection and attestation of the chips would occur at centralized inspection locations, this may reduce the overhead and logistical costs relative to on-site physical inspections, as inspectors would not be required to physically travel to data centers containing export-controlled AI chips. That said, each inspection location would need to be outfitted with the personnel and equipment necessary to inspect and attest chips, which could be costly. Moreover, a mail-in system would very likely disrupt the data center more than physical inspections because the controlled chips chosen for inspection would be out of operation for several days to over a week during shipment and inspection.

Mail-in inspections are partially verifiable and repeatable, but they are neither accurate nor secure. Mail-in inspections are only partially verifiable: although they produce evidence of an AI chip's identity, they do not clearly verify its location. They are not accurate: even a calibrated shipping window must include buffer time to account for corporate and shipping logistics, so the window very likely would not be precise enough to determine the country of origin with high confidence. Mail-in inspections are also not secure because the origin of a shipment can be somewhat easily obfuscated. For example, chips could be shipped from export-controlled countries within relatively short time periods using express shipping.⁶⁶ Finally, mail-in inspections are only partially repeatable because requiring end users to submit each chip for inspection many times per year would impose an impractical burden.

Video Inspections

Chip owners and locations can be identified (though not verified) through video inspections, where law enforcement personnel or contracted investigators remotely inspect export-controlled AI chips at a data center using tamper-resistant video cameras. Cameras would likely be shipped by law enforcement personnel or contracted investigators to a data center prior to a video inspection.⁶⁷ To perform a proper video inspection, the inspector should instruct a data center technician to plug randomly identified chips into a workstation to attest that the hardware and firmware are authentic and to send the resulting proof to the inspector—similar to an on-site physical inspection.

On their own, proper video inspections are repeatable, but they are not verifiable, accurate, or secure. They are neither verifiable nor accurate because while they attest the authenticity of a chip, they do not directly provide a location estimate. To perform location verification (and thereby become verifiable and accurate), the video inspections must use a tamper-resistant camera that integrates one of the three geolocation mechanisms discussed in this paper: GPS-based, topology-based, or ping-based. The

security of the combined system would then depend on both the geolocation module's resistance to adversarial location obfuscation and the integrity of the attestation process. If the system integrated ping-based geolocation, its geolocation process would be secure; if it instead integrated topology-based or GPS-based geolocation, it would not be secure. Furthermore, even if the data center technician appears on video to attest the chip shown, the facility could manipulate its workstation to perform the attestation on the genuine chip located in a restricted country, allowing the diverted chip to pass inspection while a dummy is shown on camera. This threat is less pronounced during an on-site physical inspection, where the inspector can use a trusted workstation under their control.

Relative to physical inspections, video inspections could eliminate recurring inspector travel costs. However, these savings would be at least partly offset by other expenses. Upfront costs, for example, could include procuring and shipping tamper-resistant cameras. Additionally, if a trusted technician is required to install or configure the cameras, costs would include technician labor and travel. Individual chip inspections may take longer because facility staff would need to perform each step slowly and clearly on camera. Overall costs would depend heavily on whether cameras were permanently installed or shipped for individual inspections and whether facility personnel could set them up without outside assistance.

Inventory Management Tools

Chip owners and locations can be identified through periodic exports from inventory management tools, such as data center infrastructure management (DCIM) software.⁶⁸ This category of methods can help identify chip owners and locations, **particularly when owners are trusted**, but cannot directly verify a chip's location unless the tool is integrated with a geolocation mechanism. BIS could require chip owners to periodically export a report from an inventory management tool to help demonstrate that they still possess certain export-controlled AI chips. These tools are commonly used by data center operators to monitor hardware (e.g., chips and server racks), network connectivity, and power usage via a central server, workstation, or cloud platform. Some of these tools, such as NVIDIA Fleet Intelligence, can cryptographically attest an AI chip's identity and output chip-specific identifiers (e.g., a serial number). However, not all inventory management tools have these capabilities, so many controlled chips could not be audited securely, if at all, without the development of new software.

Moreover, inventory management tools can help identify AI chip owners, but each tool's reliability depends primarily on how easily a bad actor could falsify an export from the tool, which in turn depends on the security of the tool's technical implementation. For

example, NVIDIA Fleet Intelligence is fairly robust because (1) it verifies each chip's identity using the chip's hardware-fused key and (2) exports from the software are cryptographically signed.⁶⁹ Still, a Fleet Intelligence report could likely be falsified by a sufficiently well-resourced threat actor. Moreover, the system is vulnerable at scale: if a threat actor breaks the software once, they could likely repeatedly generate reports falsely characterizing the ownership of large volumes of chips.

Furthermore, inventory management tools cannot directly verify chip locations. These tools can only help identify the owner of an AI chip; any information provided about the chip's location would be unverified. For example, inventory management-based location verification would likely involve the chip owner sharing with BIS an inventory report exported from software like NVIDIA Fleet Intelligence that (1) cryptographically attests which AI chips, by serial number, the owner is currently operating at which data centers and (2) contains unverified information, inputted by the owner, describing the location of each data center. This type of verification requires trusting the owner to accurately report the location of each chip in their possession and therefore would be most appropriate for verifying the locations of AI chips owned by entities that BIS deems trustworthy (e.g., a large U.S.-headquartered cloud service provider). Inventory management tools' utility is limited for location verification of AI chips owned by less trusted actors.

If an inventory management tool is integrated with a geolocation mechanism (i.e., ping-based, topology-based, or GPS-based), then the combined system would be verifiable and accurate. Security of the combined system would depend on both the resistance of the system's geolocation module to adversarial location obfuscation and the integrity of the attestation process. If the system integrated ping-based geolocation, its geolocation process would be secure; if it instead integrated topology-based or GPS-based geolocation, its geolocation process would not be secure.

Geolocation Mechanisms

Geolocation mechanisms use data generated by AI chips, servers, and data center infrastructure to identify and verify chip locations. These mechanisms can be installed via software and/or firmware or embedded in the hardware of the chips themselves. There are three general approaches to geolocation, each of which can be implemented in numerous ways with varying cost, accuracy, and robustness to attack.⁷⁰ For example, each geolocation mechanism could be implemented as either a manual or automated system.

First, **ping-based (or delay-based) location verification (PLV)** measures the time it takes for a signal to travel from an AI chip to multiple trusted “landmark” servers and back. The signal’s round-trip time (RTT) can be used to trilaterate the AI chip’s location (see Appendix B for more details on PLV accuracy). This communication would preferably be cryptographically verified and would likely require dozens to hundreds of landmark servers dispersed globally to ensure sufficient geographic coverage. This mechanism would identify roughly where the chips are located and whether their locations have changed over time.

Second, **GPS-based location verification** involves AI chips communicating with GPS satellites to determine their locations and then reporting that information to separate trusted servers. However, GPS signals are unencrypted, easily spoofable, and require a clear connection to the AI chips, which are typically housed in data centers that can obstruct GPS signals. Data center shells obstructing GPS signals could be addressed by using a single GPS receiver per data center (affixed to the roof), but doing so would create more opportunities for a bad actor to defeat the system.

Third, **topology-based (or network-based) location verification** estimates a chip’s location by analyzing its connections to nearby networks. This involves using IP addresses, Wi-Fi connections to local area networks and access points, known network identifiers such as SSID and MAC addresses, or Bluetooth Low Energy (BLE) geobeacons to estimate the chip’s location based on the known locations of those networks.⁷¹

For example, NVIDIA Fleet Intelligence, the data center management software discussed in the “Inventory Management Tools” section of this Appendix, can also serve as the backbone of a topology-based geolocation mechanism. Instead of requesting data center operators share their NVIDIA GPU serial numbers listed in Fleet Intelligence, the system could use network-based telemetry data (e.g., IP and MAC addresses) to automatically monitor GPU locations and their associated serial numbers. This would allow a data center operator to verifiably demonstrate that they still possess the controlled chips and that the chips are located at a specified IP address.⁷² This mechanism would be relatively unintrusive because it is a downloadable software service that requires no AI chip firmware or hardware modifications and leverages pre-established infrastructure. However, IP addresses and other telemetry data can be readily spoofed, significantly limiting this mechanism’s reliability.

Of the three approaches, PLV is the most robust against adversarial attacks.⁷³ While GPS- and topology-based mechanisms may be easier to implement, they face significant limitations, particularly in their susceptibility to spoofing and reliance on

infrastructure that was not designed for location verification. The next section discusses a technical implementation of PLV in more detail.

Ping-Based Location Verification (PLV) System

PLV systems can be implemented in several ways, which influences their reliability, robustness against adversarial attacks, and associated costs for chip designers. Below, we describe the five steps of how a PLV system could function. The robustness and security of a PLV system primarily depend on how the authenticity of messages coming from the AI chips is verified (i.e., how a landmark server knows whether the chips it communicates with are what they claim to be). Therefore, we outline three possible authentication implementations (ranging from least to most secure) in step two; the remaining steps are generally consistent regardless of which authentication method is chosen.

Step 1: Establish Landmark Servers

Trusted entities—either BIS, chip designers, or third parties—must establish landmark servers globally to communicate with controlled AI chips. The exact implementation, including the number of landmark servers required, will depend on the scope of the policy and who bears the responsibility of verifying chip locations, both of which are discussed in greater detail in the “Policy Design Choices” section. In any case, we estimate that a PLV system would likely require between 39 and 500 landmark servers. See Appendix B for more details.

Step 2: Configure AI Chips to Communicate with Landmark Servers

When an AI chip receives a round-trip time (RTT) challenge from a landmark server, it must cryptographically verify the challenge and then generate and cryptographically sign a response that is then verified by the landmark server. For a chip’s signature to be trusted, the chip must first be securely “booted,” which is a standard hardware-based feature that helps ensure no threat actor has tampered with the device’s firmware or software.⁷⁴ Next, the signing should take place in an AI chip’s Trusted Execution Environment (TEE), which is an isolated area within a processor that securely stores data and runs code, because the security of the entire PLV system depends on how the AI chip’s response to the RTT challenge is signed. Without a secure signature, threat actors could more easily compromise the integrity and authenticity of the AI chip’s response to each RTT challenge.⁷⁵ However, enabling the TEE may also reduce the performance of AI workloads running on each chip.⁷⁶ The first data center AI chip to incorporate a TEE was NVIDIA’s H100, unveiled in 2022. This was part of NVIDIA’s efforts to improve data security through “confidential computing,” and TEEs remain a

standard feature in subsequent generations of NVIDIA GPUs.⁷⁷ Below, we outline three ways this signing process can be implemented on AI chips that have TEEs, ranging from lowest to highest security.

Least Secure: Software-Based Signing

In this configuration, the signing of the RTT challenge occurs via software running within the user-accessible part of the AI chip's TEE.⁷⁸ While the TEE itself is attested by the chip's hardware identity key (i.e., its integrity is certified by the most secure part of the AI chip's hardware), that key is not used to cryptographically sign the AI chip's response to an RTT challenge from a landmark server.⁷⁹ This configuration is easier to implement because it only requires downloading software, but it is also less secure because (1) the RTT response is not signed by the AI chip's hardware identity key, and (2) an adversary could potentially modify the software to sign false location data.

Moderately Secure: Firmware-Based Signing

Here, the chip's response to an RTT challenge is generated and signed in the firmware of the chip's TEE, which is exclusively controlled by the chip designer and is not user-accessible.⁸⁰ To enable this function on a chip that already has a TEE (e.g., NVIDIA's latest data center GPUs), the hardware designer would need to update the TEE firmware to authenticate the RTT challenge and generate a signed response to that challenge.⁸¹ Firmware-based signing is more secure than software-based signing because (1) this method would sign the response to an RTT challenge with the AI chip's hardware identity key, and (2) a bad actor cannot modify firmware as easily as software.* Crucially, this method would not require physical, on-die changes to the hardware and can be deployed remotely to existing chips that already possess TEEs, making it more feasible for AI chips already in circulation.⁸²

Most Secure: Hardware-Based Signing

The highest level of security involves AI chip designers adding dedicated hardware to the chip to sign RTT challenge responses.⁸³ This would require changes to the AI chip's design and fabrication, so it could only be implemented on chips not yet manufactured and would be costly to the designers. Moreover, hardware changes are irreversible once fabricated. If BIS requires hardware-based signing and chip designers implement it into their chip designs but BIS later mandates a different hardware implementation,

* An AI chip's "hardware identity key" refers to a cryptographic key—derived from the chip's unique, hardware-fused identity key—which can be compared against NVIDIA's own registry of valid keys.

existing chips cannot be updated. Additionally, any vulnerabilities in hardware similarly could not be remediated after implementation.

Chips Without TEEs

Most older AI chips and chips developed by companies other than NVIDIA lack TEEs, and because TEEs rely on structures physically fabricated in the hardware, they cannot be added to chips that have already been manufactured. The software- and firmware-based signing options can still be implemented without a TEE, but the system will generally be less secure because TEEs isolate a section of the processor and protect its memory from being accessed by other parts of the AI chip or host CPU—which orchestrates AI chips in a data center.

Chips that lack a hardware identity key (in addition to lacking a TEE) would also not be able to implement the firmware-based signing option as specified; the chip would not be able to sign responses to RTT challenges as securely (see Appendix E).

Step 3: Facilitate AI Chip and Landmark Server Communication via CPUs

AI chips are unable to directly communicate with the landmark servers on their own, so they must do so through a host CPU.⁸⁴ Specifically, the AI chip's TEE would run PLV code and communicate through a CPU in the same data center server. To improve security, the host CPU would also run code in an isolated environment, similar to the AI chip's TEE. When both the AI chip and host CPU are securely booted, a secure connection can be established between them.

In the default implementation proposed here, the host CPUs are used as a pass-through, allowing the AI chip to communicate its RTT response to the landmark servers.⁸⁵ To maximize security and limit the attack surface that adversaries could exploit, the landmark servers' CPUs must also be booted in secure confidential computing environments (in the same way the AI chips and host CPUs must be securely booted).⁸⁶ Because CPUs are already pre-configured with confidential computing environments that can be used by third parties, this requires no additional effort by CPU designers.

Step 4: Implement the PLV System

To implement the PLV system, a standardized protocol is required, which all configured devices must follow. This protocol is the procedure that integrates all aforementioned steps into an operable PLV system. Due to the nature of a flexible location verification system, each responsible entity may initially develop their own PLV protocol. However,

we recommend that the U.S. government encourage the development of a standard PLV protocol through an international standards-setting organization, which would allow relevant stakeholders (e.g., AI chip designers and landmark server operators) to participate and reach consensus. Formal standards development is the best way to ensure that the protocol is feasible, robust, and functional across heterogeneous chips and that the protocol does not introduce performance issues for AI chips that would be pinged by many landmark servers on a daily or weekly basis. NIST, which already works closely with standards-setting organizations, should lead this effort.⁸⁷

Once the protocol is established, the PLV system can be implemented. First, the AI chips and landmark server CPUs will establish a secure communications tunnel between them, using the data center host CPUs as a pass-through. Second, the landmark server will issue a cryptographic challenge to verify the identity of the AI chips. Once the AI chips authenticate the cryptographic challenge sent by the landmark server, they generate and sign a response that they send back to the landmark server. The landmark server then verifies the AI chips' signature and calculates the RTT of the communication. Each AI chip would follow this communication protocol with each landmark server on a regular basis.

Step 5: Reporting to the U.S. Government

The landmark servers would forward their RTT data to centralized “reporting servers,” which would compare RTT data to estimate the location of each AI chip. These reporting servers could be maintained by the U.S. government, AI chip designers, or a third-party contractor. Regardless of who operates them, reporting servers must be able to authenticate the data and ensure it was transmitted from the landmark servers, which could be done through standard cryptographic communications infrastructure.*

* Such as public key infrastructure (PKI).

Appendix B: Accuracy of Ping-Based Location Verification

PLV systems could verify the location of AI chips through two key methods: trilateration and delay similarity. Trilateration (i.e., geolocating an object by measuring time-of-flight to known landmarks) is the simpler of the two options, so we chose it as the baseline approach for a PLV system. If BIS judges that trilateration is not accurate enough, delay similarity or a hybrid approach could improve accuracy, albeit at the expense of greater complexity and potentially higher cost.

With 500 landmark servers, a PLV system will likely achieve region-level accuracy (worst case) to city-level accuracy (best case).⁸⁸ By positioning one landmark server in every major data center hub—defined as an area containing a large concentration of data centers within a 30-kilometer (19-mile) radius—outside the United States, China, and Russia, PLV could likely achieve country-level accuracy for the vast majority of PLV-tracked chips. After all major hubs have been covered, the remaining landmark servers should be distributed among smaller data center hubs to provide baseline region-level accuracy.⁸⁹ Since accuracy gains come from proximity, the operator only needs one landmark server per data center hub. Trilateration accuracy improves as landmark servers are placed closer to AI chips, but even in the best case, accuracy is very likely limited to about 30 km (19 mi).

A global PLV system would likely need between 39 and 500 landmark servers. We find that 500 landmark servers, distributed globally, can very likely guarantee PLV accuracy of 1,600 km (about 1,000 mi) or better, which we consider to be region-level accuracy. A low-effort PLV implementation would likely require at least 39 landmark servers; if distributed globally, these servers can very likely guarantee PLV accuracy of 3,400 km (about 2,120 mi) or better, which we consider to be continent-level accuracy.

Literature Review

We reviewed six sources on PLV, which included four academic papers published in a peer-reviewed journal or conference, one think-tank paper, and one web application of a corporate technical demo (Table B1). The two documents describing trilateration indicated that accuracy could range from 10 km to 1,000 km (6 mi to 621 mi), requiring as few as 38 landmark servers or as many as 500. One paper describing delay similarity used thousands of nodes to make granular (7–10 km; 4–6 mi) location estimates in two cities and one U.S. state. Three papers used multiple geolocation approaches; for example, supplementing trilateration with network topology information.

Table B1. Number of Landmark Servers and Accuracy Reported by Different Analyses of PLV

Source	Year	Method Used	Geographic Scope	Number of Landmark Servers	Accuracy (km)
IAPS	2024, 2025	Trilateration	Global	100–500	10–1,000
Ulyssean	2025	Trilateration	Global	38	< 500
SubvectorS_Geo	2023	Delay Similarity	Cities and U.S. State ^A	1,589–2,195	7–10
VerLoc	2021	Hybrid	Europe	1,000	60
Octant	2007	Hybrid	North America	51	35 ^B
TBG	2006	Hybrid	United States	68	67

Notes:

(A) SubvectorS_Geo was tested on data from two cities and one state: Shanghai, Tokyo, and New York State.

(B) Indicates that the accuracy value was originally reported in miles and then converted to kilometers.

Sources: IAPS,⁹⁰ Ulyssean,⁹¹ SubvectorS_Geo,⁹² VerLoc,⁹³ Octant,⁹⁴ and TBG.⁹⁵

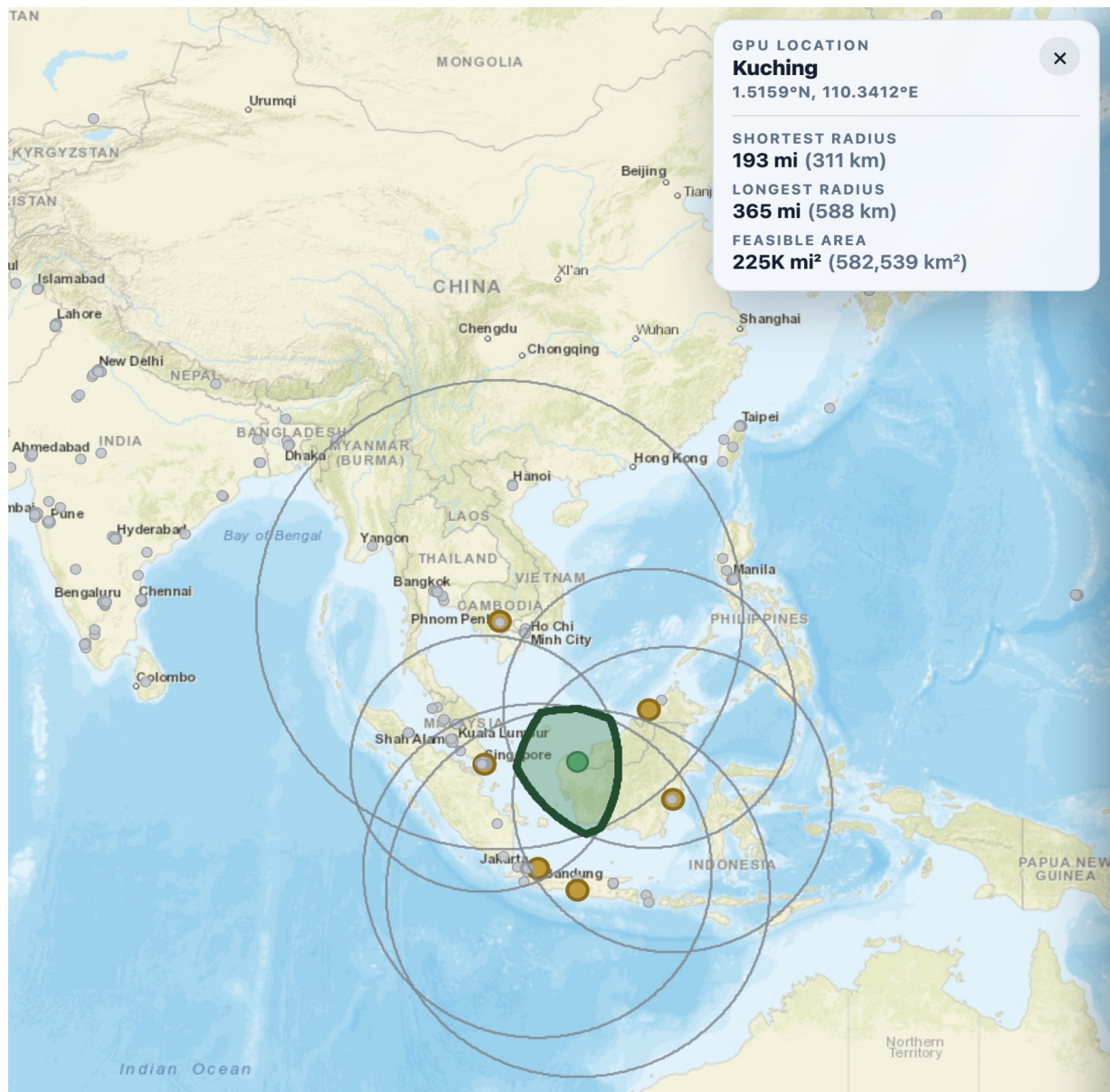
Trilateration is the simplest but least accurate approach. Each of the four papers that adopted a delay similarity or hybrid approach viewed trilateration as a baseline approach and sought to improve trilateration accuracy by leveraging additional information—either ground-truth delay information or trusted network features (i.e., locations of known network nodes).

The six sources reviewed did not provide reliable estimates for the accuracy and required number of landmark servers for a global PLV system. Four of the six documents focused on geolocation in a continent or smaller geography, which is too geographically constrained relative to a PLV system that would globally track the locations of AI chips. The two documents that provided global estimates did not clearly demonstrate how they arrived at the figures they provided. Therefore, we sought to model these figures, at least for the simplest option of trilateration.

Trilateration

Trilateration works as follows. The PLV system measures how long RTT requests and responses take to travel between the landmark server and the AI chip, using the speed of light to calculate the maximum distance the chip could be from the landmark server. This defines a disk around each landmark server within which the AI chip must be physically located (“landmark server disk”). The system would then identify the region where all landmark server disks overlap to further narrow down the area where the chip must be (“feasible region”). For a visual example of PLV geolocation, see Figure B1.

Figure B1. Illustration of PLV Geolocation



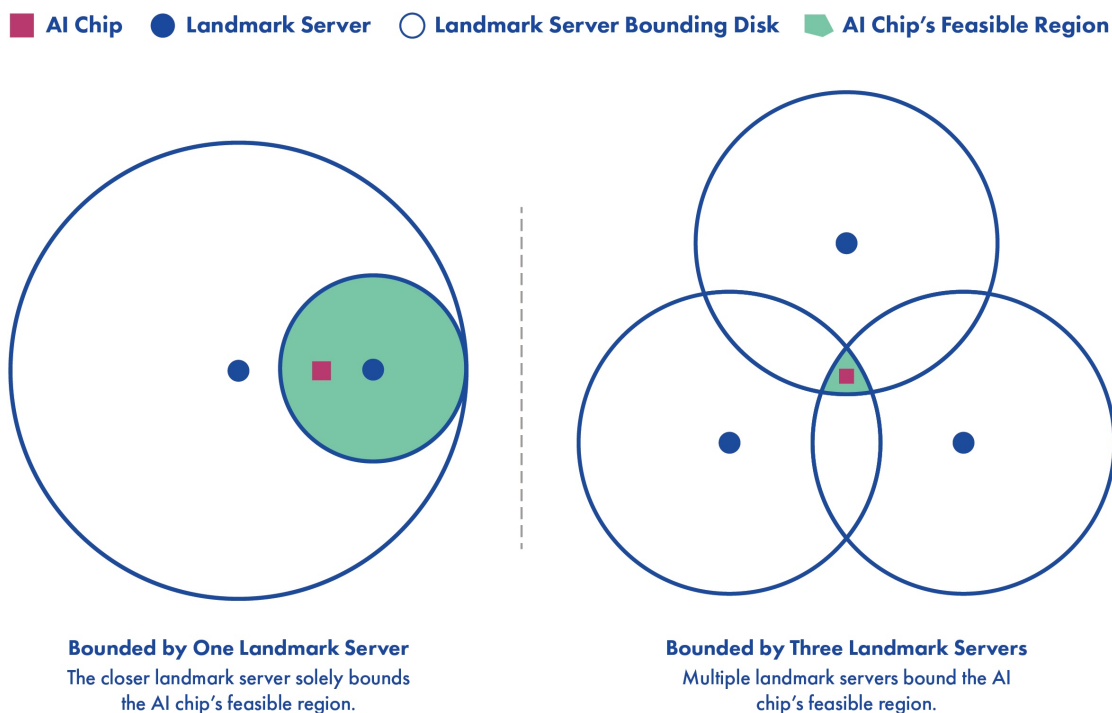
Note: In this example, the green dot is the true location of a tracked AI chip (i.e., Kuching, Malaysia). We assume landmark servers are installed in existing data centers (using datacente.rs world map), which are represented by grey and orange dots in the illustration. The orange dots represent landmark servers that are actively bounding the feasible region of the AI chip and grey dots represent landmark servers that are not actively bounding the feasible region. Therefore, we only show disks for the subset of landmark servers represented by orange dots. The green area around the tracked AI chip is the feasible area where the chip could be located based on one configuration of a PLV system.

Source: "Datacente.rs World Map," accessed on July 10, 2026, <https://map.datacente.rs/>.

Bounding PLV Accuracy

The accuracy of a PLV system is determined by the size of the feasible region around a tracked AI chip. Since this region is where all landmark server disks overlap, PLV accuracy is difficult to simulate because it depends on the exact positioning of all landmark servers relative to each tracked AI chip. However, we can easily model PLV accuracy for a single landmark server and AI chip. In cases where a landmark server is within 10 km (6 mi) of an AI chip, that landmark server's disk will likely solely define the PLV-feasible region around the AI chip.* In those cases, a model with one landmark server will accurately model PLV accuracy. In all other cases, such as when a landmark server is not located near the chip, a model with one landmark server will underestimate PLV accuracy (i.e., overestimate error) because the model does not calculate the overlap of multiple landmark servers, which would further improve accuracy (Figure B2).

Figure B2. Illustration of One Landmark Server Bounding AI Chip's Feasible Region (Left Panel) and Three Landmark Servers Bounding AI Chip's Feasible Region (Right Panel)



Source: CSET illustration.

* See subsection "When Does a Second Landmark Server Matter?"

To determine PLV accuracy, we need to calculate the maximum possible distance between the landmark server and the tracked AI chip. To do this, we multiply the RTT between the landmark server and the AI chip by the maximum possible velocity of the transmitted RTT request and response (Equation 1.1). Because the AI chip could be in any direction from the landmark server, the region where the chip could feasibly be located is represented as a disk centered on the landmark server with radius d_{max} . In Equation 1.1, we divide by two because RTT is two-way, but d_{max} is the one-way distance.

$$\text{Equation 1.1: } d_{max} = \frac{v_{max} \cdot t_{round\ trip}}{2}$$

Each additional millisecond of RTT latency increases a landmark server's disk radius by about 150 km (93 mi). Here, we set the maximum velocity equal to the speed of light (c), and we convert its units from m/s to km/ms (Equation 1.2).⁹⁶

$$\text{Equation 1.2: } d_{\Delta} = \frac{c}{2} \cdot \frac{0.001\ km/m}{1000\ ms/s} = 149.896\ km/ms$$

The RTT can be disaggregated into three components: flight time, network processing time (e.g., time required for DNS lookup and queuing time), and AI chip processing time (Equation 1.3). For the purposes of this model, we will assume that all of the AI chip processing time can be identified and communicated back to the landmark server in the RTT response; this allows us to subtract AI chip processing time from the total when calculating landmark server disk radius (Equation 1.4).

$$\text{Equation 1.3: } d_{max} = \frac{c \cdot (t_{network\ processing} + t_{flight} + t_{AI\ chip\ processing})}{2}$$

$$\text{Equation 1.4: } d_{max} = \frac{c \cdot (t_{network\ processing} + t_{flight})}{2}$$

Therefore, in the following analysis, we estimate RTT as the sum of network processing time and flight time (Equation 1.5).

$$\text{Equation 1.5: } t_{round\ trip} = t_{network\ processing} + t_{flight}$$

Using Equations 1.2 and 1.4, the final formula for calculating landmark server disk radius (i.e., PLV accuracy) is shown in Equation 1.6.

$$\text{Equation 1.6: } d_{max} = 149.896\ km/ms \cdot (t_{network\ processing} + t_{flight})$$

We calculate $t_{network\ processing}$ simply as a function of one-way network equipment latency, which we multiply by two for a round trip (Equation 1.7).

Equation 1.7: $t_{network\ processing} = 2 \cdot t_{equipment\ latency}$

We calculate t_{flight} using estimates of four variables per Equation 1.8: (a) the straight-line distance between the landmark server and AI chip ($d_{straight\ line}$), (b) the percentage of additional distance that data travels beyond the straight-line path due to how the fiber was laid (i.e., fiber path adjustment), (c) the combined distance from the landmark server and AI chip to each data center's respective local fiber access point ($d_{local\ loop}$), and (d) the average speed of communication through fiber ($v_{communication}$).

$$\text{Equation 1.8: } t_{flight} = \frac{2 \cdot d}{v} = 2 \cdot \frac{(d_{straight\ line} \cdot (1 + [fiber\ path\ adjustment])) + d_{local\ loop}}{v_{communication}}$$

To demonstrate how various factors impact RTT, we constructed four cases, then calculated RTT and landmark server radius for each (Table B2). One pair of cases assumes a true distance of 1 km (0.6 mi) between the landmark server and AI chip, while the other pair assumes a true distance of 5,000 km (3,107 mi). Each pair has a worst case and a better case (or best case) for RTT. For all cases, we assume the speed of communication is equal to the speed of light through fiber (i.e., approximately 200 km/ms).⁹⁷ The worst-cases assume 2 ms of one-way equipment latency, an additional 20% fiber distance due to deviation from the straight-line path, and a 15 km or 160 km (9 mi or 99 mi) local loop.⁹⁸ The better/best cases assume 0.1 ms of one-way equipment latency and that the fiber follows a straight-line path from the landmark server data center to the AI chip data center (i.e., 0% fiber path adjustment and 0 km local loop length).⁹⁹ For the 5,000 km cases, we refer to “better case” because, as discussed earlier, this model does not calculate the overlap of multiple landmark servers, which could substantially improve PLV accuracy beyond the better-case accuracy. For the 1 km cases, we refer to “best case,” because realistically, the 1 km better case could not be improved by adding more landmark servers, so it represents the best-case scenario.¹⁰⁰

For each case, we calculate PLV accuracy (i.e., landmark server radius) using Equation 1.6 (Table B2). When the distance between the landmark server and the AI chip is 5,000 km (3,107 mi), the PLV accuracy ranges from the better case of 7,525 km (4,676 mi) to the worst case of 9,833 km (6,110 mi). When the distance between the landmark server and the AI chip is 1 km, the PLV accuracy ranges from the best case of 31 km (19 mi) to the worst case of 624 km (388 mi). **Decreasing the distance between the landmark server and the AI chip improves PLV accuracy.** Therefore, a PLV operator could place landmark servers closer to the true location of AI chips to improve accuracy.

Table B2. Round-Trip Time and PLV Accuracy (i.e., Landmark Server Disk Radius) for Four Example Cases

Distance from Landmark Server to AI Chip ^A	1 km		5,000 km	
	Best	Worst	Better	Worst
Equipment Latency (ms) ^B	0.1	2.0	0.1	2.0
Estimated RTT: Network Processing (ms)	0.2	4.0	0.2	4.0
Fiber Path Adjustment	0%	20%	0%	20%
Local Loop Length (km) ^C	0	15	0	160
Speed of Communication (km/ms) ^D	200	200	200	200
Estimated RTT: Flight (ms)	0.01	0.16	50.00	61.60
Estimated RTT: Total (ms)	0.21	4.16	50.20	65.60
PLV Accuracy (i.e., Landmark Server Disk Radius; km) ^E	31	624	7,525	9,833

Notes:

(A) The distance from landmark server to AI chip is represented by $d_{straight\ line}$ in Equation 1.8.

(B) The equipment latency is represented by $t_{equipment\ latency}$ in Equation 1.7.

(C) The local loop length is represented by $d_{local\ loop}$ in Equation 1.8.

(D) The speed of communication (in km/ms) is represented by $v_{communication}$ in Equation 1.8.

(E) PLV accuracy is calculated using Equation 1.6.

Sources: CSET calculations (available in this Google Sheets workbook:

<https://docs.google.com/spreadsheets/d/1VQth7sMuOVd9tI7eAAYHFuSKIWI0-yll4O7fMW3qE9k/edit?gid=0#gid=0>); WAN Latency Calculator, <https://wintelguy.com/wanlat.html>.

For a given AI chip and landmark server, the amount of error in a landmark server's disk radius depends on how much higher the measured RTT is than the minimum theoretical RTT. Every additional millisecond of measured RTT adds about 150 km (93 mi) of error to the landmark server's disk radius (Equation 1.2), increasing the size of the AI chip's feasible region. When added RTT comes from network processing, all of that time becomes additional error. When that additional RTT comes from flight time, the error is proportional to the difference between actual flight speed and the maximum flight speed (i.e., v_{max} , which we have defined as the speed of light in a vacuum).

For each of the four cases presented in Table B2, we calculate minimum theoretical RTT using Equation 1.9 (Table B3). The minimum theoretical RTT is two times the straight-line distance between the landmark server and AI chip ($d_{straight\ line}$) divided by the speed of light in a vacuum (c).

$$\text{Equation 1.9: } t_{min\ round\ trip} = \frac{2 \cdot d_{straight\ line}}{c}$$

Table B3. Landmark Server Radius Error for Four Example Cases

Distance from Landmark Server to AI Chip ^A		1 km		5,000 km	
		Best	Worst	Better	Worst
Minimum Theoretical RTT (ms) ^B		0.007	0.007	33.356	33.356
Error	Distance (km)	30	623	2,525	4,833
	Percentage ^C	96.8%	99.8%	33.6%	49.2%
Error Share of Total by Type	Network Processing	98.4%	96.3%	1.2%	12.4%
	Flight	1.6%	3.7%	98.8%	87.6%

Notes:

(A) The distance from landmark server to AI chip is represented by $d_{straight\ line}$ in Equation 1.8.

(B) Minimum theoretical RTT is calculated using Equation 1.9.

(C) RTT error percentage is calculated as $1 - \frac{\text{Distance from Landmark Server to AI Chip}}{\text{Landmark Server Radius}}$.

Source: CSET calculations (available in this Google Sheets workbook:

<https://docs.google.com/spreadsheets/d/1VQth7sMuOVd9tI7eAAYHfFuSKIWI0-yII4O7fMW3qE9k/edit?gid=0#gid=0>).

The closer the landmark server is to the AI chip, the more the error comes from network processing rather than time of flight. When the distance between the landmark server and the AI chip is 5,000 km (3,107 mi), time of flight is responsible for 88% to 99% of the landmark server radius error. When the distance between the landmark server and the AI chip is 1 km (0.6 mi), network processing is responsible for nearly all (96–98%) of the radius error (Table B3).

Realistically, the best-case PLV accuracy is 30 km (19 mi), which we consider to be city-level accuracy. To calculate this, we assume that the landmark server and AI chip

are in directly adjacent data centers, meaning the flight time between them is effectively zero. We also assume that the communication incurs a best-case, one-way equipment latency of 0.1 ms, which creates a landmark server disk with radius of 30 km (19 mi).¹⁰¹ Adding additional landmark servers around the AI chip cannot improve this best-case accuracy because each additional kilometer between a landmark server and AI chip increases disk radius by 1.5 km (0.9 mi).¹⁰² Therefore, the overlap with any additional landmark server disks will still equal the area of the directly adjacent landmark server's disk. This best-case accuracy could only be improved by reducing the one-way equipment latency below 0.1 ms.

Bad actors could easily introduce latency into the system to increase the size of an AI chip's feasible region, but this does not undermine the value of the system. If an AI chip's feasible region is larger than expected for the chip's expected location and overlaps with China, for example, it should indicate to the responsible entity that diversion may be occurring, and further investigation may be necessary. The benefit of trilateration is that, so long as the landmark servers themselves are not compromised, BIS can be sure that the tracked chip is located within the feasible area. However, the downside of trilateration is that the feasible area it outputs can be very large. Another approach, delay similarity (discussed later), can likely achieve greater precision.

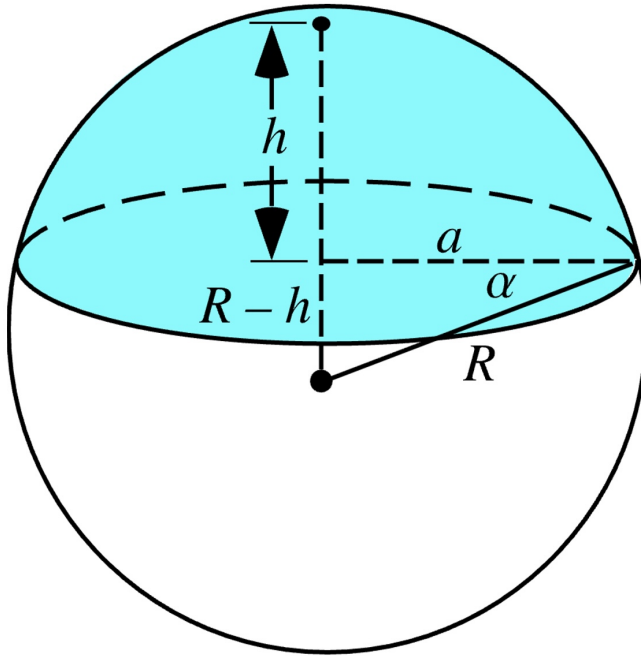
Number of Landmark Servers

Next, given a specified number of landmark servers, we estimate the worst-case distance from any AI chip on Earth to its nearest landmark server as well as worst-case and better-case PLV accuracy (Table B4). Naively, to guarantee a baseline PLV accuracy at any non-restricted location around the world, we would equally distribute landmark servers so that no point in a covered location is more than a specified distance from a landmark server.¹⁰³ We can treat this as a sphere-covering problem to calculate the worst-case distance from any AI chip on Earth to its nearest landmark server, given a specified number of landmark servers; each landmark server is a cap and the Earth is a sphere.¹⁰⁴

We start with the formula for a cap on a sphere, where R is the radius of the sphere, h is the height of the cap, and A_{cap} is the area of the cap (Equation 2.1).

Equation 2.1: $A_{cap} = 2\pi Rh$

Figure B3. Spherical Cap



Note: In Equations 2.2–2.8.3, α denotes the complement of the angle labeled α in this figure—that is, 90 degrees ($\pi/2$ radians) minus the angle shown.

Source: “Spherical Cap,” Wolfram MathWorld, accessed July 8, 2026, <https://mathworld.wolfram.com/SphericalCap.html> [<https://perma.cc/QAQ9-D49T>].

We rewrite the height of the cap (h) in terms of the sphere’s radius as $R(1 - \cos(\alpha))$, where α is the angle (in radians) from vertical to the edge of the cap (Equation 2.2).

$$\text{Equation 2.2: } A_{cap} = 2\pi R^2(1 - \cos(\alpha))$$

We rewrite the cap angle α in terms of the radius of the cap r (i.e., the maximum distance that we will allow between an AI chip and a landmark server). The length of an arc r is given as $r = R\alpha$, where α is the cap angle in radians. Therefore, $\alpha = \frac{r}{R}$, which we can substitute into Equation 2.2 to create Equation 2.3.

$$\text{Equation 2.3: } A_{cap} = 2\pi R^2(1 - \cos(\frac{r}{R}))$$

Next, the number of caps of radius r needed to cover the surface of the sphere is approximately equal to the area of the sphere A_{Earth} divided by the cap area A_{cap} and multiplied by a covering factor θ (Equation 2.4).

$$\text{Equation 2.4: } N_{cap} \approx \frac{\theta \cdot A_{Earth}}{A_{cap}}$$

The number of caps (i.e., landmark servers) required to cover the sphere is minimized when caps are laid out on the sphere at the vertices of a lattice of equilateral triangles (i.e., a regular triangular lattice or regular hexagonal lattice). This yields the smallest known covering density factor $\theta = \frac{2\pi}{3\sqrt{3}} \approx 1.209$ (Equation 2.5).¹⁰⁵ We multiply θ by two as a margin of error to account for two ways that the positioning of landmark servers for PLV deviates from a pure sphere-covering problem: (a) while landmark servers could be laid in a triangular lattice, they certainly could not be laid in a *regular* triangular lattice without building new data centers, which we do not recommend, and (b) the covered land masses will be shaped irregularly, not as a perfect sphere, because we will exclude the oceans and certain countries.* Multiplying θ by two assumes that the real positioning of landmark servers (in an irregular triangular lattice on irregularly-shaped land masses) would require no more than double the number of landmarks to cover the surface of the sphere, relative to a scenario where landmarks are laid in a regular triangular lattice on a perfect sphere.¹⁰⁶

$$\text{Equation 2.5: } N_{cap} \approx \frac{\frac{2 \cdot 2\pi}{3\sqrt{3}} \cdot A_{Earth}}{A_{cap}}$$

We substitute A_{cap} using Equation 2.3 (Equation 2.6.1).

$$\text{Equation 2.6.1: } N_{cap} \approx \frac{\frac{2 \cdot 2\pi}{3\sqrt{3}} \cdot A_{Earth}}{2\pi R^2 (1 - \cos(\frac{r}{R}))}$$

We rearrange the equation so r is the output and N_{cap} is an input (Equations 2.6.2 and 2.6.3).

$$\text{Equation 2.6.2: } 1 - \cos(\frac{r}{R}) \approx \frac{\frac{2 \cdot 2\pi}{3\sqrt{3}} \cdot A_{Earth}}{2\pi R^2 N_{cap}}$$

$$\text{Equation 2.6.3: } r \approx \arccos\left(1 - \frac{\frac{2 \cdot 2\pi}{3\sqrt{3}} \cdot A_{Earth}}{2\pi R^2 N_{cap}}\right) \cdot R$$

Now, we can substitute R and A_{Earth} for their respective values. We need to make one modification: calculating A_{Earth} based on the radius of the Earth would give us the entire surface area of the Earth, but we want to exclude the oceans, Antarctica, China, Russia, and the United States.¹⁰⁷ Therefore, we calculate A_{Earth} per Equation 2.7, where $A_{Earth\ land} = 148,940,000 \text{ km}^2$, $A_{Antarctica} = 14,200,000 \text{ km}^2$, $A_{Russia} =$

* See Equation 2.7.

17,075,400 km², $A_{China} = 9,572,900$ km², and $A_{United\ States} = 9,148,663$ km².¹⁰⁸ For R , we use Earth's shortest radius, which is at the poles: $R = 6,356.752$ km.¹⁰⁹

Equation 2.7:

$$A_{Earth} = A_{Earth\ land} - A_{Antarctica} - A_{Russia} - A_{China} - A_{United\ States} = 98,943,037 \text{ km}^2$$

$$\text{Equation 2.8.1: } r \approx \arccos\left(1 - \frac{\frac{2 \cdot 2\pi}{3\sqrt{3}} \cdot A_{Earth}}{2\pi R^2 N_{cap}}\right) \cdot R = \arccos\left(1 - \frac{\frac{2 \cdot 2\pi}{3\sqrt{3}} \cdot 98,943,037}{2\pi (6,356.752)^2 N_{cap}}\right) \cdot (6,356.752)$$

$$\text{Equation 2.8.2: } r \approx \arccos\left(1 - \frac{0.942459754}{N_{cap}}\right) \cdot (6,356.752)$$

Using Equation 2.8.2, given a specified number of landmark servers, we estimate the worst-case distance from any AI chip in a tracked location to its nearest landmark server as well as worst-case and better-case PLV accuracy (Table B4).

In addition, Equation 2.8.3 gives the approximate maximum spacing between neighboring landmark servers. In our model, we assume landmark servers are arranged in a regular triangular lattice, such that three neighboring servers form an equilateral triangle. When the radius of each landmark server cap (r)—the maximum distance from any point to its nearest landmark server—is a small percentage of Earth's radius, r is approximately equal to the circumradius of an equilateral triangle of neighboring landmark servers. Therefore, the approximate maximum distance (d) between neighboring landmark servers is represented by Equation 2.8.3.¹¹⁰ The actual distances between neighboring landmark servers may be somewhat smaller than the maximum because our density factor (θ) includes a conservative buffer to account for irregularly placed landmark servers and irregularly shaped land masses.

$$\text{Equation 2.8.3: } d \approx \sqrt{3} \cdot r$$

Table B4. PLV Accuracy and Maximum Distance between AI Chip and Landmark Server for Fixed Number of Landmark Servers

Number of Landmark Servers	Maximum Distance between AI Chip and Landmark Server (km) ^A	PLV Accuracy (km) ^B	
		Better	Worst
250	553	859	1,834
500	391	616	1,543
750	319	508	1,413
1,000	277	445	1,338

Notes: As in Table B2, the worst-case assumes 2 ms of one-way equipment latency, an additional 20% fiber distance due to deviation from the straight-line path, and a 160 km (99 mi) local loop. The better case assumes 0.1 ms of one-way equipment latency and that the fiber follows a straight-line path from the landmark server data center to the AI chip data center (i.e., 0% fiber path adjustment and 0 km local loop length). We refer to “better case” because, as discussed earlier, this model does not calculate the overlap of multiple landmark servers, which could substantially improve PLV accuracy beyond the better-case accuracy.

(A) Maximum distance between AI chip and landmark server is calculated using Equation 2.8.2.

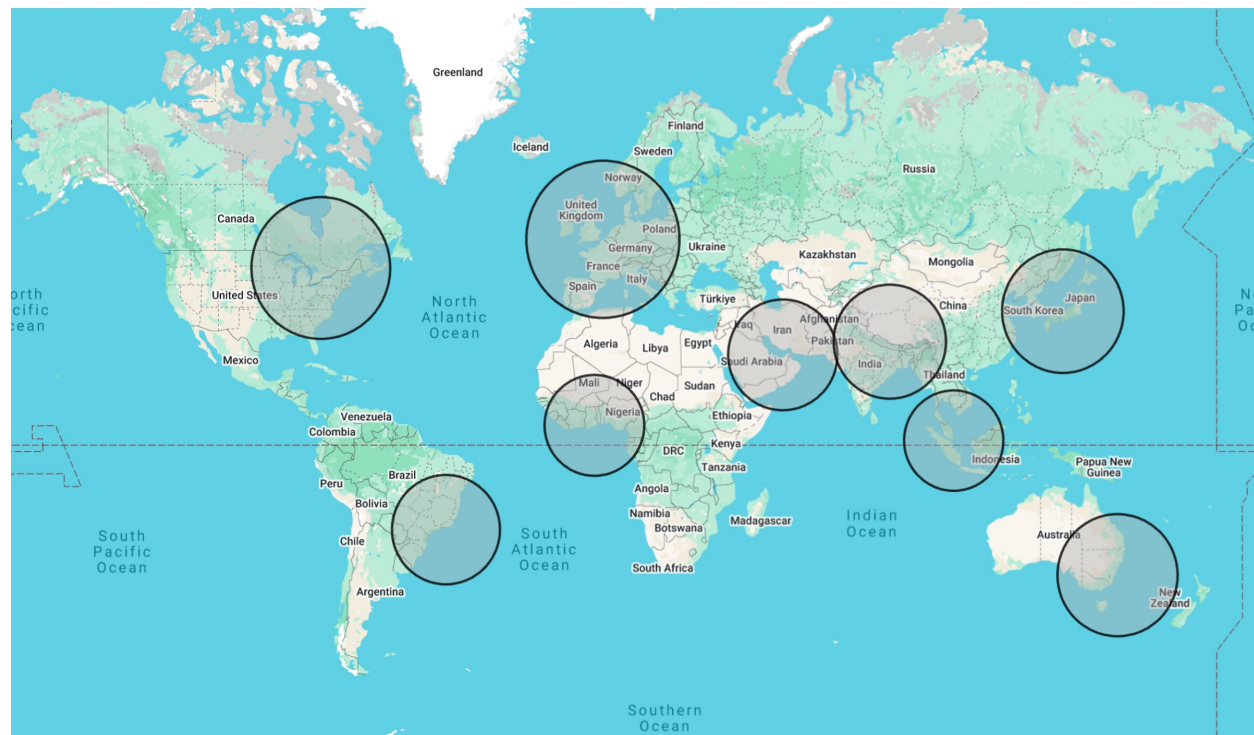
(B) PLV accuracy is calculated using Equation 1.6.

Source: CSET calculations (available in this Google Sheets workbook:

<https://docs.google.com/spreadsheets/d/1VQth7sMuOVd9tI7eAAYHFuSKIWI0-yII4O7fMW3qE9k/edit?gid=1841821404#gid=1841821404>).

Based on this analysis, 500 landmark servers is a reasonable upper bound for a global PLV system. We find that 500 landmark servers, distributed globally, can very likely guarantee PLV-feasible regions with a radius of 1,600 km (about 1,000 mi) or less, which we consider to be region-level accuracy. For example, imagine that you have an AI chip 391 km (243 mi) from a landmark server in Paris. Worst case, the PLV system should show that the feasible region of that AI chip is within a disk centered in Paris with a radius 1,600 km; the disk would enclose most of Europe (Figure B4). As discussed earlier, however, this model does not calculate the overlap of multiple landmark servers, which could substantially improve PLV accuracy beyond even the better-case accuracy. Furthermore, strategically positioning each landmark server within a data center hub would improve accuracy at the locations where most tracked AI chips are installed.¹¹¹

Figure B4. Example Landmark Server Disks with Radius of 1,600 km (about 1,000 mi)



Note: Imagine that there are nine AI chips, each 391 km (243 mi) from a respective landmark server in one of the following nine cities: (a) Toronto, Canada, (b) Rio de Janeiro, Brazil, (c) Paris, France, (d) Accra, Ghana, (e) Abu Dhabi, United Arab Emirates, (f) Kathmandu, Nepal, (g) Singapore, (h) Osaka, Japan, and (i) Sydney, Australia. Worst case, the PLV system should show that the feasible region of the AI chip is within a disk centered at each landmark server with a radius 1,600 km (about 1,000 mi). Lastly, note that the 2D disks vary in size due to their distance from the equator (i.e., as a result of the effects of the Mercator projection).

Source: Map Developers “Draw a Circle” Tool (saved map available [here](#)).

Thirty-nine landmark servers is a reasonable lower bound for a global PLV system.

The simplest viable implementation of trilateration would likely use a single U.S. cloud service provider (CSP), instantiating a single landmark server in each of the CSP’s geographic regions. Among the three largest CSPs by market share (i.e., AWS, Microsoft Azure, and Google Cloud), AWS has the lowest number of geographic regions at 39 (Table B5).

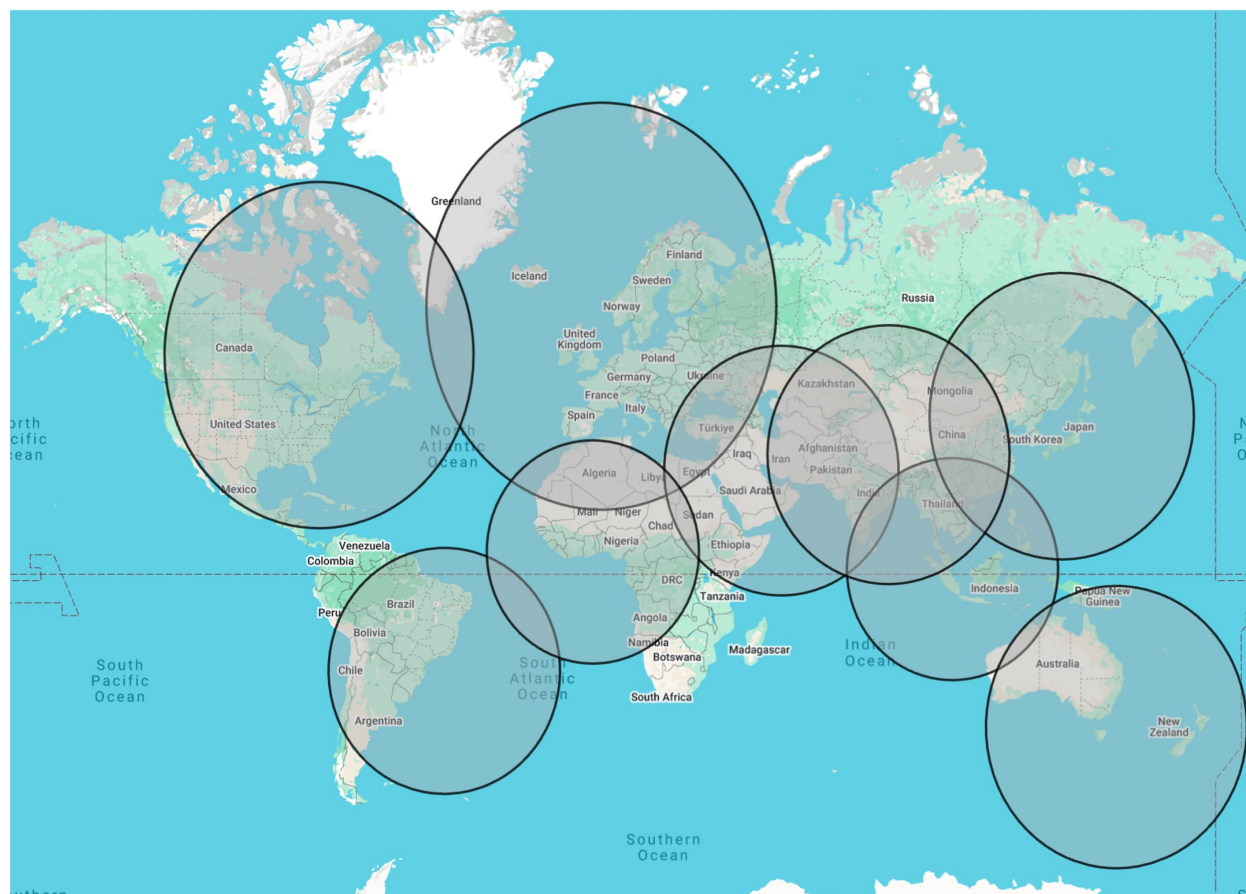
Table B5. Number of Geographic Regions Maintained by the Three Largest Cloud Service Providers

Cloud Service Provider (CSP)	Market Share (Q1 2026)	Number of Geographic Regions
Amazon Web Services (AWS)	28%	39
Microsoft Azure	21%	57
Google Cloud	14%	43

Sources: Statista,¹¹² Amazon Web Services,¹¹³ Microsoft,¹¹⁴ Google Cloud.¹¹⁵

We find that 39 landmark servers, distributed globally, can very likely guarantee PLV-feasible regions with a radius of 3,400 km (about 2,120 mi) or less, which we consider to be continent-level accuracy. Thirty-nine landmark servers can be distributed so that no place on Earth (where the U.S. government would want to verify AI chip locations) is more than 1,401 km (871 mi) from a landmark server. Imagine an AI chip is 1,401 km from a landmark server in Paris. Worst case, the PLV system should show that the feasible region of that AI chip is within a disk centered in Paris with a radius 3,400 km; the disk would enclose all of Europe, North Africa, and parts of the Middle East (Figure B5). As discussed earlier, however, this model does not calculate the overlap of multiple landmark servers, which could substantially improve PLV accuracy beyond even the better-case accuracy.

Figure B5. Example Landmark Server Disks with Radius of 3,400 km (about 2,120 mi)



Note: Imagine that there are nine AI chips, each 1,401 km from a respective landmark server in one of the following nine cities: (a) Toronto, Canada, (b) Rio de Janeiro, Brazil, (c) Paris, France, (d) Accra, Ghana, (e) Abu Dhabi, United Arab Emirates, (f) Kathmandu, Nepal, (g) Singapore, (h) Osaka, Japan, and (i) Sydney, Australia. Worst case, the PLV system should show that the feasible region of the AI chip is within a disk centered at each landmark server with a radius 3,400 km (about 2,120 mi). Lastly, note that the 2D disks vary in size due to their distance from the equator (i.e., as a result of the effects of the Mercator projection).

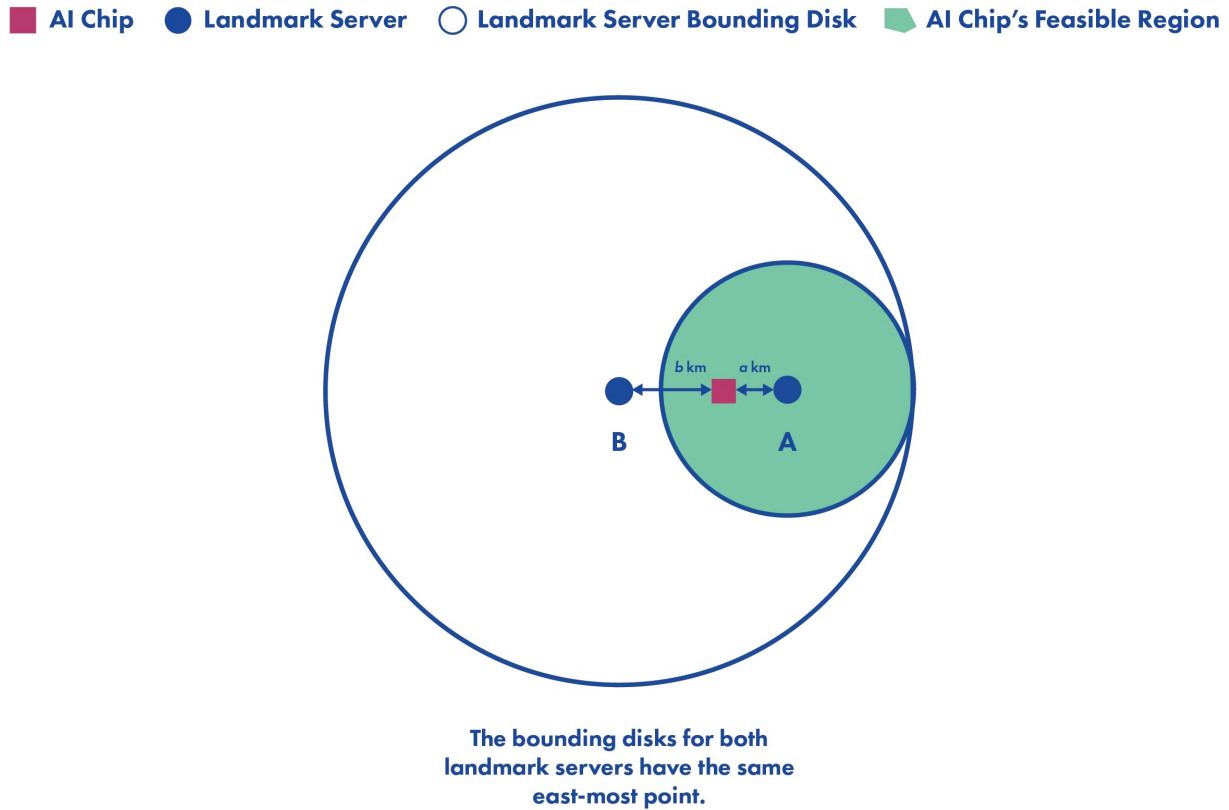
Source: Map Developers “Draw a Circle” Tool (saved map available [here](#)).

When Does a Second Landmark Server Matter?

As explained in the “Bounding PLV Accuracy” section of this Appendix, when a landmark server is positioned within 10 km (6 mi) of an AI chip, that server’s disk will likely solely define the PLV-feasible region around the AI chip. This is because PLV error increases as the distance between a landmark server and the AI chip increases, meaning that landmark servers further away from an AI chip are less useful than ones close by (Table B3).

Assume landmark server A is positioned near an AI chip, specifically a km east of the chip, where a is 20 km (12 mi) or less. Landmark server B is positioned b km west of the chip. We want to calculate how close landmark server B needs to be to the AI chip for its disk to start bounding the feasible region of the AI chip (Figure B6).¹¹⁶

Figure B6. Landmark Server B 's Disk Starts to Bound the Feasible Region of the AI Chip when the East-Most Points of the Two Landmark Server Disks Intersect



Source: CSET illustration.

Landmark server B 's disk starts to bound the feasible region when the east-most points of the two landmark server disks are at the same location. For simplicity, we assume the two landmark servers have the same $t_{equipment\ latency}$ and therefore the same $t_{network\ processing}$ (Equation 1.7). We also assume that each landmark server has a straight-line path to the AI chip data center (i.e., 0% fiber path adjustment and 0 km local loop length). Therefore, the equation for t_{flight} (Equation 1.8) simplifies to Equation 3.1.

$$\text{Equation 3.1: } t_{flight} = \frac{2 \cdot d_{straight\ line}}{v_{communication}}$$

Using Equations 1.6 and 3.1, the location of the east-most point of landmark server A is given by Equation 3.2, and the location of the east-most point of landmark server B is given by Equation 3.3.

$$\text{Equation 3.2: } p_A = 149.896 \text{ km/ms} \cdot \left(t_{\text{network processing}} + \frac{2 \cdot d_{\text{landmark A to AI chip}}}{v_{\text{communication}}} \right) + d_{\text{landmark A to AI chip}}$$

$$\text{Equation 3.3: } p_B = 149.896 \text{ km/ms} \cdot \left(t_{\text{network processing}} + \frac{2 \cdot d_{\text{landmark B to AI chip}}}{v_{\text{communication}}} \right) - d_{\text{landmark B to AI chip}}$$

Next, we set Equations 3.2 and 3.3 equal to each other to identify when $p_A = p_B$ (Equation 3.4.1). We set $d_{\text{landmark A to AI chip}}$ equal to a and $d_{\text{landmark B to AI chip}}$ equal to b (Equation 3.4.2). Then we solve for b (Equations 3.4.3, 3.4.4, and 3.4.5).

Equation 3.4.1:

$$149.896 \cdot \left(t_{\text{network processing}} + \frac{2 \cdot d_{\text{landmark A to AI chip}}}{v_{\text{communication}}} \right) + d_{\text{landmark A to AI chip}} = 149.896 \cdot \left(t_{\text{network processing}} + \frac{2 \cdot d_{\text{landmark B to AI chip}}}{v_{\text{communication}}} \right) - d_{\text{landmark B to AI chip}}$$

$$\text{Equation 3.4.2: } 149.896 \cdot \left(t_{\text{network processing}} + \frac{2a}{v_{\text{communication}}} \right) + a = 149.896 \cdot \left(t_{\text{network processing}} + \frac{2b}{v_{\text{communication}}} \right) - b$$

$$\text{Equation 3.4.3: } \frac{299.792a}{v_{\text{communication}}} + a = \frac{299.792b}{v_{\text{communication}}} - b$$

$$\text{Equation 3.4.4: } \frac{a \cdot (299.792 + v_{\text{communication}})}{v_{\text{communication}}} = \frac{b \cdot (299.792 - v_{\text{communication}})}{v_{\text{communication}}}$$

$$\text{Equation 3.4.5: } b = \frac{a \cdot (299.792 + v_{\text{communication}})}{299.792 - v_{\text{communication}}}$$

We set $v_{\text{communication}}$ equal to the speed of light in fiber. Then, for a specified distance (1 to 20 km) between the closer landmark server and the AI chip (a), we calculate the maximum distance between the farther landmark server and the AI chip (b) required for the farther landmark to bound the AI chip's feasible region (Table B6).

Table B6. In Two-Landmark Server Model, Maximum Bounding Distance of Farther Landmark Server by Distance to Nearer Landmark Server

Distance from Closer Landmark Server to AI Chip (a , km)	Maximum Distance for Farther Landmark Server to Bound AI Chip's Feasible Region (b , km)
1	5.0
2	10.0
5	25.0
10	50.1
20	100.2

Note: We use Equation 3.4.5.

Source: CSET calculations (available in this Google Sheets workbook:

<https://docs.google.com/spreadsheets/d/1VQth7sMuOVd9tI7eAAYHFuSKIWI0-yll4O7fMW3qE9k/edit?gid=1042034955#gid=1042034955>).

Under these conditions, the farther landmark server only helps narrow the chip's possible location if the farther server is no more than about 5x farther away than the closer landmark server. This means, for example, that if a landmark server is positioned 10 km (6 mi) from an AI chip, adding a second landmark server will not help bound the chip's feasible region unless it is positioned within 50 km (31 mi) from the chip (in the opposite direction). Table B4 shows that 500 landmark servers could be distributed such that a tracked AI chip in any covered location is within 391 km (243 mi) of a landmark server. According to the model presented in subsection "Number of Landmark Servers," landmark servers would be distributed as far as 677 km (421 mi) from one another (Equation 2.8.3). Since landmark servers would likely need to be hundreds of kilometers apart to guarantee minimum accuracy, a landmark operator likely would not place multiple landmark servers within 100 km (62 mi) of each other. Therefore, when a landmark server is within 10 km (6 mi) of an AI chip, that server's disk will likely solely define the PLV-feasible region around the AI chip. This means a second PLV landmark server does not increase geographic resolution within a metropolitan area. **In most cases, a PLV landmark server operator should place no more than one landmark server per metropolitan area.**

Positioning Landmark Servers

PLV could likely achieve country-level accuracy for the vast majority of PLV-tracked chips by positioning one landmark server in every major data center hub outside the

United States, China, and Russia. We define a data center hub as an area containing multiple data centers within a 30-kilometer (19-mile) radius, and a major data center hub as one containing a particularly large concentration of data centers within that radius.¹¹⁷ Once all major hubs have been covered, the remaining landmark servers should be distributed among smaller data center hubs to provide baseline regional-level accuracy elsewhere.

We define country-level accuracy as an error circle small enough to indicate that an AI chip is located either in its expected country or in a neighboring country, while ruling out countries farther away. An error-circle radius of 187 km (116 mi) or less—equivalent to the size of the 100th-largest country—is a reasonable threshold for country-level accuracy.¹¹⁸

This hub-based positioning strategy would work because the vast majority of PLV-tracked AI chips are likely to be located in major data center hubs. PLV tracking would apply only to data center AI chips, so even relatively small chip deployments would generally be housed in data centers, which are themselves often concentrated near one another.¹¹⁹ Placing a landmark server in each major hub therefore would reduce the distance between tracked AI chips in that hub and their nearest landmark server. If landmark servers are positioned in this way, PLV could achieve country-level accuracy—defined as an error-circle radius of 187 km (116 mi) or less—provided that one-way equipment latency does not exceed 0.4 milliseconds and the fiber route between the AI chip's data center and the landmark server adds less than 15 km (9 mi) beyond the straight-line distance.¹²⁰ Both parameters are likely achievable for AI chips located at such short distances from the landmark servers (i.e., 30 km or less).

If BIS knew the locations of all data centers that contain tracked AI chips, it could require the landmark server operator to place servers only in those hubs to maximize accuracy. The locations of large clusters of controlled AI chips are available in public and commercial datasets, but we are not aware of any source that currently does the same for smaller deployments.¹²¹ If BIS imposed a global licensing requirement alongside a location verification requirement, the Bureau would learn the locations of new, small deployments of AI chips over time, but it does not currently have that information in-house.

Finally, some locations of AI chips are beyond the reach of this strategy. BIS could address poor PLV accuracy at new, remote data centers by requiring on-site landmark servers at those facilities. Because data centers currently exist in a limited set of places, PLV cannot always achieve a worst-case accuracy of 1,600 km (about 1,000 mi) at each location where an AI chip could potentially be installed, no matter how many

landmark servers are used; worst-case accuracy is effectively capped in some regions of the world.¹²² PLV cannot, by default, guarantee worst-case 1,600 km accuracy for AI chips located either (a) very far from existing data centers (e.g., parts of Central Asia) or (b) in one of the removed areas where we do not expect to position landmark servers (i.e., the oceans, Antarctica, China, Russia, and the United States). For example, many locations in Mongolia and western Kazakhstan are more than 391 km (243 mi) from a data center and therefore would not be guaranteed a worst-case accuracy of 1,600 km. If BIS receives a license application to export an AI chip to a data center in a remote location, the Bureau could require the owner to permit installation and maintenance of a landmark server on-site.¹²³ Doing so would substantially improve the worst-case PLV accuracy for AI chips at that location by decreasing the distance to the nearest landmark server from more than 500 km to less than 1 km (more than 311 mi to less than 0.6 mi).

Delay Similarity

Delay similarity involves comparing the RTT from a landmark server to the AI chip with a separate RTT from the same landmark server to another trusted device near the AI chip's supposed location. Essentially, this method uses communication latency with the trusted device as a ground truth signal. If the two times are within a specified error bound of one another, then it indicates that the AI chip could be in the correct location. If the RTT times between the AI chip and each landmark server are consistent with each respective ground truth RTT, it would indicate that the AI chip is very likely in the expected location. Conversely, if an AI chip RTT to any landmark server does not match up with the ground truth RTT, it could indicate that the chip is not in its expected location.

While delay similarity would likely allow for more precise geolocation than trilateration, it would also require collecting ground truth telemetry, which could increase costs, depending on the quantity of trusted assets and the price to access them. Specifically, when each landmark server pings an AI chip, it would also need to ping a trusted device close to that AI chip in order to collect each ground truth RTT.

Appendix C: Centralized vs. Decentralized Location Verification

In the main text, we focus on a centralized system for location verification based on our assessment of the strengths and weaknesses of each. We elaborate on our evaluation here.

Under a **centralized** configuration, a single entity remains responsible for verifying the location of an AI chip throughout its lifecycle, even as the chip changes hands. The responsible entity should be a single entity (i.e., BIS, the chip designer, or a trusted third-party contractor).

When responsibility for location verification is **decentralized**, the last exporter assumes responsibility for verifying a chip's location with each subsequent sale of the chip. When the designer sells a chip to a customer, the responsibility remains with the chip designer until the customer resells the chip to downstream customers. In this system, when a company exports a controlled chip, it must also take charge of verifying the chip's location following its export until the chip is re-exported.

To illustrate how both configurations would work in practice, it is important to consider AI chip market dynamics. Chip designers sell to a variety of customers, many of whom are not the end user of the chips themselves. For instance, NVIDIA identifies two types of customers: direct and indirect.¹²⁴ “Direct customers” are those who purchase directly from NVIDIA, such as add-in board (AIB) manufacturers, distributors, system integrators, original device manufacturers (ODMs), and original equipment manufacturers (OEMs).¹²⁵ Some of NVIDIA's direct customers likely include Quanta Cloud Technology (QCT), Supermicro, and Dell.¹²⁶ “Indirect customers”—who purchase products through the “direct customers”—include cloud service providers (CSPs) (e.g., Amazon Web Services, Microsoft Azure, and Google Cloud Platform), consumer internet companies, enterprises, and public sector entities.¹²⁷

Consider the following hypothetical example for each system:

- (1) A chip designer sells 72 chips to an OEM in Malaysia. Under existing regulations, no export license is required for this initial export.
- (2) The OEM then assembles the 72 chips into a server rack.¹²⁸
- (3) The OEM then sells the server rack to an AI company, who intends to install the rack in a data center in Singapore.¹²⁹ No export license is required for this purchase under existing regulations.

In the **centralized configuration**, when the OEM makes its initial purchase from the chip designer, the OEM likely would not know to whom the final server rack will be sold.¹³⁰ Once the OEM sells the server rack containing the 72 chips to the AI company in Singapore, the centralized regulatory configuration would require the OEM to report this sale to the chip designer, including the intended location and end user of the chips.* Any subsequent sale by the AI company or downstream sellers would also have to be reported to the chip designer, which would be responsible for updating the database of intended chip locations and end users. This centralized regulatory configuration would place a burden on the chip designer, who is required to update the reporting servers with the intended location of each chip.

In the **decentralized configuration**, when the chip designer makes the initial sale to the OEM, the chip designer would be responsible for verifying the location of the 72 chips until the OEM sells the server rack to the AI company. At that point, the OEM, as the last exporter, would assume responsibility for verifying the location of the chips sold to the AI company in Singapore. The responsibility would continue to shift down to the last exporter with each subsequent transaction.

We recommend a centralized configuration to limit the system's vulnerability to undetected non-compliance. In a decentralized configuration, if a downstream customer is not a trustworthy actor, the system can be undermined in the same way that bad actors undermine the current export control system: through non-compliance. A bad actor could attest to having exported the chip to a permitted location and configured a location verification system when in reality they exported the chip to an unauthorized location without a verification mechanism in place. A centralized configuration, by contrast, limits this threat vector as a small number of entities are in charge of verifying locations.

* Major chip designers are already aware of who their major indirect customers are; however, this is likely only true for the major customers.

Appendix D: Comparative Cost-Benefit Analysis of Approaches for Location Verification of AI Chips

To determine whether PLV or physical inspections is more cost-effective, we modeled their benefits and costs in over ten million simulated scenarios. We built the model in Python and published the code on [GitHub](#).

We hypothesized that PLV would be more cost-effective than physical inspections; therefore, we deliberately chose model parameters favorable to physical inspections to determine whether PLV is more cost-effective even under those conditions. Moreover, we modeled two separate costs for PLV: one modeled a PLV system in which landmark servers are owned and operated directly by BIS, chip designers, or a third party, while the second modeled a PLV system in which landmark servers are rented from CSPs.

We compared the two methods using detections-per-cost: the expected number of diverted chips a method detects divided by the dollar cost (i.e., the higher the detections-per-cost, the more detections of diverted chips per dollar). **We found that PLV is more cost-effective than physical inspections. Physical inspections did not have more detections-per-cost than renting or owning PLV in any of the 10.5 million scenarios we simulated.** In the vast majority of modeled scenarios, both PLV options had strictly higher detections-per-cost; in the remaining scenarios, physical inspections tied one or both PLV options but never outperformed either. We conducted sensitivity checks on all key constants and concluded that the model outcome is not sensitive to any of our choices.

See our technical CSET blog post “[Tracking AI Chips: What Does It Cost?](#)” for a full breakdown of our model design, cost estimates, and sensitivity checks.

Appendix E: Chips Without TEEs

The PLV system outlined in this analysis involves configuring AI chips to securely (a) authenticate RTT challenges sent by the landmark servers and (b) generate and sign responses to each challenge. As explained in Appendix A, the security of this process is improved when it occurs within a trusted execution environment (TEE); however, not all AI chips have TEEs, especially those developed by companies other than NVIDIA. To contextualize how this may or may not be problematic for a PLV system, this appendix reviews the security features of three prominent brands of data center AI chips, each of which lacks dedicated TEEs: AMD Instinct GPUs, Google TPUs, and AWS Trainium ASICs.* Specifically, this appendix seeks to determine (1) whether each AI chip brand claims to have security features that approximate a TEE, and (2) whether software-based or firmware-based signing of each AI chip's response to an RTT challenge could proceed without hardware changes. While this paper only assesses the security features of four prominent AI chip brands (including NVIDIA chips), future work should do the same for chips developed by other designers, such as Cerebras Wafer Scale Engine (WSE), SambaNova RDU, Intel data center GPUs, Microsoft Maia, Meta MTIA, and OpenAI Jalapeño.

Google and AWS claim their respective TPUs and Trainium chips have a full set of security features that approximate a TEE, but AMD Instinct does not.¹³¹ **If policymakers choose to require a TEE for PLV implementation, they will need to decide whether features that claim to approximate the security of a TEE satisfy that requirement.**

Google's TPUs and AWS's Trainium chips do not appear to have an on-die hardware identity key, but they are integrated with other processors that do.¹³² It is unclear whether AMD's Instinct data center GPUs have a hardware identity key. **If policymakers choose to require a firmware-based PLV implementation, they will need to decide whether implementations like Google's and AWS' satisfy the requirement for secure signing using a hardware identity key.**

TEE-Like Security Features

AMD's Instinct data center GPUs do not have TEEs, and they do not claim to have other features that approximate the security of a TEE. Aside from secure boot and device attestation—features discussed later in this section—AMD does not appear to advertise other TEE-like security features for their Instinct GPUs.¹³³

* TPU stands for tensor processing unit. ASIC stands for application-specific integrated circuit.

Google's TPUs lack TEEs, but they claim to have other features that approximate the security of a TEE. Historically, Google has not directly sold their TPUs to external customers; however, Google has suggested that they may begin selling directly to external customers, making TPUs more relevant to location verification efforts.¹³⁴ Google's strategy for confidential computing revolves around securing their cloud infrastructure, through which users remotely interact with TPUs, as opposed to securing individual TPUs that are purchased directly by customers and installed in data centers outside of Google Cloud. To create a secure environment, Google uses AMD CPUs with TEEs "to encrypt and isolate memory and processing from the host," as well as their own Titanium Intelligence Enclave to create "a hardened TPU platform that delivers privacy and security properties comparable to a typical TEE."¹³⁵ This enclave acts as a "fortified space" for isolating and privately processing data within Google Cloud; it does not involve TPU TEEs because Google controls the chips, and their users' security and privacy can be built in at a higher level. If Google were to export TPUs, however, the lack of TEEs would present the same issue as AMD chips do.

AWS' Trainium chips also lack TEEs but, like Google's TPUs, they offer other off-chip features that approximate the security of a TEE at the system level. Also similar to Google TPUs, AWS historically has not directly sold their Trainium chips to external customers; however, AWS has suggested that they may begin selling directly to foreign, external customers, making Trainium more relevant to location verification efforts.¹³⁶ Trainium chips are integrated into the AWS Nitro System, which offloads security features (among other things) to dedicated hardware that works alongside CPUs and Trainium chips. Here, chip attestation is conducted by the Nitro System hardware, which provides hardware root of trust and cryptographic attestation for the whole system.¹³⁷ AWS also provides Nitro Enclaves that "further protect and securely process highly sensitive data such as personally identifiable information (PII), healthcare, financial, and intellectual property data within their Amazon EC2 instances."¹³⁸ Overall, the Nitro System provides TEE-like security features, but because they are off-chip and at the system level, the lack of dedicated on-chip TEEs would present the same issue as AMD and TPU AI chips.

Software-Based or Firmware-Based Signing Without a TEE

The key consideration is whether AMD, Google, and AWS AI chips have hardware identity keys, which are required for firmware-based signing. Software-based signing does not depend on a hardware identity key, nor any other security feature, so software-based signing can proceed on AMD Instinct GPUs, Google TPUs, and AWS Trainium ASICs without firmware or hardware changes. Firmware-based signing, however, requires that the RTT challenge response is signed by the AI chip's hardware

identity key, which is a unique private key fused into each device during production.¹³⁹ Therefore, if AMD, Google, and AWS AI chips do not have a hardware identity key, they cannot support firmware-based signing as specified.

AMD's Instinct data center GPUs may have hardware identity keys. For example, AMD's Instinct MI350 series chips can securely boot and update to "ensure only trusted firmware runs, while Platform-Level DICE Identity & Attestation verifies GPU authenticity to prevent unauthorized access."¹⁴⁰ This may mean that AMD's latest Instinct GPUs are equipped with a hardware identity key that underlies device attestation. According to an IAPS report, AMD's AI chips "are already equipped with security modules that could likely be used to implement this functionality via a firmware update" and they "also include some form of platform security processor that could be used to implement location attestation."¹⁴¹ Notwithstanding these claims, if AMD's Instinct GPUs do not have a unique hardware identity key, then designers will not be able to implement secure firmware-based signing on those GPUs (as specified in Appendix A).

Google's TPUs do not appear to have an on-die hardware identity key, but they are integrated with other processors that do. Google's TPUs are deployed alongside separate Titan chips, each of which contains a hardware identity key among other cryptographic features.¹⁴² This implementation contrasts with NVIDIA's latest data center AI chips where the hardware identity key is directly fused into the AI chip's die. It is likely that Google's solution is lower cost but also lower security than NVIDIA's.

AWS' Trainium chips also do not appear to have on-die hardware identity keys but, like Google TPUs, they are integrated with other processors that do. As previously stated, chip attestation and hardware root of trust are provided by Nitro System hardware.¹⁴³ Like Google TPUs, this contrasts with NVIDIA's latest data center AI chips where the hardware identity key is directly fused into the AI chip's die.

Overall, the lack of a TEE does not necessarily prevent an AI chip from being tracked in a PLV system, but it does reduce the security of the implementation. AMD Instinct GPUs, Google TPUs, and AWS Trainium ASICs do not feature dedicated TEEs, meaning PLV implemented on any of these chips would likely be less secure than on chips with TEEs. These chips can support software-based signing, though this is the least secure approach. Alternatively, firmware-based signing may also be feasible for AMD chips if they have hardware identity keys, for Google TPUs through a companion Titan chip, and for AWS Trainium chips through Nitro System hardware. Policymakers will need to determine the minimum acceptable security of a PLV system, which in turn will define which existing chips could feasibly participate in a PLV system.

Endnotes

¹ Office of Public Affairs, “Chinese National and Two U.S. Citizens Charged with Conspiring to Smuggle Artificial Intelligence Technology to China,” U.S. Department of Justice, March 25, 2026, <https://www.justice.gov/opa/pr/chinese-national-and-two-us-citizens-charged-conspiring-smuggle-artificial-intelligence>; Office of Public Affairs, “Two Chinese Nationals Arrested on Complaint Alleging They Illegally Shipped to China Sensitive Microchips Used in AI Applications,” U.S. Department of Justice, August 5, 2025, <https://www.justice.gov/opa/pr/two-chinese-nationals-arrested-complaint-alleging-they-illegally-shipped-china-sensitive>.

² Qianer Liu, “Nvidia AI Chip Smuggling to China Becomes an Industry,” *The Information*, August 2024, <https://www.theinformation.com/articles/nvidia-ai-chip-smuggling-to-china-becomes-an-industry>.

³ Isabel Juniewicz, “Diversion and Resale: Estimating Compute Smuggling to China,” *Epoch AI*, April 29, 2026, <https://epoch.ai/publications/chip-smuggling>.

⁴ In Q1 2026, the largest AI cluster tracked by Epoch AI was Anthropic-Amazon New Carlisle, which had 471,000 H100-equivalents of compute at the time. The second largest cluster was Microsoft Fairwater Atlanta, which had 374,000 H100-equivalents. Therefore, we judge that 663,000 H100-equivalents of smuggled compute was roughly 1–2 frontier-level clusters’ worth of compute in Q1 2026. “AI Data Centers,” *Epoch AI*, accessed July 8, 2026, <https://epoch.ai/data/ai-data-centers>.

⁵ While the smuggled 663,000 H100-equivalent GPUs were almost certainly installed in more than two clusters, and therefore could not have been used for a single training run, this example shows what these chips could be capable of. To calculate the lower- and upper-bound wall-clock time to train a GPT-5-equivalent model on 331,500 H100 SXM GPUs, for each primary stage of model training (i.e., pre-training and post-training) we divide GPT-5 FLOP (estimated by Epoch AI) by the aggregate effective FLOP/s of 331,500 H100 GPUs. Our calculation assumes the model is trained in FP16 precision using the H100’s tensor cores. To calculate the H100’s effective FLOP/s, we assume pre-training model FLOP utilization (MFU) ranges from 30–45% and post-training (e.g., reinforcement learning) MFU ranges from 5–20%, although these ranges can shift depending on the AI chips used, type of post-training, model architecture, and precision. Moreover, the lower ranges for post-training are primarily due to low reinforcement learning (RL) MFU. While supervised fine-tuning (SFT) MFU may be higher on average than RL MFU, for the purposes of our calculation, we assume that RL will dominate post-training compute and therefore we use a range biased towards RL. Our calculations are available in this Google Sheets workbook:

<https://docs.google.com/spreadsheets/d/1p8w67MQDaZPjrisXcZrwkMWeqQx6dQbSe8Q0380Uw/edit?gid=1319530489#gid=1319530489>; Aakanksha Chowdhery, Sharan Narang, Jacob Devlin et al., “PaLM: Scaling Language Modeling with Pathways,” arXiv preprint arXiv:2204.02311 (2022), <https://arxiv.org/abs/2204.02311>; Houwen Peng, Kan Wu, Yixuan Wei et al., “FP8-LM: Training FP8 Large Language Models,” arXiv preprint arXiv:2310.18313 (2023), <https://arxiv.org/abs/2310.18313>; Zhengding Hu, Hehua Ouyang, Chang Chen et al., “JigsawRL: Assembling RL Pipelines for Efficient LLM Post-Training,” arXiv preprint arXiv:2604.23838 (2026), <https://arxiv.org/abs/2604.23838>; Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri et al., “The Llama 3 Herd of Models,” arXiv preprint arXiv:2407.21783 (2024), <https://arxiv.org/abs/2407.21783>; Jean-Stanislas Denain and Cheryl Wu,

“Final Training Runs Account for a Minority of R&D Compute Spending,” *Epoch AI*, March 23, 2026, <https://epoch.ai/gradient-updates/r-and-d-vs-training-compute>.

⁶ Office of Public Affairs, “Three Charged with Conspiring to Unlawfully Divert Cutting Edge U.S. Artificial Intelligence Technology to China,” U.S. Department of Justice, March 19, 2026, <https://www.justice.gov/opa/pr/three-charged-conspiring-unlawfully-divert-cutting-edge-us-artificial-intelligence>; “Sealed Redacted Indictment,” *United States v. Liaw*, No. 1:26-cr-00100 (S.D.N.Y. March 17, 2026), ECF No. 2, <https://storage.courtlistener.com/recap/gov.uscourts.nysd.660030/gov.uscourts.nysd.660030.2.0.pdf> [<https://perma.cc/YQ3U-GAB5>].

⁷ This upper-bound estimate assumes that all \$2.5 billion of chips were NVIDIA B200 GPUs integrated into GB200 NVL72 rack-scale servers, which contain 72 GPUs each. In reality, according to the indictment, some percentage of diverted chips were lower-performance NVIDIA H200 GPUs. SemiAnalysis estimates that each GB200 NVL72 server alone (without additional infrastructure costs) costs \$3.1 million. “H100 vs GB200 NVL72 Training Benchmarks,” *SemiAnalysis*, August 20, 2025, <https://newsletter.semianalysis.com/p/h100-vs-gb200-nvl72-training-benchmarks>. Dividing \$2.5 billion by \$3.1 million per 72-GPU rack yields approximately 806 racks, which contain approximately 58,032 B200 GPUs. 58,032 B200 GPUs are equivalent to approximately 146,620 H100 GPUs, weighted by their relative peak FP16 performance over dense matrices ($58,032 \times \frac{2,500}{989.5} = 146,620$ H100-equivalent GPUs). We round this rough upper-bound estimate to 145,000 H100-equivalents since its purpose is to illustrate scale rather than provide a precise count. “Datasheet for NVIDIA Blackwell Architecture,” NVIDIA, accessed March 20, 2025, <https://nvdam.widen.net/s/wwnsxrh2w/blackwell-datasheet-3384703> [<https://perma.cc/VB8S-NRFK>]; “NVIDIA H100 GPU Datasheet,” NVIDIA, accessed May 7, 2026, <https://resources.nvidia.com/en-us-gpu-resources/h100-datasheet-24306> [<https://perma.cc/3QQY-B9PT>].

⁸ Chip Security Act, H.R. 3447, 119th Cong. (2025), <https://www.congress.gov/bill/119th-congress/house-bill/3447/text>.

⁹ Maggie Eastland and Oma Seddiq, “Chip Security Bill Advances in House After Super Micro Case,” *Bloomberg*, March 26, 2026, <https://www.bloomberg.com/news/articles/2026-03-26/chip-security-bill-advances-in-house-following-super-micro-case>; “Tech: Chips bills poised for NDAA ride,” *Punchbowl News*, July 14, 2026, <https://punchbowl.news/article/tech/chips-bills-ndaa-ride/>. The text of the bill was amended from the original bill introduced in May 2025. For the latest version of the text, see “Amendment in the Nature of a Substitute to H.R. 3447,” U.S. House Committee on Foreign Affairs, March 26, 2026, <https://docs.house.gov/meetings/FA/FA00/20260326/118999/BILLS-1193447ANSih-U2.pdf>.

¹⁰ “Amendment in the Nature of a Substitute to H.R. 3447,” U.S. House Committee on Foreign Affairs, March 26, 2026, <https://docs.house.gov/meetings/FA/FA00/20260326/118999/BILLS-1193447ANSih-U2.pdf>. Note that periodic verification is required to ensure AI chips are not smuggled after they are initially geolocated, so mechanisms such as external GPS trackers placed in chip packaging during shipment are excluded from the scope of the Act. Fanny Potkin, Karen Freifeld, and Jun Yuan Yong, “Exclusive: US Embeds Trackers in AI Chip Shipments to Catch Diversions to China, Sources Say,” *Reuters*, August 13, 2025, <https://www.reuters.com/world/china/us-embeds-trackers-ai-chip-shipments-catch-diversions-china-sources-say-2025-08-13/>.

¹¹ “Amendment in the Nature of a Substitute to H.R. 3447,” U.S. House Committee on Foreign Affairs, March 26, 2026, <https://docs.house.gov/meetings/FA/FA00/20260326/118999/BILLS-1193447ANSih-U2.pdf>.

¹² The bill identifies three potential chip security mechanisms: “(A) periodic on-site audits or inventories at the end user’s approved destination for the covered integrated circuit product,” “(B) periodic attestations by a U.S.-headquartered entity, or its subsidiaries, confirming that all covered integrated circuit products are accounted for, provided the Secretary determines that the U.S.-headquartered entity or its subsidiaries verifiably certifies that the U.S.-headquartered entity or its subsidiaries maintain continuous and sufficiently secure control, operation, repair (to the extent such repair is conducted by or under the direct supervision of the U.S.-headquartered entity or its subsidiaries), and disposal of the covered integrated circuit products,” and “(C) ping-based location verification through a trusted landmark server utilizing secure software- or firmware-enabled mechanisms.” The bill also states that these are not exhaustive and chip security mechanisms could be “various other mechanisms, or combinations of mechanisms, that the Secretary determines can verifiably demonstrate with significant confidence that the covered integrated circuit product has not been illegally diverted to a destination of concern.” “Amendment in the Nature of a Substitute to H.R. 3447,” U.S. House Committee on Foreign Affairs, March 26, 2026, <https://docs.house.gov/meetings/FA/FA00/20260326/118999/BILLS-1193447ANSih-U2.pdf>.

¹³ The legislation would also task Commerce with assessing enhancements to chip security mechanisms that would detect smuggling and prevent the unauthorized use of smuggled chips. H.R. 3447 § 6(b) requires the Secretary of Commerce, in consultation with the Secretary of State, the Secretary of Defense, the Secretary of Energy as well as the public, to conduct an assessment to identify what enhancements should be used to improve the chip security mechanisms implemented to enhance compliance; to detect illegal diversion; to identify and monitor smuggling intermediaries; to ensure U.S. technological leadership; to ensure the orderly and effective implementation of chip security mechanisms; and to address industry feedback. “Amendment in the Nature of a Substitute to H.R. 3447,” U.S. House Committee on Foreign Affairs, March 26, 2026, <https://docs.house.gov/meetings/FA/FA00/20260326/118999/BILLS-1193447ANSih-U2.pdf>.

¹⁴ White House, *Winning the Race: America’s AI Action Plan* (Washington, DC: White House, July 2025), <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.

¹⁵ Semiconductor Industry Association, “SIA Statement on Chip Security Act,” statement of John Neuffer, press release, March 2, 2026, <https://www.semiconductors.org/sia-statement-on-chip-security-act/>.

¹⁶ Onni Aarne, Tim Fist, and Caleb Withers, “Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing” (Center for a New American Security, January 2024), <https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf>.

¹⁷ Monica J. White, “How to Find Your GPU’s Serial Number (And Why You Might Need To),” *SlashGear*, June 17, 2025, <https://www.slashgear.com/1887100/gpu-serial-number-how-to-find-what-for-explained/>.

¹⁸ IAPS estimates that physical checks would cost about \$2.5 million per million chips per year to detect 95% of chip diversion within 6 months. We estimate that there are roughly 3 million data center chips subject to advanced computing export controls. We come to this value by using Epoch AI's "AI Chip Owners" dataset and excluding chips (a) sold to China and (b) owned by U.S.-headquartered CSPs or other U.S. hyperscalers. Assuming that the locations of 3 million chips would be subject to location verification, the IAPS estimates suggest it would cost about \$7.5 million annually to detect 95% of chip diversion within 6 months. Using our own cost model alongside IAPS assumptions for the number of clusters and chips inspected, we estimate a lower bound of \$6.3 million and an upper bound of \$14.7 million. IAPS's calculation looks similar to our lower-bound estimate. Our calculations are available in this Google Sheets workbook:

<https://docs.google.com/spreadsheets/d/1jz9wLM0gWkwgCuqcufysz3ly9UF-vnBHUDxVp3ViNAw/edit?gid=1183540240#gid=1183540240>. However, IAPS's model for detection of chip smuggling likely overestimates the probability of detecting smuggled chips through physical checks, especially in cases where only a portion of chips within a given cluster are smuggled. IAPS assumes full cluster diversion when diversion occurs and does not consider diversion of portions of clusters. Therefore, to account for those scenarios and other uncertainties, we modeled costs and benefits for physical inspections across ten million simulated scenarios (see Appendix D). Onni Aarne and Erich Grunewald, "Export Auditors as Market-Powered Export Enforcement" (Institute for AI Policy and Strategy, March 2026), <https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/69bc642c9cbbd038d27a9660/1773954092111/Export+Auditors+as+Market-Powered+Export+Enforcement.pdf>; "AI Chip Owners," Epoch AI, accessed July 8, 2026, <https://epoch.ai/data/ai-chip-owners?tab=count>.

¹⁹ This attestation would likely involve disconnecting a server or compute sled that contains AI chips from its rack, plugging it into a mobile workstation, and using software toolkits such as nvTrust to attest the authenticity of each chip. NVIDIA, "nvTrust: Ancillary Software for NVIDIA Trusted Computing Solutions," GitHub, version 2026.06.04.001, released May 4, 2026, <https://github.com/NVIDIA/nvtrust>; "NVIDIA Attestation," NVIDIA, October 28, 2024, <https://docs.nvidia.com/attestation/index.html>; White, "How to Find Your GPU's Serial Number."

²⁰ A useful anecdote is that according to the March 2026 Department of Justice press release announcing the indictment of three chip smugglers, "a hair dryer [was used] to remove and affix labels and serial number stickers to the server boxes and to the dummy servers themselves" and the smugglers "re-packag[ed] the dummy servers in the U.S. Manufacturer's boxes," demonstrating that less-intrusive checks that do not examine individual chips are insufficient. U.S. Attorney's Office, Southern District of New York, "Three Charged with Conspiring to Unlawfully Divert U.S. Artificial Intelligence Technology to China," U.S. Department of Justice, March 19, 2026, <https://www.justice.gov/usao-sdny/pr/three-charged-conspiring-unlawfully-divert-us-artificial-intelligence-technology-china>.

²¹ Asher Brass and Onni Aarne, "Location Verification for AI Chips" (Institute for AI Policy and Strategy, May 2024), <https://www.iaps.ai/research/location-verification-for-ai-chips>; Aarne, Fist, and Withers, "Secure, Governable Chips"; Samuel Hammond, "The Chip Security Act: A Bipartisan Solution to Chip Smuggling," Foundation for American Innovation, July 18, 2025, <https://www.thefai.org/posts/the-chip-security-act-a-bipartisan-solution-to-chip-smuggling>.

²² “4 Geolocation Technologies Compared,” Sensolus, accessed July 8, 2026, <https://www.sensolus.com/knowledge-hub/behind-the-tech/four-geolocation-technologies-compared-how-can-they-improve-your-operational-efficiency/>; Emma Rivero and Ed Sperling, “Location Verification Becomes Much Bigger Concern for Chips,” *Semiconductor Engineering*, October 9, 2025, <https://semiengineering.com/location-verification-becomes-much-bigger-concern-for-chips/?cmid=73314953-d2d8-4d9d-ae4c-a76a2b245cce>.

²³ “What Is RTT in Networking?,” Amazon Web Services, accessed July 8, 2026, <https://aws.amazon.com/what-is/rtt-in-networking/> [<https://perma.cc/QU3Q-5A9T>]; “Trilateration,” ScienceDirect Topics, Elsevier, accessed July 8, 2026, <https://www.sciencedirect.com/topics/engineering/trilateration>.

²⁴ An error-circle radius of 187 km (116 miles) or less—equivalent to the size of the 100th-largest country—is a reasonable threshold for country-level accuracy. See subsection “Positioning Landmark Servers” in Appendix B for more details.

²⁵ We are less concerned about vulnerabilities that cannot be readily exploited at scale. For example, if a vulnerability can be exploited to obfuscate chip locations but it requires extensive resources compared to the chip’s value, then bad actors would incur significant costs if they sought to exploit it across many chips.

²⁶ While the false-positive rate for physical inspections is likely lower than for PLV, it is not zero. An AI chip selected for physical inspection may not be in its expected data center because it has been removed for legitimate reasons; the chip may have been temporarily removed for maintenance, moved to a different cluster for legitimate reasons, or decommissioned.

²⁷ But if BIS were to allow U.S.-headquartered CSPs to submit inventory reports instead of implementing a PLV system, it would raise two questions: (a) should foreign branches and subsidiaries of these firms be trusted to the same degree, and (b) how would PLV be enabled on controlled AI chips re-sold by a trusted CSP?

²⁸ Worst-case accuracy is not guaranteed for data center AI chips located in bodies of water, Antarctica, the United States, China, and Russia (i.e., locations where landmark servers likely would not be placed). This does not undermine the value of the system, however, because if an AI chip’s feasible region is larger than expected for the chip’s expected location and overlaps with China, for example, it should indicate to the responsible entity that diversion may be occurring and further investigation may be necessary.

Furthermore, worst-case accuracy is not, by default, guaranteed for AI chips installed in a remote region (i.e., more than 391 km from an existing data center). BIS could address this limitation by requiring new, remote data centers to host on-site landmark servers. See Appendix B for more details.

²⁹ Secure Technology Alliance, *Trusted Execution Environment (TEE) 101: A Primer*, version 1.1 (June 2018), <https://www.securetechalliance.org/wp-content/uploads/TEE-101-White-Paper-V1.1-FINAL-June-2018.pdf>.

³⁰ Jianwei Zhu, Hang Yin, Peng Deng, Aline Almeida, and Shunfan Zhou, “Confidential Computing on NVIDIA Hopper GPUs: A Performance Benchmark Study,” arXiv preprint arXiv:2409.03992 (2024), <https://arxiv.org/abs/2409.03992>; Marcin Chrapek, Marcin Copik, Etienne Mettaz, and Torsten Hoefer, “Confidential LLM Inference: Performance and Cost Across CPU and GPU TEEs,” arXiv preprint arXiv:2509.18886 (2025), <https://arxiv.org/abs/2509.18886>.

³¹ The TEE is optional, but the chip must have a hardware identity key or some similarly secure way of attesting the device’s identity.

³² Dan Goodin, “New Physical Attacks Are Quickly Diluting Secure Enclave Defenses from Nvidia, AMD, and Intel,” *Ars Technica*, October 29, 2025, <https://arstechnica.com/security/2025/10/new-physical-attacks-are-quickly-diluting-secure-enclave-defenses-from-nvidia-amd-and-intel/>; Alex Seto et al., “WireTap: Breaking Server SGX via DRAM Bus Interposition,” in *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, 2025), <https://wiretap.fail/>.

³³ Another alternative would be to require operating licenses, which would remotely shut chips off if they are unresponsive. However, this is outside the scope of this analysis, which focuses on location verification mechanisms only.

³⁴ Brass and Aarne, “Location Verification for AI Chips.”

³⁵ While this attack is easy to implement, it is unlikely to succeed in convincing U.S. authorities that a smuggled chip is still in its intended location. If an AI chip’s feasible region is larger than expected for the chip’s expected location and overlaps with China, for example, it should automatically indicate to the responsible entity that diversion may be occurring, and further investigation may be necessary. Furthermore, so long as the maximum theoretical velocity of RTT requests and responses is set at the speed of light in a vacuum, the AI chip’s feasible area will always include the true location of the AI chip. For the feasible area to not include the true location of the AI chip, the bounding disk around at least one landmark server would need to be small enough that it does not encompass the true location of the AI chip. However, that condition can only be met with faster-than-light travel, so long as the equation used to calculate each landmark server’s bounding disk assumes that the RTT request and response traveled at the speed of light in a vacuum. Therefore, the feasible region of each AI chip must include its true location, and if the chip has been smuggled, the PLV system would indicate that the AI chip could be in an unauthorized location.

³⁶ Anton Shilov, “Price of Nvidia’s Vera Rubin NVL72 Racks Skyrockets to as Much as \$8.8 Million Apiece, but Server Makers’ Margins Will Be Tight — Nvidia Is Moving Closer to Shipping Entire Full-Scale Systems,” *Tom’s Hardware*, March 24, 2026, <https://www.tomshardware.com/tech-industry/artificial-intelligence/price-of-nvidias-vera-rubin-nvl72-racks-skyrockets-to-as-much-as-usd8-8-million-apiece-but-server-makers-margins-will-be-tight-nvidia-is-moving-closer-to-shipping-entire-full-scale-systems>.

³⁷ Data Center Map, “Paris Data Centers,” accessed July 8, 2026, <https://www.datacentermap.com/france/paris/>; Data Center Map, “Singapore Data Centers,” accessed July 8, 2026, <https://www.datacentermap.com/singapore/singapore/>; Data Center Map, “Kuala Lumpur Data Centers,” accessed July 8, 2026, <https://www.datacentermap.com/malaysia/kuala-lumpur/>.

³⁸ This does not undermine the value of the system. If an AI chip's feasible region overlaps with China, for example, it should indicate to the responsible entity that diversion may be occurring, and further investigation may be necessary.

³⁹ According to Synergy Research Group, colocated data centers make up an estimated 20% of the global data center capacity market as of 2025. Furthermore, while we are not aware of instances where companies have transferred hardware between cages colocated in the same data center, it remains a vulnerability in a location verification system. Mark Haranas, "Data Center Market Share Face-Off: Hyperscalers vs. Colocation vs. Enterprise," CRN, April 10, 2026, <https://www.crn.com/news/data-center/2026/data-center-market-share-face-off-hyperscalers-vs-colocation-vs-enterprise>.

⁴⁰ CSET calculations (available in this Google Sheets workbook: <https://docs.google.com/spreadsheets/d/1am0WdvZpuFG6utluAuhwEXTmKz95NnN2xk1jcVQL2g8/edit?gid=2098198945#gid=2098198945>); BIS's enacted FY2026 total budget was \$235M. Bureau of Industry and Security, *Budget Estimates, Fiscal Year 2027: President's Budget Request* (Washington, DC: U.S. Department of Commerce, April 2026), BIS-5, <https://www.commerce.gov/sites/default/files/2026-04/FY2027-BIS-CJ-Submission.pdf> [<https://perma.cc/FZT4-2VH5>].

⁴¹ For simplicity, our cost estimates for both owning and renting servers assume AI chips receive a firmware update to follow the PLV protocol; we do not consider costs of software or hardware updates. In addition to the costs of designing the PLV protocol and updating the firmware, our estimates for owning the landmark servers include the costs of (a) purchasing the servers and hardware security modules, (b) general overhead to purchase peripheral and replacement equipment, (c) the salaries of full-time employees maintaining the PLV system, including but not limited to the landmark servers, and (d) colocation costs of the servers (i.e., power, physical space, and network bandwidth). Our estimates for renting servers from CSPs include the costs of (a) renting different types of servers and hardware security modules, and (b) the salaries of full-time employees maintaining the landmark servers. We note that hardware security modules are the primary area where the costs of owning versus renting diverge. The annual cost of owning a single hardware security module is \$1,102–\$8,500, while the cost of renting a hardware security module is \$12,264–\$53,086.

⁴² The TEE is optional, but the chip must have a hardware identity key or some similarly secure way of attesting the device's identity.

⁴³ Dan Goodin, "New Physical Attacks Are Quickly Diluting Secure Enclave Defenses from Nvidia, AMD, and Intel," *Ars Technica*, October 29, 2025, <https://arstechnica.com/security/2025/10/new-physical-attacks-are-quickly-diluting-secure-enclave-defenses-from-nvidia-amd-and-intel/>; Seto, "WireTap."

⁴⁴ ECCN 3A090 includes integrated circuits ("ICs") that meet the performance thresholds defined in the Commerce Control List. ECCN 4A090 covers computers, electronic assemblies, and components that contain ICs that are covered under 3A090. 5A002.z are encryption items that contain ICs meeting the 3A090 thresholds.

⁴⁵ The first category refers to data center AI chips that meet the combination of peak performance (i.e., total processing performance, TPP) and performance density (i.e., TPP divided by die area) specifications that define ECCNs 3A090.a and 3A090.b.

⁴⁶ The peak performance of NVIDIA's Rubin GPU in Tensor FP4 precision over dense matrices is 35 PFLOPS, which equals 140,000 total processing performance (TPP). The first category (i.e., data center AI chips above a specified performance threshold) includes chips with TPP as low as 1,600.

$$100 \times \frac{1,600}{140,000} = 1.1\%$$

"NVIDIA Vera Rubin NVL72," NVIDIA, accessed January 23, 2026, <https://www.nvidia.com/en-us/data-center/vera-rubin-nvl72/> [<https://perma.cc/W9AV-E7P6>].

⁴⁷ BIS determines which countries are subject to which restrictions by categorizing them into Country Groups. Country Group D:5 and Macau (abbreviated as "D:5M") refers to Macau and U.S.-arms-embargoed countries, including Russia, Iran, and China.

⁴⁸ The second category refers to non-data center AI chips that meet the peak performance (i.e., total processing performance, TPP) threshold that defines ECCN 3A090.a.1.

⁴⁹ This would present a particular challenge for a worldwide policy scope but not a country-of-concern policy scope. See section "Reporting Scope" for more.

⁵⁰ "How Long Do GPUs Last?," RePCB, July 30, 2025, <https://www.repcb.co.uk/how-long-do-gpus-last/>; Anton Shilov, "Datacenter GPU Service Life Can Be Surprisingly Short—Only One to Three Years Is Expected According to Unnamed Google Architect," *Tom's Hardware*, October 24, 2024, <https://www.tomshardware.com/pc-components/gpus/datacenter-gpu-service-life-can-be-surprisingly-short-only-one-to-three-years-is-expected-according-to-unnamed-google-architect>.

⁵¹ Bureau of Industry and Security, "Foreign Availability Assessments," Department of Commerce, accessed July 8, 2026, <https://www.bis.gov/about-bis/bis-leadership-and-offices/OTE/foreign-availability-assessments>.

⁵² There are several possible configurations: (1) AI chip designers could each operate their own network of landmark servers while BIS operates the reporting server; (2) BIS could operate a single network of landmark servers while chip designers each operate their own reporting server; or (3) either entity (i.e., BIS and chip designers) could operate both. In each case, the responsible entity may either directly operate the servers or contract with a third party rather than operate the infrastructure directly. In theory, multiple AI chip designers could also contract with the same company to operate landmark and/or reporting servers.

⁵³ See also compliance in the environmental regulation context. Under the Renewable Fuel Standard program (RFS), EPA requires renewable fuel producers to engage and pay for independent third-party engineers and auditors to verify compliance with program requirements. The EPA accredits the third parties and receives the reports directly. [40 C.F.R. § 80.1450\(b\)\(2\)](#); [40 C.F.R. § 80.1464](#).

⁵⁴ IAPS researchers suggest this approach is unsuitable because the AI chip market is heavily concentrated, meaning that AI chip designers with high market share would have significant financial leverage over their supposedly independent auditor, a significant conflict of interest. However, this risk may be overstated considering that AI chip designers often do not directly export data center hardware to the final customer. Therefore, even if NVIDIA holds 90% of the AI accelerator market, the client lists of third-party auditors would also include server OEMs, who purchase NVIDIA hardware, and their

distributors (Appendix C). Onni Aarne and Erich Grunewald, “Export Auditors as Market-Powered Export Enforcement” (Institute for AI Policy and Strategy, March 2026), <https://www.iaps.ai/research/export-auditors-as-market-powered-export-enforcement>.

⁵⁵ These estimates include PLV maintenance but exclude the cost of standards and firmware development. See section “Benefits and Costs of a PLV System” for more details on our cost estimates.

⁵⁶ 15 C.F.R. pt. 740, supp. no. 1 (2026), <https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-740/appendix-Supplement%20No.%201%20to%20Part%20740>.

⁵⁷ 15 C.F.R. pt. 740, supp. no. 1, <https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-740/appendix-Supplement%20No.%201%20to%20Part%20740>.

⁵⁸ Under a worldwide scope, the system would not be able to detect when a U.S. chip is sold or resold to an export-controlled end user, including entities located or headquartered in China. Under a country-of-concern scope, the system likewise would be unable to detect when U.S. chips are diverted to China or other countries of concern.

⁵⁹ This is unlikely to be a significant issue because the ICTS office is housed within BIS; coordination on these complementary policies should be feasible.

⁶⁰ 15 C.F.R. pt. 791 (2026), <https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-E/part-791>; David Plotinsky and Jiazhen Guo, “Securing the ICTS Supply Chain: Commerce Issues Final Rules Pursuant to EO 13873,” Morgan Lewis, December 2024, <https://www.morganlewis.com/pubs/2024/12/securing-the-icts-supply-chain-commerce-issues-final-rules-pursuant-to-eo-13873>; “Securing the Information and Communications Technology and Services Supply Chain,” 89 FR 96872 (December 6, 2024), <https://www.federalregister.gov/documents/2024/12/06/2024-28335/securing-the-information-and-communications-technology-and-services-supply-chain>; Ian Cohen, “Lawmakers Want to Codify ICTS Authorities, BIS Chief Says,” *International Trade Today*, March 28, 2025, https://internationaltradetoday.com/article/2025/03/28/lawmakers-want-to-codify-icts-authorities-bis-chief-says-2503280064?BC=bc_685d990daf901.

⁶¹ U.S. persons controls could provide another potential avenue for regulating U.S. chips; however, this would require amending existing export controls on U.S. persons, and analyzing whether and how to do so is beyond the scope of this paper.

⁶² Office of Public Affairs, “Three Charged with Conspiring to Unlawfully Divert Cutting Edge U.S. Artificial Intelligence Technology to China,” U.S. Department of Justice, March 19, 2026, <https://www.justice.gov/opa/pr/three-charged-conspiring-unlawfully-divert-cutting-edge-us-artificial-intelligence>.

⁶³ “Trade Compliance Flash: Senate Committee Releases Report Assessing Commerce’s Enforcement of Export Controls,” Miller & Chevalier, January 16, 2025, <https://www.millerchevalier.com/publication/trade-compliance-flash-senate-committee-releases-report-assessing-commerces-enforcement>; Majority Staff of the Senate Permanent Subcommittee on Investigations, 118th Cong., *The U.S. Technology Fueling Russia’s War in Ukraine: Examining the Bureau of Industry and Security’s Enforcement of Semiconductor Export Controls* (Washington, DC: U.S.

Senate Committee on Homeland Security and Governmental Affairs, December 18, 2024), <https://www.hsgac.senate.gov/wp-content/uploads/The-U.S.-Technology-Fueling-Russias-War-in-Ukraine-Examining-BISs-Enforcement-of-Semiconductor-Export-Controls.pdf>; Commerce, Justice, Science; Energy and Water Development; and Interior and Environment Appropriations Act, 2026, Pub. L. No. 119-74 (2026), <https://www.congress.gov/bill/119th-congress/house-bill/6938/text>; White House, “Congressional Bill H.R. 4323 and H.R. 6938 Signed into Law,” January 23, 2026, <https://www.whitehouse.gov/briefings-statements/2026/01/congressional-bill-h-r-4323-and-h-r-6938-signed-into-law/>.

⁶⁴ NVIDIA, “nvTrust.”; NVIDIA, “NVIDIA Attestation.”; White, “How to Find Your GPU’s Serial Number.”

⁶⁵ Mail-in inspections can also help identify (but not verify) chip owners. For more details on mail-in inspections, see: Erich Grunewald and Bruna Avellar, “Near-Term Verification Methods for AI Chip Exports” (Institute for AI Policy and Strategy, August 2026), <https://www.iaps.ai/research/near-term-verification-methods-for-ai-chip-exports>; Tim Fist and Erich Grunewald, “Preventing AI Chip Smuggling to China: A Working Paper” (Center for a New American Security, October 2023), <https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china>.

⁶⁶ To conceal the country from which a smuggled AI chip was actually shipped, a bad actor would likely need to route it in two stages: first, from the export-controlled country where it was located to the country where it was expected to be; and second, from that country to the designated inspection office.

⁶⁷ Grunewald and Avellar, “Near-Term Verification Methods for AI Chip Exports.”

⁶⁸ “What Is DCIM?,” Sunbird DCIM, accessed July 8, 2026, <https://www.sunbirddcim.com/what-dcim>; “Opt-In Software Enables Data Center Fleet Management,” NVIDIA Blog, December 10, 2025, <https://blogs.nvidia.com/blog/optional-data-center-fleet-management-software/>.

⁶⁹ NVIDIA Fleet Intelligence allows a user to monitor their chips as well as export a signed report detailing information about those chips. Fleet Intelligence, which is built on NVIDIA’s Data Center GPU Manager (DCGM) and NVIDIA Device Identity Services (NDIS), determines whether a connected chip is authentic by comparing a cryptographic key—derived from the chip’s unique, hardware-fused identity key—with NVIDIA’s own registry of valid keys. As such, a Fleet Intelligence report could list verified product serial numbers or count numbers of verified, controlled chips by product type (SKU). Christian Shrauder and Gail Frederick, “Introducing NVIDIA Fleet Intelligence for Real-Time GPU Fleet Visibility and Optimization,” *NVIDIA Technical Blog*, May 11, 2026, <https://developer.nvidia.com/blog/introducing-nvidia-fleet-intelligence-for-real-time-gpu-fleet-visibility-and-optimization/> [<https://perma.cc/W5C5-HAVR>]; “Opt-In NVIDIA Software Enables Data Center Fleet Management,” *NVIDIA Newsroom*, December 10, 2025, <https://blogs.nvidia.com/blog/optional-data-center-fleet-management-software/> [<https://perma.cc/3Y3R-7537>]; “Get inventory report,” NVIDIA Fleet Intelligence Documentation, accessed August 17, 2026, <https://docs.nvidia.com/fleet-intel/latest/api-explorer/reports/get-v-1-reports-inventory/> [<https://perma.cc/6PWU-FUJL>]; “NVIDIA DCGM,” NVIDIA Developer, accessed August 17, 2026, <https://developer.nvidia.com/dcgm> [<https://perma.cc/PKV3-DNJZ>]; “Field Identifiers,” NVIDIA DCGM Documentation, accessed July 8, 2026, <https://docs.nvidia.com/datacenter/dcgm/latest/dcgm-api/dcgm-api-field-ids.html> [<https://perma.cc/5WMP-93WL>].

⁷⁰ Brass and Aarne, “Location Verification for AI Chips”; Aarne, Fist, and Withers, “Secure, Governable Chips”; Samuel Hammond, “The Chip Security Act: A Bipartisan Solution to Chip Smuggling” (Foundation for American Innovation (FAI), July 18, 2025), <https://www.thefai.org/posts/the-chip-security-act-a-bipartisan-solution-to-chip-smuggling>.

⁷¹ “4 Geolocation Technologies Compared,” Sensolus, July 23, 2021, <https://www.sensolus.com/knowledge-hub/behind-the-tech/four-geolocation-technologies-compared-how-can-they-improve-your-operational-efficiency/>; Emma Rivero and Ed Sperling, “Location Verification Becomes Much Bigger Concern for Chips,” *Semiconductor Engineering*, October 9, 2025, <https://semiengineering.com/location-verification-becomes-much-bigger-concern-for-chips/?cmid=73314953-d2d8-4d9d-ae4c-a76a2b245cce>.

⁷² NVIDIA Fleet Intelligence allows a user to monitor their chips as well as export a signed report detailing information about those chips. Fleet Intelligence, which is built on NVIDIA’s Data Center GPU Manager (DCGM) and NVIDIA Device Identity Services (NDIS), determines whether a connected chip is authentic by comparing a cryptographic key—derived from the chip’s unique, hardware-fused identity key—with NVIDIA’s own registry of valid keys. As such, a Fleet Intelligence report could list verified product serial numbers or count numbers of verified, controlled chips by product type (SKU). However, while this method could somewhat reliably confirm the operator of a specified export-controlled chip, its method of location verification (i.e., network-based telemetry data) is less robust; IP addresses can be easily spoofed. Christian Shrauder and Gail Frederick, “Introducing NVIDIA Fleet Intelligence”; “Opt-In NVIDIA Software Enables Data Center Fleet Management”; “Get inventory report”; “NVIDIA DCGM”; “Field Identifiers”.

⁷³ Brass and Aarne, “Location Verification for AI Chips.”

⁷⁴ A GPU is securely booted by using a hardware root of trust, which is “an inherently trusted combination of hardware and firmware that maintains the integrity of information.” “Hardware Root of Trust,” Computer Security Resource Center Glossary, National Institute of Standards and Technology, accessed July 8, 2026, https://csrc.nist.gov/glossary/term/hardware_root_of_trust.

⁷⁵ “Attestation Types and Scenarios,” Microsoft Learn, accessed August 18, 2026, <https://learn.microsoft.com/en-us/azure/confidential-computing/attestation-solutions> [<https://perma.cc/JQ94-DMV2>]; Jämes Ménétre, Christian Göttel, Anum Khurshid, Marcelo Pasin, Pascal Felber, Valerio Schiavoni, and Shahid Raza, “Attestation Mechanisms for Trusted Execution Environments Demystified,” arXiv preprint arXiv:2206.03780 (2022), <https://arxiv.org/pdf/2206.03780>.

⁷⁶ Jianwei Zhu, Hang Yin, Peng Deng, Aline Almeida, and Shunfan Zhou, “Confidential Computing on NVIDIA Hopper GPUs: A Performance Benchmark Study,” arXiv preprint arXiv:2409.03992 (2024), <https://arxiv.org/abs/2409.03992>; Marcin Chrapek, Marcin Copik, Etienne Mettaz, and Torsten Hoefler, “Confidential LLM Inference: Performance and Cost Across CPU and GPU TEEs,” arXiv preprint arXiv:2509.18886 (2025), <https://arxiv.org/abs/2509.18886>.

⁷⁷ “Confidential computing is the protection of data in use by performing computation in a hardware-based, attested trusted execution environment (TEE).” Emily Apsey, Phil Rogers, Michael O’Connor, and Rob Nertney, “Confidential Computing on H100 GPUs for Secure and Trustworthy AI,” NVIDIA Developer Blog, August 3, 2023, <https://developer.nvidia.com/blog/confidential-computing-on-h100-gpus-for-secure-and-trustworthy-ai>.

⁷⁸ Specifically, the software running in the user-accessible part of the GPU TEE would be a CUDA kernel.

⁷⁹ Attestation is providing cryptographic proof that a hardware or software component is authentic, untampered, and running in a trusted state; it is “the evidence that you use to evaluate whether or not to trust a Confidential Computing program or environment.” Attesting a modern NVIDIA data center GPU involves comparing a cryptographic key—derived from the chip’s unique, hardware-fused identity key—with NVIDIA’s own registry of valid keys. Alec Fernandez, “Why Is Attestation Required for Confidential Computing?,” Confidential Computing Consortium, April 6, 2023, <https://confidentialcomputing.io/2023/04/06/why-is-attestation-required-for-confidential-computing/>; “Hardware Root of Trust,” Computer Security Resource Center Glossary, National Institute of Standards and Technology, accessed July 8, 2026, https://csrc.nist.gov/glossary/term/hardware_root_of_trust.

⁸⁰ Users can run code in the TEE but cannot modify the structure of the TEE.

⁸¹ We assume that every chip with a TEE will also have a hardware identity key, which can be used to attest the AI chip’s authenticity and integrity. This is at least the case for each NVIDIA GPU that contains a TEE. Zhongshu Gu et al., “Blueprint, Bootstrap, and Bridge: A Security Look at NVIDIA GPU Confidential Computing,” arXiv preprint arXiv:2507.02770v2 (2026), <https://arxiv.org/abs/2507.02770v2> [<https://perma.cc/ZA66-FD8W>].

To enable firmware-based signing on a chip that already has a TEE, NVIDIA, for example, would need to modify its GPU firmware to expose a new application programming interface (API) that can (a) authenticate and respond to the RTT challenge and (b) sign the response using the private hardware identity key physically fused into the chip during fabrication. Zhongshu Gu et al., “Blueprint, Bootstrap, and Bridge.” In addition to this firmware change, the GPU designer would need to provide a modified kernel driver, which users would need to install on the host CPU.

NVIDIA’s current TEE firmware does expose an API for attesting that the GPU (a) is an authentic device (b) with valid firmware and (c) was booted securely. Therefore, it is possible that a PLV system with firmware-based signing could be configured for NVIDIA GPUs without requiring the chip designer to modify its GPU firmware. Specifically, the nonce in NVIDIA’s existing attestation protocol could potentially be repurposed as an RTT response. “GPU Claims,” NVIDIA Attestation Advanced Documentation, NVIDIA, June 5, 2026, https://docs.nvidia.com/attestation/advanced-documentation/latest/claims-guide/gpu_claims.html [<https://perma.cc/DY4A-D96X>]. If this is possible, it would be less secure overall because the process of *generating* a response to the RTT challenge would still occur in software. Furthermore, this option would only be possible for NVIDIA’s Hopper generation GPUs and later, since the attestation protocol was added at the same time as the TEE, as part of NVIDIA’s confidential computing (CC) feature set. “NDIS Device Root Certificate,” NVIDIA Device Identity Services (NDIS) 1.0 Documentation, NVIDIA, accessed July 8, 2026, <https://docs.ndis.nvidia.com/NDIS%20Certificate%20Chains/NDIS%20Device%20Identity.html> [<https://perma.cc/DM4F-Y7A2>].

⁸² Note that this would still require customers to install a firmware update (and corresponding software update). There must also be enough memory space in the AI chip to install the firmware update, but we do not speculate on whether or not this would be an issue.

⁸³ Physical unclonable functions (PUFs) would be added to the AI chip's hardware. PUFs are objects that provide physically defined "digital fingerprint" outputs that serve as unique identifiers. Vincent van der Leest, "What Is a Physical Unclonable Function?" Synopsys, accessed July 8, 2026, <https://www.synopsys.com/glossary/what-is-a-physical-unclonable-function.html>.

⁸⁴ Many data center servers include 8x AI chips and 2x CPUs. The CPUs control and orchestrate the AI chips.

⁸⁵ The system could instead be configured to have RTT challenge response signing take place on the host CPU. For the least secure implementation (i.e., software-based signing), this would mean signing all of the challenge responses in the CPU's confidential computing environment—with minimal involvement from the AI chip, beyond producing a regular AI chip attestation report that would be transmitted to landmark servers along with the RTT challenge response. For the firmware and hardware implementations, this would require the chip designer to extend the AI chip's hardware root of trust to the CPU (because these implementations are based on challenge response signing with hardware root of trust from the GPUs). However, for all proposed implementations, performing the challenge response signing on the CPU would reduce the security of the system by expanding the attack surface (e.g., exposing the process of generating and signing the RTT challenge response to the GPU-CPU interface).

⁸⁶ The landmark server CPUs would ideally not initiate contact with the AI chips' host CPUs, because doing so would expose the securely booted AI chips' host CPUs' network interfaces to the open internet and expand the attack surface that adversaries could exploit. Instead, the host CPU of each AI chip should initiate secure mTLS tunnels to the landmark servers, and then the landmark servers would send the RTT challenge through the tunnel; the AI chip's response would also be communicated back through the same tunnel.

⁸⁷ NIST already works closely with standards-setting organizations such as the Internet Engineering Task Force (IETF). Information Technology Laboratory, "Standards Activities," National Institute of Standards & Technology, accessed July 10, 2026, <https://www.nist.gov/itl/standards-activities-0> [<https://perma.cc/B23R-LP8K>].

⁸⁸ Worst-case accuracy is not guaranteed for data center AI chips located in bodies of water, Antarctica, the United States, China, and Russia (i.e., locations where landmark servers likely would not be placed). This does not undermine the value of the system, however, because if an AI chip's feasible region is larger than expected for the chip's expected location and overlaps with China, for example, it should indicate to the responsible entity that diversion may be occurring and further investigation may be necessary.

Furthermore, worst-case accuracy is not, by default, guaranteed for AI chips installed in a remote region (i.e., more than 391 km from an existing data center). BIS could address this limitation by requiring new, remote data centers to host on-site landmark servers. See subsection "Positioning Landmark Servers."

⁸⁹ See subsection "Positioning Landmark Servers" in Appendix B.

⁹⁰ Asher Brass, "Location Verification for AI Chips" (Institute for AI Policy and Strategy, May 16, 2025), <https://www.iaps.ai/research/location-verification-for-ai-chips>; Brass and Aarne, "Location Verification for AI Chips."

⁹¹ Ping-based location verification demonstration website, accessed July 10, 2026, <https://ping-location.info/>.

⁹² Zhaorui Ma, Xinhao Hu, Shicheng Zhang, Na Li et al., “SubvectorS_Geo: A Neural-Network-Based IPv6 Geolocation Algorithm,” *Applied Sciences* 13, no. 2 (January 2023): 754, <https://doi.org/10.3390/app13020754>.

⁹³ Katharina Kohls and Claudia Diaz, “VerLoc: Verifiable Localization in Decentralized Systems,” in *Proceedings of the 31st USENIX Security Symposium* (2022), <https://www.usenix.org/conference/usenixsecurity22/presentation/kohls>; Katharina Kohls and Claudia Diaz, “VerLoc: Verifiable Localization in Decentralized Systems,” arXiv preprint arXiv:2105.11928 (2021), <https://arxiv.org/abs/2105.11928>.

⁹⁴ Bernard Wong, Ivan Stoyanov, and Emin Gün Sirer, “Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts,” in *Proceedings of the 4th USENIX Symposium on Networked Systems Design and Implementation* (2007), <https://www.usenix.org/conference/nsdi-07/presentation/octant-comprehensive-framework-geolocalization-internet-hosts> [https://www.usenix.org/legacy/event/nsdi07/tech/full_papers/wong/wong.pdf].

⁹⁵ TBG combines delay measurements with information on network topology. Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe, “Towards IP Geolocation Using Delay and Topology Measurements,” in *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement* (2006), <https://dl.acm.org/doi/10.1145/1177080.1177090>.

⁹⁶ We use the speed of light through a vacuum because it is a safer upper bound than the speed of light through fiber; RTT requests and responses could, in theory, make use of satellite or out-of-band radio communication, which would exceed the speed of light through fiber. However, since the maximum speed of light through fiber is about two-thirds that of light through a vacuum, using the lower value as the maximum velocity would shrink the upper-bound radius for a given landmark server by approximately 33% (i.e., one-third). For further discussion of the trade-offs, see page 18 of the following paper: Brass and Arne, “Location Verification for AI Chips.”

⁹⁷ Kevin Miller, “Calculating Optical Fiber Latency,” M2Optics, January 9, 2012, <https://www.m2optics.com/blog/bid/70587/calculating-optical-fiber-latency> [<https://perma.cc/HQ2N-48L4>]; “WAN Latency Calculator,” WintelGuy.com, accessed July 8, 2026, <https://wintelguy.com/wanlat.html>.

⁹⁸ According to the WAN Latency Calculator, 1 ms is a common estimate for one-way equipment latency; we double that figure to arrive at 2 ms. When the true distance is 1 km, we assume a worst-case local loop of 15 km, because the AI chip and landmark server are already so close together. When the true distance is 5,000 km, we assume a worst-case local loop of 160 km, which is double the 80 km upper bound reported by WAN Latency Calculator. “WAN Latency Calculator,” WintelGuy.com, accessed July 8, 2026, <https://wintelguy.com/wanlat.html>.

⁹⁹ According to the WAN Latency Calculator, 1 ms is a common estimate for one-way equipment latency; we take one-tenth of that figure to arrive at 0.1 ms. “WAN Latency Calculator,” WintelGuy.com, accessed July 8, 2026, <https://wintelguy.com/wanlat.html>.

¹⁰⁰ See subsection “When Does a Second Landmark Server Matter?” A landmark server positioned 1 km from the AI chip could only be improved by another landmark server no more than 5 km from the chip (in the opposite direction). As discussed in the subsection, it is unlikely that landmark servers would be positioned within 100 km of each other.

¹⁰¹ According to the WAN Latency Calculator, 1 ms is a common estimate for one-way equipment latency; we take one-tenth of that figure to arrive at a best-case of 0.1 ms. “WAN Latency Calculator,” WintelGuy.com, accessed July 8, 2026, <https://wintelguy.com/wanlat.html>. Using Equation 1.7: network processing latency is 0.2 ms. Using Equation 1.6: $0.2 \text{ ms} \times 149.896 \text{ km/ms} = 30 \text{ km}$.

¹⁰² Using Equation 1.8: [change in flight time] = $2 \times \frac{1 \text{ km}}{200 \text{ km/ms}} = 0.01 \text{ ms}$. Using Equation 1.6: $0.01 \text{ ms} \times 149.896 \text{ km/ms} = 1.5 \text{ km}$.

¹⁰³ Since we exclude bodies of water, Antarctica, the United States, China, and Russia, the specified worst-case accuracy is not guaranteed for data center AI chips in those locations. This does not undermine the value of the system, however, because if an AI chip’s feasible region is larger than expected for the chip’s expected location and overlaps with China, for example, it should indicate to the responsible entity that diversion may be occurring and further investigation may be necessary. Furthermore, worst-case accuracy is not, by default, guaranteed for AI chips installed in a remote region (i.e., more than 391 km from an existing data center). BIS could address this limitation by requiring new, remote data centers to host on-site landmark servers. See subsection “Positioning Landmark Servers.”

¹⁰⁴ Eric W. Weisstein, “Spherical Cap,” Wolfram MathWorld, accessed July 8, 2026, <https://mathworld.wolfram.com/SphericalCap.html> [<https://perma.cc/QAQ9-D49T>]; Gábor Fejes Tóth, “Packing and Covering,” in *Handbook of Discrete and Computational Geometry*, 3rd ed. (Boca Raton, FL: CRC Press, 2017), <https://www.csun.edu/~ctoth/Handbook/chap2.pdf>.

¹⁰⁵ Richard Kershner, “The Number of Circles Covering a Set,” *American Journal of Mathematics* 61, no. 3 (July 1939): 665–671, <https://www.jstor.org/stable/2371320>; W. Kuperberg, “Covering the Plane with Congruent Copies of a Convex Body,” *Bulletin of the London Mathematical Society* 21, no. 1 (January 1989): 82–86, <https://londmathsoc.onlinelibrary.wiley.com/doi/10.1112/blms/21.1.82>; T. Tarnai and Zs. Gáspár, “Covering a sphere by equal circles, and the rigidity of its graph,” *Mathematical Proceedings of the Cambridge Philosophical Society* 110, no. 1 (July 1991): 71–89.

¹⁰⁶ Even if doubling θ is not sufficient, it should not matter because this worst-case estimate is overly conservative. In remote regions where data centers are sparse, the landmark server operator could not add more servers to improve accuracy. By default, AI chips located in those remote regions would not be guaranteed the worst-case accuracy specified for a fixed number of 250 landmark servers or more (see Table B4). Functionally, remote regions could be subtracted out of the land mass that needs to be covered with landmark server caps (but we can’t do so because we don’t know how much land mass remote regions take up). Therefore, some share of the 250 landmark servers would not be needed unless new data centers with AI chips were built in those remote regions. In short, this formulation of

the problem overestimates the number of landmark servers required to achieve any given accuracy better than some threshold.

BIS could address poor PLV accuracy at new, remote data centers by requiring on-site landmark servers at those facilities. If BIS receives a license application to export an AI chip to a data center in a remote location not covered by a landmark server cap, the Bureau could require the owner to permit installation and maintenance of a landmark server on-site. See subsection “Positioning Landmark Servers.”

¹⁰⁷ In Equation 2.5, we multiply the covering factor (θ) by two to account, in part, for the irregular shapes of land masses left after removing the oceans, Antarctica, China, Russia, and the United States.

¹⁰⁸ Central Intelligence Agency, “World,” in *The World Factbook 2004* (Washington, DC: Central Intelligence Agency, 2004), <https://user.iiasa.ac.at/~marek/fbook/04/print/xx.html>; Lori Perkins, “Compare the Size of Antarctica to the Continental United States,” NASA Scientific Visualization Studio, accessed July 8, 2026, <https://svs.gsfc.nasa.gov/3540/>; “List of the Total Areas of the World’s Countries, Dependencies, and Territories,” *Encyclopaedia Britannica*, accessed July 8, 2026, <https://www.britannica.com/topic/list-of-the-total-areas-of-the-worlds-countries-dependencies-and-territories-2130540>.

¹⁰⁹ We use the radius at Earth’s poles so that we underestimate the size of each cap and therefore overestimate the number of landmark servers. This is more appropriate than using the radius at Earth’s equator or the average radius since we are building worst-case estimates for PLV accuracy. “The Shape of Earth,” Geoscience Community Codes, University of South Florida, accessed July 8, 2026, https://gscommunitycodes.usf.edu/geoscicommunitycodes/public/geophysics/Gravity/earth_shape.php.

¹¹⁰ “Circumradius,” Art of Problem Solving (AoPS) Wiki, accessed September 3, 2026, <https://artofproblemsolving.com/wiki/index.php/Circumradius>.

¹¹¹ See subsection “Positioning Landmark Servers.”

¹¹² Felix Richter, “Big Three Hold Dominant Lead in Accelerating Cloud Market,” Statista, May 5, 2026, <https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/> [<https://perma.cc/PN8B-HM8C>].

¹¹³ “Regions and Availability Zones,” Amazon Web Services, accessed July 10, 2026, https://aws.amazon.com/about-aws/global-infrastructure/regions_az/ [<https://perma.cc/2HA9-TDPW>].

¹¹⁴ “Azure Regions List,” Microsoft Learn, Microsoft, May 29, 2026, <https://learn.microsoft.com/en-us/azure/reliability/regions-list?tabs=all> [<https://perma.cc/CM8B-QNE6>].

¹¹⁵ “Cloud Locations,” Google Cloud, June 30, 2026, <https://cloud.google.com/about/locations> [<https://perma.cc/XV5Q-T4Z3>].

¹¹⁶ Positioning landmark server B opposite landmark server A (with the AI chip in the middle) maximizes the distance b (i.e., distance from landmark B to the AI chip server) before landmark server B’s disk starts to bound the chip’s feasible region. For example, if landmark server B was instead positioned on a vertical line passing through the AI chip, landmark server B would need to be closer to the chip to start

bounding its feasible region compared to if landmark server B was opposite landmark server A. A toy model that demonstrates this is available on Desmos: <https://www.desmos.com/calculator/atwhdj96n9>.

In the toy model, landmark server A's disk is colored blue. An instance of landmark server B that lies opposite landmark server A is colored red. Three other potential locations for landmark B (with the same distance from the AI chip) are shown in grey. As distance b is decreased, the edge of the red disk intersects the edge of the blue disk before the edge of any grey disk does the same.

In the toy model, $r_B = 1.5b$ represents that increasing the distance of a landmark server from the AI chip by 1 km increases disk radius by 1.5 km: using Equation 1.8: $[\text{change in flight time}] = 2 \times \frac{1 \text{ km}}{200 \text{ km/ms}} = 0.01 \text{ ms}$. Using Equation 1.6: $0.01 \text{ ms} \times 149.896 \text{ km/ms} = 1.5 \text{ km}$.

¹¹⁷ Larger cities may contain more than one data center hub.

¹¹⁸ Cuba is the 100th-largest country by land mass, with 109,884 km² (42,426 mi²). As a circle, Cuba's land mass would correspond to a radius of 187 km (116 miles). "List of Countries and Dependencies by Area," Wikipedia, accessed July 8, 2026, https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area; the same list is available in this Google Sheets workbook as well: <https://docs.google.com/spreadsheets/d/1KLeamlwrhVG1LF9kGBQv8co9CdufoNVLzmSr2sz3b5jk/edit?gid=0#gid=0>. Still, a circle of that size would not be small enough to determine whether an AI chip is located in a small country (e.g., Liechtenstein) or its neighboring countries. Nor would it likely be able to conclusively identify the country where an AI chip is located when the chip is close to the border between two large countries.

¹¹⁹ For example, data centers located in Frankfurt, Germany and Dublin, Ireland are almost entirely enclosed by a single 30-km circle in each city (see [map](#)). "Frankfurt Data Centers," Data Center Map, accessed July 11, 2026, <https://www.datacentermap.com/germany/frankfurt/>; "Dublin Data Centers," Data Center Map, accessed July 11, 2026, <https://www.datacentermap.com/ireland/dublin/>.

Frankfurt contained 55 data-center sites as of October 2024, representing approximately 30% of all German data centers. Lennart Bolwin et al., "Rechenzentren in Frankfurt am Main und der Region: Standort- und regionalökonomische Wirkungszusammenhänge," *IW Consult GmbH and Detecon International GmbH* (2025), https://www.frankfurt-business.net/fileadmin/user_upload/IT/RZ-Studie/mgmsum_Studie_Rechenzentren_final-ENGL.pdf.

In Ireland, 33 of the country's 36 operational data center sites and 69 of its 72 operational data-center buildings were located in the Greater Dublin Area in 2025. "The Value of Data Centres to Ireland," (KPMG, March 2026), <https://enterprise.gov.ie/en/publications/publication-files/the-value-of-data-centres-to-ireland.pdf>.

¹²⁰ Calculations are available in this Google Sheets workbook: <https://docs.google.com/spreadsheets/d/1VQth7sMuOVd9tI7eAAYHfFuSKIWI0-yll4O7fMW3qE9k/edit?gid=448061102#gid=448061102>.

¹²¹ SemiAnalysis, "Datacenter Industry Model," accessed July 8, 2026, <https://semianalysis.com/datacenter-industry-model/>; "AI Data Centers," *Epoch AI*, accessed July 8, 2026, <https://epoch.ai/data/ai-data-centers>.

¹²² This [map](https://map.datacente.rs/) shows known data centers in and near Kazakhstan and Mongolia; a cap with a radius of 391 km is drawn around each data center. Locations in Kazakhstan and Mongolia that are not covered by a cap would not be guaranteed the worst-case accuracy of 1,600 km calculated for 500 landmark servers (Table B4). Data center locations are sourced from “Datacente.rs World Map,” accessed on July 10, 2026, <https://map.datacente.rs/>.

Imagine a landmark server operator is trying to improve worst-case accuracy in Kazakhstan below 1,600 km (Table B4). After they set up landmark servers in existing data centers in each of the following five cities, there are no other existing data center locations where additional landmark servers would help: (a) Astana, Kazakhstan, (b) Semey, Kazakhstan, (c) Tashkent, Uzbekistan, (d) Bishkek, Kyrgyzstan, and (e) Baku, Azerbaijan.

¹²³ Furthermore, to ensure the new landmark server isn’t moved, the landmark server operator could use PLV to geolocate the new landmark server using other, more trusted landmark servers.

¹²⁴ NVIDIA Corporation, Annual Report (Form 10-K) for the fiscal year ended January 26, 2025 (filed February 26, 2025), Item 7, “Concentration of Revenue,” <https://www.sec.gov/Archives/edgar/data/1045810/000104581025000023/nvda-20250126.htm>.

¹²⁵ NVIDIA Corporation, Annual Report (Form 10-K) for the fiscal year ended January 26, 2025 (filed February 26, 2025), Item 7, “Concentration of Revenue,” <https://www.sec.gov/Archives/edgar/data/1045810/000104581025000023/nvda-20250126.htm>.

¹²⁶ Amanda Gerut, “Supermicro owes its rapid rise to \$4 trillion Nvidia—but China smuggling allegations and a high-profile arrest could blow up the partnership,” *Fortune*, April 7, 2026, <https://finance.yahoo.com/markets/stocks/articles/supermicro-soared-because-4-trillion-070200827.html>. “QCT RTX PRO Server Powers Enterprise AI Factory Transformation,” press release, Quanta Cloud Technology, accessed July 8, 2026, <https://www.qct.io/Press-Releases/index/PR/Server/QCT-RTX-PRO-Server-Powers-Enterprise-AI-Factory-Transformation>; Scott Martin, “How Dell Technologies Is Building the Engines of AI Factories With NVIDIA Blackwell,” NVIDIA Blog, May 27, 2025, <https://blogs.nvidia.com/blog/dell-technologies-ai-factories-blackwell/>.

¹²⁷ NVIDIA Partner Network Locator, enterprise partners located in the United States, accessed August 18, 2026, <https://marketplace.nvidia.com/en-us/enterprise/partners/?location=US&page=1&limit=15>. “For fiscal year 2025, an indirect customer which primarily purchases our products through system integrators and distributors, including through Direct Customer B, is estimated to represent 10% or more of total revenue, attributable to the Compute & Networking segment.” NVIDIA Corporation, Annual Report (Form 10-K) for the fiscal year ended January 26, 2025 (filed February 26, 2025), Item 7, “Concentration of Revenue,” <https://www.sec.gov/Archives/edgar/data/1045810/000104581025000023/nvda-20250126.htm>.

¹²⁸ “Supermicro NVIDIA GB300 NVL72,” Supermicro, accessed July 8, 2026, <https://www.supermicro.com/en/products/system/gpu/48u/srs-gb300-nvl72>.

¹²⁹ “Stock of the Week: Super Micro Computer,” XTB, October 9, 2025, <https://www.xtb.com/int/market-analysis/news-and-research/stock-of-the-week-super-micro-computer-09-10-2025>; Meta, “Meta’s

Singapore Data Centre,” February 2025, https://datacenters.atmeta.com/wp-content/uploads/2025/02/Meta_s-Singapore-Data-Center.pdf.

¹³⁰ Supermicro sold to over 1,000 customers in over 100 countries in fiscal years 2023, 2024, and 2025, though sales are concentrated in a few large customers. Four customers accounted for 10% or more of the net sales in fiscal year 2025. Super Micro Computer, Inc., Annual Report (Form 10-K) for the fiscal year ended June 30, 2025, <https://d18rn0p25nwr6d.cloudfront.net/CIK-0001375365/1a475a6e-49d9-4d51-8431-5f3647a70cf0.pdf>. “Direct customers,” like Supermicro, often purchase excess supply to preempt potential shortages or take advantage of bulk discounts.

¹³¹ Google, “Google Private AI Compute: Extending On-Device Privacy with the Power of the Cloud” (Google, 2025), https://services.google.com/fh/files/misc/private_ai_compute_technical_brief.pdf; AMD, “AMD Instinct MI355X GPU,” product brief, accessed July 8, 2026, <https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/amd-instinct-mi355x-gpu-brochure.pdf>.

¹³² “Titan Hardware Chip,” Google Cloud, January 2025, <https://docs.cloud.google.com/docs/security/titan-hardware-chip> [<https://perma.cc/7CHM-8MDR>]; “AWS Nitro System,” Amazon Web Services, accessed July 8, 2026, <https://aws.amazon.com/ec2/nitro/> [<https://perma.cc/ES25-DRDM>]; “AWS Nitro Enclaves,” Amazon Web Services, accessed July 8, 2026, <https://aws.amazon.com/ec2/nitro/nitro-enclaves/> [<https://perma.cc/YS3T-MVKB>].

¹³³ AMD, “AMD Instinct MI355X GPU”; Trusted Computing Group, “DICE Attestation Architecture,” Version 1.00, Revision 0.22, September 18, 2020, https://trustedcomputinggroup.org/wp-content/uploads/TCG_DICE_Attestation_Architecture_r22_02dec2020.pdf.

¹³⁴ Georgia Butler, “Google to Sell TPUs to a ‘Select Group of Customers’ for Their Data Centers, Increases Capex Forecast,” Data Center Dynamics, April 30, 2026, <https://www.datacenterdynamics.com/en/news/google-to-sell-tpus-to-a-select-group-of-customers-for-their-data-centers/>.

¹³⁵ Google, “Google Private AI Compute: Extending On-Device Privacy with the Power of the Cloud” (Google, 2025), https://services.google.com/fh/files/misc/private_ai_compute_technical_brief.pdf; “Titanium,” Google Cloud, accessed July 8, 2026, <https://cloud.google.com/titanium>; “Confidential VM Overview,” Google Cloud, July 7, 2026, <https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/confidential-vm-overview>.

¹³⁶ “Amazon in Talks to Sell AI Chips Competing with Nvidia to Other Companies,” *Bloomberg*, June 18, 2026, <https://www.bloomberg.com/news/articles/2026-06-18/amazon-is-in-talks-to-sell-nvidia-rival-chips-to-other-companies>.

¹³⁷ “AWS Nitro System,” Amazon Web Services, accessed July 8, 2026, <https://aws.amazon.com/ec2/nitro/> [<https://perma.cc/ES25-DRDM>]; “AWS Nitro Enclaves,” Amazon Web Services, accessed July 8, 2026, <https://aws.amazon.com/ec2/nitro/nitro-enclaves/> [<https://perma.cc/YS3T-MVKB>]; “Cryptographic Attestation,” AWS Nitro Enclaves User Guide, Amazon Web Services, accessed July 8, 2026, <https://docs.aws.amazon.com/enclaves/latest/user/set-up-attestation.html> [<https://perma.cc/2W84-CWVX>]; Amazon Web Services, “The Components of the Nitro

System,” in “The Security Design of the AWS Nitro System” (Amazon Web Services, November 2022), <https://docs.aws.amazon.com/whitepapers/latest/security-design-of-aws-nitro-system/the-components-of-the-nitro-system.html> [<https://perma.cc/MLE4-DSQZ>].

¹³⁸ “AWS Nitro System”; “AWS Nitro Enclaves.”

¹³⁹ More specifically, firmware-based signing requires that the RTT challenge response is signed by a cryptographic derivative of the AI chip’s hardware identity key.

¹⁴⁰ AMD, “AMD Instinct MI355X GPU”; Trusted Computing Group, “DICE Attestation Architecture,” Version 1.00, Revision 0.22, September 18, 2020, https://trustedcomputinggroup.org/wp-content/uploads/TCG_DICE_Attestation_Architecture_r22_02dec2020.pdf.

¹⁴¹ Brass and Aarne, “Location Verification for AI Chips.”

¹⁴² “Titan Hardware Chip,” Google Cloud, January 2025, <https://docs.cloud.google.com/docs/security/titan-hardware-chip> [<https://perma.cc/7CHM-8MDR>].

¹⁴³ “AWS Nitro System”; “AWS Nitro Enclaves.”