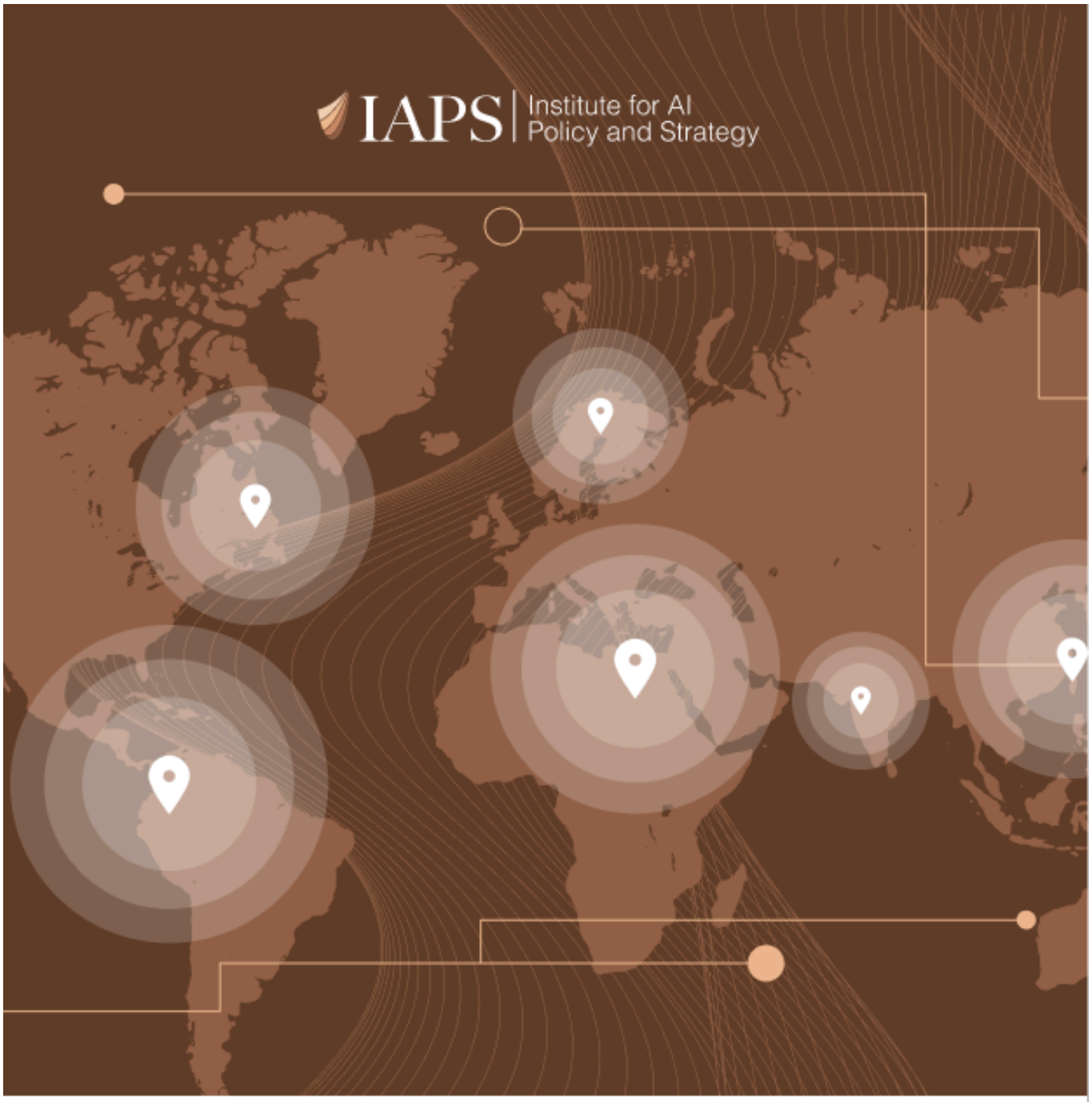


April 2024

Location Verification for AI Chips

By Asher Brass & Onni Aarne

Table of Contents

Executive Summary

Advanced, general-purpose AI models are emerging as catalysts for economic development and scientific advancement, and they might end up fundamentally altering many different fields and aspects of daily life in unexpected ways. These models are being built and deployed using cutting-edge AI chips, designed and fabricated by just a handful of companies. The AI chips in question are currently classified as export-controlled goods by the United States and other countries, who have taken steps to avoid the proliferation of this technology to actors who might misuse it in dangerous ways. However, export control officials and regulators at-large have no mechanism by which they can accurately and reliably receive information regarding the actual deployment location of these chips.

Adding location verification features to AI chips could unlock new governance mechanisms for regulators, help enforce existing and future export controls by deterring and catching smuggling attempts, and enable post-sale verification of chip locations.

Delay-based schemes appear to be the most promising approach to location verification. Combining such a scheme with additional governance mechanisms, such as a centralized chip registry, would provide regulators with more effective tools to verify that chips are not deployed in restricted locations.

Compared to more commonly utilized methods like GPS, delay-based location verification would be difficult to "spoof," i.e., falsify. While delay-based location verification is much less precise than GPS, it is still sufficiently precise to be useful for export control enforcement.

This paper is meant to serve as an initial introduction to location verification use-cases for AI chips, comparing different methods that could be utilized for location verification and briefly discussing both obstacles and requirements for creating a secure version of this solution.

Our main finding is that it seems both feasible and relatively cheap to implement pure-software delay-based solutions on chips in the near future. These solutions are likely to aid current AI chip export control enforcement efforts. A solution of this kind would likely cost less than $1,000,000 to set up and maintain for several years. Chip design companies should consider taking the initiative by experimenting with this technology on chips that are at a particularly high risk of diversion. More proactive use of technology to prevent the diversion of chips could bolster international trade by reducing the need for broad country-level export bans or license requirements.

Detailed Summary

Location verification refers to the process by which the location of an asset, such as an AI chip[1], can be verified via cryptographic or other verifiably secure mechanisms. The goal is to assure the verifier that the asset's reported location is genuine and hasn't been manipulated or spoofed. **Location verification is a potentially useful feature to add to AI chips for governance purposes.**

Regulators could use location information to focus investigative resources and locate stolen or missing chips (more below). In the future, pending the introduction of on-chip governance mechanisms (Aarne et al., 2024), regulators could also take action regarding chips that are misbehaving or are in restricted locations via remotely locking down chips or throttling certain capabilities.

Location verification of AI chips requires geolocation methods that are not only reliable and fairly accurate, but also secure from adversaries who may have a strong motivation to attempt to provide false or misleading location information for chips in their possession[2].

It is important to note that the mechanisms this report discusses would not allow any kind of invasive, unilateral monitoring. Instead, these mechanisms would give chip owners the ability to make verifiable claims about the approximate location of their chips.

In essence, this paper discusses two main problems that may face regulators interested in implementing a location verification solution:

1. The **General Geolocation Problem**, meaning how do we satisfy all of the use-cases described while reliably and securely geolocating chips without any particular additional considerations? There are several unresolved issues that require further work in order to set up such a scheme (more below).
2. The **Anti-Smuggling Geolocation Problem**, a solution to which only satisfies the narrower use-case of verifying that chips are not in restricted countries and is probably worth adding to chips as soon as possible (more below).

More generally, location verification could allow regulators[3] to verify that a given chip is in a particular location[4], verify that it is not in a restricted location[5], verify the approximate location of the

[1] For the purposes of this report, "asset" is used to refer to a device, such as a GPU chip, graphics card, or server.
[2] Such motivations may include not getting caught smuggling chips, avoiding remote shut-down by regulators, evading taxes to internalize safety externalities, or evading other forms of additional scrutiny.
[3] The term "regulator," in this report, refers to parties who have an interest in verifying or regulating the use and/or location of AI chips.
[4] E.g., that it matches some registry of chip locations.
[5] E.g., in violation of export controls or contradicting public claims about the location of chips.

chip[6], identify post-deployment changes in the location of the chip[7], and uncover suspicious behavior[8] (more below).

Location Verification Could Enable Regulators To:
Verify that a given chip is not in a specific location or set of locations
Verify that a given chip is in a specific known location or set of locations
Geolocate a given chip without knowing ahead of time where it might be
Discover chips that are behaving suspiciously with regards to their location reporting
Discover chips that have likely been moved to another location post initial deployment

A central component of location verification is the geolocation method used in order to determine where the chip is located at any given time. These methods can be divided into three broad categories (more below):

1. **Asset-reported**: The asset utilizes external signals to determine its own location and then sends that location to a verifier. GPS is a prominent example of this type of location verification, which can provide extremely precise positioning information (~3m-~100m in most cases).

2. **Topology-based**: The asset reports certain measurements of radio-frequency emitters that are within range and widely deployed in the target region in known locations–such as cell towers, wifi routers, or similar radio frequency (RF) beacons. These measurements and the known locations of the emitters are then used to calculate the location of the asset. Network-based mobile phone tracking is a prominent example of this type of location verification, which can provide very accurate positioning information (~50m-~500m in most urban areas).

3. **Delay-based**: The asset sends a "ping"–or some other short transmission that includes the current time–to a series of servers or devices in known locations[9]. The location of the asset is then calculated by estimating the possible distance the transmission could have traveled in the elapsed time and then combining the results using various algorithms. These

[6] Without reference to registries, estimated location, or any other forms of evidence.
[7] I.e., that the chip used to be in one geographic location and has since been moved away to a different location.
[8] E.g., a chip reporting a location that is infeasible like the middle of the ocean, indicating that someone is probably manipulating the reported location.
[9] The communication could be initiated by either the asset or the landmark server. This is an implementation detail that could have some security implications, but for now we can simplify and simply state that the asset is the one sending the ping initially.

methods are less well-known and are primarily useful for cloud computing scenarios, where they can provide broad positioning information (~10km-~1,000km, depending on the exact algorithm used and other considerations).

Method	Description	Advantages	Limitations
Asset-Reported	Involves the asset determining its own location via GPS or similar technologies and reporting it to a trusted server.	• Commonly used, well-understood technology. • Can be very accurate: (~3m) under optimal conditions.	• GPS can be easily spoofed. • Requires unobstructed signals from satellites, which may not be reliable indoors or in data centers.
Topology-Based	Maps digital characteristics of the asset (e.g., IP address, wifi SSIDs) to known locations via the internet or other public infrastructures.	• Utilizes existing digital infrastructure. • Can be quite accurate (~50m-~500m), given optimal conditions.	• Relies on unreliable and insecure protocols and infrastructure. • Typically requires adding (at minimum) an antenna to the die.
Delay-Based	Uses a cryptographic challenge-response between the asset and landmarks to calculate location based on communication travel time (RTT).	• Overall system security to spoofing and other manipulation techniques. • Probably does not require adding new hardware to chips.	• Difficult to achieve precise geolocations. • Requires excellent distribution and coverage of landmarks–also requires setting up and maintaining a network of secure landmark servers.

Asset-reported and topology-based methods have well-established track records of insecurity and manipulation by adversarial actors. **GPS, for example, can be trivially spoofed for as little as ~$200. Therefore, it is not recommended that these methods be relied upon exclusively in order to perform location verification for advanced AI chips.**

In short, current asset-based and topology-based methods are probably sufficiently secure in cases where the potential adversary is very worried about being caught[10], but not for cases where an adversary has a moderate or high incentive to cheat even if they might get caught (more below).

Delay-based methods appear to have the most relevant security features that could be useful for preventing adversarial obfuscation–such as their encrypted communication and their relative lack of reliance on public infrastructure. However, these methods are still susceptible to manipulation by moderately and highly advanced technological adversaries (more below).

For instance, utilizing dark fiber could allow adversaries to consistently achieve faster communication speeds than expected due to reduced delays from congestion, routing, etc.,because the speed of communication in dark fiber is typically closer to the speed of light in fiber, without additional delays caused by congestion and routing.

Nevertheless, there appears to be significant value in implementing delay-based location verification on AI chips, which would probably dissuade most adversaries from even attempting to spoof or obfuscate the location of the chip and would enable regulatory actions for misbehavior or being caught cheating for adversaries who are not dissuaded (more below).

Combining location verification with a centralized chip-registry[11] would make location spoofing at scale substantially harder and potentially more expensive for adversaries, to a degree that probably makes it impractical for most non nation-states to even attempt to circumvent restrictions.

[D]elay-based location verification could, with relatively little other investment, be invaluable in helping authorities both identify illicit chip smuggling activity and focus their resources more efficiently.

However, creating a landmark[12] deployment map[13] is not a trivial issue, and many tactical obstacles need to be overcome. Several of these obstacles are discussed, but a precise solution is out-of-scope for the current paper, and any such system would need to be periodically revised and updated as global communication patterns and trends change.

[10] For example, an international company that is strictly supervised by a government or regulatory body, or heavily relies on its good standing in the eyes of regulators.
[11] As suggested by Shavit, 2023 and Baker, 2023.
[12] "Landmarks" here refers to servers or other assets that are geographically dispersed in known locations, can communicate with the asset, and can communicate with each other or with a back-end server in order to perform location calculations. These are typically private servers owned by cloud providers or other large telecommunication firms located in data centers, but that is not a requirement for the geolocation function per se.
[13] Meaning a coverage scheme that utilizes the least amount of landmarks while allowing regulators to glean useful insights regarding the location of the chips.

Specifically, as an enforcement and verification mechanism for anti-smuggling in the context of export controls, delay-based location verification could, with relatively little other investment, be invaluable in helping authorities both identify illicit chip smuggling activity and focus their resources more efficiently (more below).

An illustrative case study is presented based on real-world ping data[14], meant to show how delay/RR-based location verification could help regulators determine, in practice, whether chips can be said to be located where their owners claim they are located or not (more below). Finally, a brief description of the solution requirements is also sketched, including both on-chip modifications and server specifications (more below).

This paper is meant to serve as an initial introduction to location verification use-cases for AI chips, comparing different methods that could be utilized for location verification and briefly discussing both obstacles and requirements for creating a secure version of this solution. As such, there are still several open questions left to solve before a solution of this nature is viable, and many of these are included in the "Future Research Directions" section.

Location verification and potential use-cases for AI Governance

Remote Attestation is a security concept primarily used in the context of trusted computing. It allows a device (or a computing environment) to prove to a remote party that it is in a certain state, running specific software, or other internal details without revealing more information than necessary. This information is shared in such a way that the remote party can trust the authenticity and integrity of the information received.

Central Equipment Identity Registers (CEIR) are an example of a ubiquitous remote attestation technique: every mobile phone has a globally unique identification number (an IMEI), which is transmitted in an encrypted fashion to the mobile network upon registration. The IMEI is compared to a blocklist on the CEIR, and if it has previously been reported as stolen, the network registration attempt will be rejected[15].

Location verification is a specific kind of remote-attestation that could be a particularly useful feature to add to cutting-edge AI chips. Successful implementation of location verification mechanisms could enable regulators to:

1. **Verify that a given chip is not in a specific location or set of locations** (more below).

[14] Generously provided by www.wondernetwork.com

[15] Digital cellular telecommunications system; International Mobile station Equipment Identities (IMEI) (GSM 02.16)

a. Assist regulators in making sure that chips that are export-controlled are not making their way to restricted countries or organizations via smuggling or other illicit methods (Grunewald & Aird, 2023).

b. Allow regulators to identify bad actors in supply chains for further investigation, such as shipping companies, shell companies, fronts, and so on.

2. **Verify that a given chip[16] is in a specific known location or set of locations.**

    a. Combined with a centralized chip registry[17]–as suggested by Shavit, 2023, and Baker, 2023–this could allow regulators to make sure chips are indeed where they are supposed to be and take action if that is not the case.

3. **Geolocate a given chip without knowing ahead of time where it might be** (more below).

    a. Regulators could detect large clusters of chips active within a given geographic region, which might indicate that they are being used in order to train AI models. For example, the October 2023 Executive Order introduces reporting requirements for clusters with a theoretical maximum computing capacity of 10^{20} FLOP/s. This method could also allow for a ballpark estimate of the magnitude of the model being trained.

4. **Discover chips that are behaving suspiciously with regard to their location reporting.** E.g., chips that are not reporting their location at all, that previously reported their location but have ceased doing so, for which a location cannot be accurately estimated, or attest to locations that seem infeasible[18].

5. **Discover chips that have likely been moved to another location post initial deployment.**[19] If a given chip has established a baseline of *n* time to communicate with the landmark and then *n* suddenly grows or shrinks, this could indicate that the chip has been moved, possibly in an illicit or covert manner. Obviously, this sort of result could also happen due to data center connectivity changes, global interconnection shifts, and so on–but it could at least be an indication that regulators should investigate further.

Location Verification Could Enable Regulators To:

[16] Regulators would probably prefer to monitor groups of chips with common features (purchased by the same end-user, shipped via the same company, etc.) rather than individual chips–the term "given chip" is used for simplicity.
[17] Broadly, a chip registry might contain a unique identifier for each chip (e.g., serial number), the identity of the company that purchased the chip, the deployment location or country of the chip, and other relevant meta-data.
[18] Like in the middle of the Pacific Ocean, for example.
[19] Credit and thanks to Lennart Heim for this idea.

Verify that a given chip is not in a specific location or set of locations
Verify that a given chip is in a specific known location or set of locations
Geolocate a given chip without knowing ahead of time where it might be
Discover chips which are behaving suspiciously with regards to their location reporting
Discover chips which have likely been moved to another location post initial deployment

Given the information provided by location verification, if chips are not located where they are supposed to be, if they are located in restricted regions, or if their location cannot be accurately determined, etc., regulators could then take action, for instance, by:

1. Utilizing "region locking" features, which completely disable the function of the chip if its location cannot be verified.[20] These features are commonly used to restrict software and hardware from being deployed or accessed from restricted locations in order to comply with local or international regulations. Printer cartridges, for example, are commonly restricted to specific geographic regions and will not function unless they match the printer's region.[21]
2. Alternatively, this "region locking" could simply limit certain capabilities on the chip (e.g., reduce interconnect bandwidth, as suggested by Kulp et al., 2023), pending further clarification from the chip owner that they are engaging in legitimate activity.
3. Focusing investigative resources and efforts on chips that report suspicious location information.
4. Tracking down stolen or missing chips that are currently in use and might be in the possession of unauthorized actors either domestically or internationally.

[20] The implementation details of "region locking" on chips are out of scope for the current paper. See discussion of "location verification" and "operating licenses" in Aarne et al. (2024).
[21] Hey, did you know most inkjet printers are region-locked?

Solution requirements, threat models, and additional considerations

Level of precision required: For some of the above-stated purposes, a country-level geolocation might be good enough (i.e., a geolocation resolution that is precise enough to determine the general country in which a particular point or area lies but not necessarily detailed enough to specify finer details such as the exact city, region, or address).

However, there seem to be clear advantages in improving the geolocation resolution as much as possible[22]. For example, detecting large clusters would be impractical if the geolocation in question is larger than several kilometers. Population centers that are near country borders could also be a significant issue, depending on the size and accuracy of the geolocation and other geographic factors.[23]

Integration in production process: The location verification solution would need to be added to the chips either during the fabrication process (if it requires a physical component, such as a secure chip on die, a GPS receiver, or any similar antenna) or as an additional pure software component prior to final sale. Neither of these options appears like it would be infeasible or particularly onerous to chip manufacturers, although manufacturers probably prefer not to add additional hardware into the already miniscule chip space.

Threat model: Three classes of adversaries[24] seem relevant when discussing the security of the feature, borrowing the adversary classification terminology used by Aarne, Fist & Withers 2024:

1. Minimally adversarial: "*attackers do not spend much on attacks, and are very averse to being discovered attempting to compromise mechanisms,*" e.g., U.S. tech companies.
2. Covertly adversarial: *"attackers are more willing to spend substantial resources to compromise mechanisms, but still want to avoid being caught doing so,"* e.g., international shell companies or companies with suspicious ties to adversarial governments, such as Huawei[25] or VK[26].

[22] To be fair, there may also be some drawbacks to granular location resolutions, particularly around privacy and data sensitivity issues.
[23] Illustrative examples include Gdansk, Poland <-> Kaliningrad, Russian (~100 km), Qingdao, China <-> Seoul, South Korea (~600 km), and Singapore <-> Kuala Lumpur, Malaysia (~300 km).
[24] Meaning actors who might want to attempt to circumvent geographic restrictions or insulate themselves from possible chip inspections.
[25] Addition of Entities to the Entity List—Huawei
[26] U.S. Treasury Imposes Immediate Economic Costs in Response to Actions in the Donetsk and Luhansk Regions

3. Openly adversarial: *"attackers are willing to spend very significant resources to compromise mechanisms, and are indifferent to this being discovered"* e.g., rogue nation-states or associated intelligence agencies.

Overview of Adversary Categories			
Adversary Category	**Key Properties**	**Protections Required**	**Example Applications**
Minimally Adversarial	Low resources, strongly prefers not to be discovered	Basic security measures	Domestic regulation
Covertly Adversarial	Moderate to high resources, prefers not be discovered	Exceptionally secure software and/or hardware	Export control enforcement against large international companies
Openly Adversarial	High resources, does not mind being discovered	Provably secure software and/or hardware	More challenging cases of export control enforcement and treaty verification, where other deterrence fails

For minimally adversarial contexts, insecure but accurate location verification methods will probably suffice due to the threat of punishment if circumvention attempts are detected. Furthermore, these actors are likely to only gain a limited amount from circumvention, and lack access to specialized resources[27] which would be required for successful circumvention. For covert and openly adversarial contexts, the location verification solution needs to be not only accurate but also secure.

Location verification systems could be attacked by either *blocking* or *spoofing* location reporting. Adversaries of these kinds could attempt to prevent the chip from reporting its current location or attempt to block communication between the chip and regulators. Alternatively, adversaries might attempt to spoof the location of the chip, so the chip appears to be either in another specific "known" location. For example, a Russian company might deploy a chip in the Kaliningrad exclave and spoof the location to be ~100km away in Gdansk, Poland, or in an undetermined adjacent location. For example, a Chinese company might deploy a chip in China but tamper with location parameters such that the chip appears to be somewhere outside of China—in this example, the

[27] Vulnerability researchers, cyber-physical malicious infrastructure, location protocol experts, etc.

company might not care where the geolocation lands as long as it is outside China[28] or they might be unable to accurately calculate the resulting location ahead of time.

A brief survey of modern location verification methods

The problem of geolocating remote assets[29] is not unique to AI governance. A majority of public academic and scientific contributions that deal with this issue do so for the purposes of tracking and verifying the geographic location of cloud data (Jiang et al., 2021) or for supply-chain asset tracking (Ahmed et al., 2020). Notably, academic work on these topics mostly does not address potential security concerns that might be utilized to obfuscate chip locations.

Geolocating remote assets is also useful for various covert and SIGINT operations, but these methods are not publicly available and generally rely on significantly different assumptions and requirements that make them less relevant to the task at hand, even though they might have more stringent threat models. For example, in determining where a supply-chain tampered asset has been deployed, an intelligence agency might rely on bespoke RF base stations in a given country or use wardriving to triangulate the position of the asset based on nearby public wifi networks. These methods need to be extremely secure—in that they must not be discovered by other actors—but they need not withstand adversarial attempts to spoof the location of the asset itself.

[28] Or at least plausibly outside of China.

[29] Items, equipment, data, or any other specific resources that are located away from the primary location or central point of operation and management. These assets are often spread across different geographical locations and are not physically accessible or directly observable by the asset owners or managers on a day-to-day basis.

Method	Description	Advantages	Limitations	Use-Cases
Asset-Reported	Involves the asset determining its own location via GPS or similar technologies and reporting it to a trusted server.	Commonly used, well understood technology. Can be very accurate (~3m) under optimal conditions.	GPS can be easily spoofed. Requires unobstructed signals from satellites, which may not be reliable indoors or in data centers.	More suitable for environments with low adversarial presence and clear sky visibility.
Topology-Based	Maps digital characteristics of the asset (e.g., IP address, wifi SSIDs) to known locations via the internet or other public infrastructures.	Utilizes existing digital infrastructure. Can be quite accurate (~50-~500m), given optimal conditions.	Relies on unreliable and insecure protocols and infrastructure. Typically requires adding (at minimum) an antenna to the die.	More suitable for geolocation in less security-sensitive applications that require connectivity to public infrastructure anyway.
Delay-Based	Uses a cryptographic challenge-response between the asset and landmarks to calculate location based on communication travel time (RTT).	Overall system security to spoofing and other manipulation techniques. Probably does not require adding new hardware to chips.	Difficult to achieve precise geolocations. Requires excellent distribution and coverage of landmarks—also requires setting up and maintaining a network of secure landmark servers.	Best suited for regulatory verification and high-security applications.

Modern methods for geolocating a given asset[30] can be broken down into three categories[31] (Jiang et al., 2021, Esposito et al., 2018):

1. **Asset-reported:** These methods involve the asset determining its own location via GPS or similar publicly available technologies and then reporting that location to a trusted server. However, GPS is an unencrypted and insecure protocol that can be trivially spoofed by even a relatively

[30] Excluding simplistic methods such as examining content-artifacts like language packs, since these are trivially easy to manipulate and also seem irrelevant for the given use-case.
[31] Note that some methods involve aspects from more than one category, for example, A-GPS which relies on fusing asset-reported GPS with topology based cell-tower information.

unsophisticated adversary (Tippenhauer, 2011). For a practical demonstration of this phenomenon, see also UT Austin Researchers Successfully Spoof an $80 million Yacht at Sea, wherein researchers managed to steer a yacht off its intended course in a manner completely undetected by the ship's instruments.

GPS is an unencrypted and insecure protocol which can be trivially spoofed by even a relatively unsophisticated adversary.

Two approaches for avoiding GPS spoofing seem promising but are probably not feasible for geolocating AI chips at the moment: spoofing-resistant GPS receivers and encrypted GPS frequencies. Given significant resource dedication and further academic/technological research, these could perhaps prove sufficient for the purposes of thwarting covertly and openly adversarial actors. It remains unclear who might invest in these technologies and whether it is worth trying to advance research into them specifically for the purposes of compute governance:

a. **Spoofing-resistant receivers**, such as the one described by Ranganathan, 2016, have not seen widespread adoption or development and typically require additional hardware (such as multiple antennae or gyroscopes) that is probably unlikely to fit on cutting-edge AI chips due to the limited space available on the die as-is.

b. **Encrypted GPS frequencies**[32] or private dedicated satellite positioning systems[33] are either undeveloped or reserved for military use (Teunissen & Montenbruck, 2017)[34].

Additionally, GPS-based methods rely on chips being able to receive unobstructed GPS signals from satellites. Given that the kind of chips discussed are likely to be exclusively deployed in data centers and similar facilities, it seems unlikely that they would consistently be able to do so, considering the amount of RF interference, bare metal, and similar obstructions involved.

Alternative short-range technologies, such as NFC (e.g., as used in contactless payment methods) or RFID (e.g., as used in e-passports), are not good fits since they only operate in relatively close proximity (~1cm - ~100m) and require the deployment of additional technological components adjacent to the asset itself[35] which would need to be pre-deployed to the data centers in question. Unless there is already an agreement to use such systems universally and methods to perform more secure attestation on these components, such methods do not appear to be useful for our purposes.

[32] Such as the US military M-code GPS.
[33] Perhaps using uplink RTT methods as described below and dedicated satellite dishes in data centers—although we are not aware of any current real-world systems that utilize such methods.
[34] And even these might also be vulnerable to replay attacks, depending on implementation.
[35] "Beacons" or base stations.

2. **Topology-based:** These methods attempt to map internet or other digital characteristics of the asset (typically an IP address, visible wifi SSIDs, cell towers, etc.) to known locations via Whois, DNS records, and other internet/digital infrastructures. For example, cellular providers can geolocate individual mobile handsets by first determining which cell towers it is in range of and then combining that information with signal strength measurements and the known coordinates of the cell towers. These methods are widely considered to be unreliable and insecure against technologically advanced actors (Jia et al., 2019), and many of the protocols (SUPL, SS7, Diameter, DNS, etc.) utilized by these methods have long histories of adversarial exploitation[36].

3. **Delay-based:** These methods involve a cryptographic challenge-response between the asset and a landmark or a series of landmarks. Given a successful verification of the asset via the cryptographic handshake, the possible location of the asset can be calculated via the amount of time it takes for the communication to travel between the components (round-trip time or RTT[37]).

   There are a number of different kinds of delay/RTT-based mechanisms that have been proposed or are currently in use, with varying features that affect the overall security, reliability, and accuracy of the measurement (Jiang et al., 2021). Recent algorithms have managed to reduce the median error distance range to <10km, given excellent landmark distribution and coverage (Ma et al., 2023). Although this range seems quite large in comparison to the previously described methods, it should suffice for many of the use-cases outlined earlier.

   See Case Study: Japan and China for an illustrative example of how delay-based geolocation with real-world public landmark data could assist regulators in verifying whether chips are likely deployed where they should be.

[36] SS7, for example.
[37] Round-trip time, or round-trip delay, refers to the overall amount of time it takes a given piece of data to travel from one point to another and back again. One-way trips are also referred to as “pings”

Example landmark deployment map with 15 landmarks servers. Made with http://openmaptiles.org/ & https://carto.com/

For minimally adversarial contexts[38], "tried and true" solutions such as GPS, cell tower geolocation, or utilizing public landmarks for delay-based methods might be sufficient for reliably proving or verifying geographic information to regulators. However, as discussed earlier, these methods are easily spoofable and might also be difficult to implement within noisy data center environments. Without inserting additional hardware into the chips, rudimentary delay-based methods with public landmarks or perhaps even simplistic IP/traceroute-based tracking could be used, but these are only likely to catch actors who are especially careless with their manipulations.

For covertly and openly adversarial contexts[39], as mentioned, a much higher level of security is required: In a case where the adversary is faced with a GPS-enabled, "region locked" chip, they could simply spoof the GPS signal to make the chip believe it is currently located in an allowed location, and regulators would

[38] As defined earlier: "*attackers do not spend much on attacks, and are very averse to being discovered attempting to compromise mechanisms.*"

[39] As defined earlier, covertly adversarial refers to: *"attackers [who] are more willing to spend substantial resources to compromise mechanisms, but still want to avoid being caught doing so,"* and openly adversarial actors are: *"attackers [who] are willing to spend very significant resources to compromise mechanisms, and are indifferent to this being discovered"*

have no way of knowing the chip was misrepresenting its own location. One proof-of-concept for such an attack cost the researchers ~$200 to develop and implement (Zeng et al., 2018). Similarly, if an adversary is faced with a topology-based solution on the chip, the adversary could turn to VPNs, SDRs, or other proxy technologies in order to make the chip calculate that it is somewhere else entirely. One proof-of-concept for such an attack cost the researcher less than $120 to implement.

For dealing with these actors, delay-based methods seem to be the most promising class of solutions, and it is worth discussing them further in some detail.

Delay-based methods for The General Geolocation Problem

There are many proposed delay-based schemes for geolocating remote assets. The majority of these schemes were developed in the context of cloud-data migration and, therefore, assume an adversary who aims to reduce costs by migrating user data internationally despite contractual or legal obligations (Paladi, 2014). This phenomenon maps fairly well to our "minimally adversarial" actor but not to the other two adversaries, and that fact is worth keeping in mind as we discuss security considerations.

The vast majority of currently proposed delay-based techniques are composed of two components:

1. **Geopositioning**, where delays between asset/landmark pairs are used to determine the maximum possible distance between them. This information is combined to determine a region where the asset may possibly be or is likely to be.
2. **Cryptographic proof of identity**, which is used to verify that the replies the landmarks are getting are actually coming from the real asset.

Geopositioning

For a specific asset/landmark pair, if the time it takes for communication to occur between them is known, the landmark (and/or the asset) can determine the maximum possible distance the asset could be located. This calculation relies on the propagation speed of network packets in standard fiber-optic cables (approximately two-thirds the speed of light in a vacuum, or ~200,000 kilometers per second) and can also include other estimated variables or constraints, depending on the exact technique. A simple calculation of upper-bounds can be represented thus:

1. Let *D* represent the maximum distance (in kilometers) from the device to the landmark.
2. Let *T* represent the one-way ping time (in milliseconds).

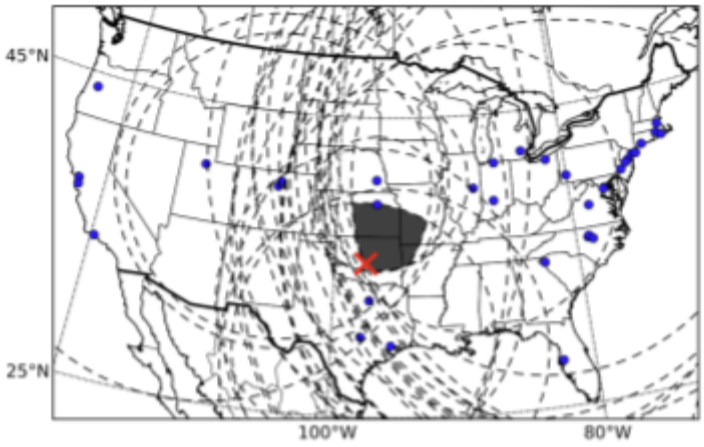
3. *D* is equal to *T* multiplied by the speed of light in fiber-optic cables, which is 200,000km/s: $D = T * 200$.

In other words, if a ping from an asset takes *T* seconds to reach the landmark, we can say that it's at most *D* kms away from the landmark—if it was located further away, then *T* would be larger.

[I]f a ping from an asset takes T seconds to reach the landmark, we can say that it's at most D kms away from the landmark—if it was located further away, then T would be larger.

Alternatively, the speed of light in a vacuum (~300,000km/s) can be used in order to account for worst-case scenarios such as advanced out-of-band radio transmissions, satellite communication, etc. The major advantage of using this number instead of the former is that it is now physically impossible for an adversary to "beat" the limit established. However, this approach may also run into problems in practice when assets are unable to establish communications quickly enough with landmarks, even though they are relatively close by—thereby generating false negative results.

From a moderate amount of experimentation with real landmark ping times, it seems likely that these false negatives will occur quite often, perhaps even in more than 50% of cases. As discussed later, it might be possible to calibrate landmark delay factors in a way that would mitigate this rate of false negatives somewhat, but implementing this scheme in practice would require further efforts to solve this issue. Various algorithms can then be used to narrow down the most likely area that the asset can be located in based on the intersection between these radii and additional factors such as known transmission delays, estimated consistency of RTTs, etc.

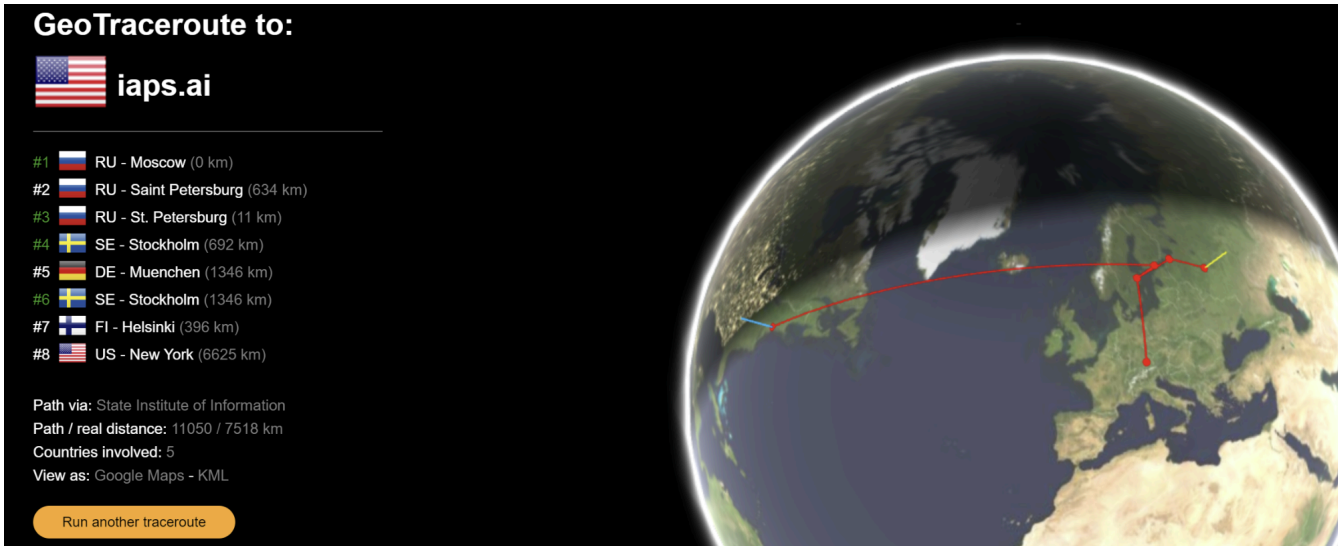


(The blue dots represent landmark servers, the circles are the distance upper-bounds for the asset, the shaded grey area is the calculated location of the asset, and the red X is the known location of the asset. Gondree & Peterson, 2013)

It is worth noting that the nature of the internet and international communication make *precise geolocation* using these methods difficult. Internet traffic does not travel in a straight line, nor does it always take the shortest path to the target. Rather, internet routing is based on complicated relationships between different internet peers, routing arrangements, congestion, fiber access, and other variables (Gondree & Peterson, 2013)[40]. These complex relationships affect the consistency of RTT measurements and also make timing-distance translation sensitive to regional nuances.

GeoTraceroute.com is a fantastic tool for visualizing the physical paths packets can take between two locations. For example, a packet originating from Moscow and destined for www.iaps.ai hosted in New York might pass through St. Petersburg, Stockholm, Munich, back to Stockholm, and Helsinki before finally arriving at its intended destination. As the crow flies, the distance between these two points is approximately 7,500km, but in practice, this packet has traveled 11,050km.

[40] One way to calibrate the distance-to-delay function (as in CBG) is to have the landmarks measure communication delays between each other multiple times to establish baselines for network delays.



Despite the challenge of providing a precise geolocation, these methods can nonetheless be used to determine that a chip is NOT located further than *D* away from the landmark and, therefore, could be put to use presently as an export control enforcement mechanism.

Cryptographic proof of identity

Various cryptographic methods can be used to establish that the asset is indeed the one communicating with the landmarks and that the contents of its communication have not been tampered with in any way. In order to establish these factors, the asset could use a combination of secret keys and internal information (possibly secured further by a Trusted Platform Module or TPM) and utilize a challenge-response protocol with the landmarks (Jiang et al., 2021 demonstrate one such arrangement utilizing an Intel SGX secure enclave on both the asset and the landmark which can securely communicate with each other, although there are other alternative methods that seem reliably secure which could be used[41]).

The cryptographic portion of the solutions relies on the typical security assumptions that are relevant to all remote-attestation and/or on-chip mechanisms: Namely, the asset must be able to keep its secret key and other information confidential from adversaries. See related research on the security of hardware-enabled mechanisms on AI chips (Aarne, Fist & Withers 2024).

This report primarily focuses on the geolocation portion of the solution, but it seems clear that there are numerous open questions and issues to resolve regarding the cryptographic component as well, which can hopefully be addressed in future research.

[41] It seems fairly reasonable to draw on other industries with reliable and secure AAA protocols, such as 4G/5G mobile networks. For example, the DIAMETER protocol serves a similar function within 4G/LTE networks.

Delay attacks

While delay-based methods offer promising security features, they are not immune to adversarial manipulation and attacks, mostly via timing delay manipulations.

Adversarial Strategy	Objective	Method	Implications and Challenges
Increasing Timing Delays	To increase the estimated geolocation area, making precise location identification difficult and decreasing confidence in results.	Deliberately routing traffic in a circuitous path, or deliberately delaying transmission, in order to artificially raise RTT measurements.	Increases uncertainty in asset's geolocation, potentially covering a wider area. Effectiveness varies; for short distances, the likelihood of detection decreases. Requires universal application to all communications for maximum effect.
Decreasing Timing Delays	To make the resulting geolocation appear to be in a different location than it actually is, with increased confidence.	Utilizing dark fiber and other high-speed interconnects or out-of-band methods to artificially lower RTT measurements.	Can falsely enhance confidence in a spoofed geolocation. The impact varies with the distribution of landmarks; non-uniform distributions can lead to skewed geolocations. Potentially shifts geolocation towards denser landmark areas.
Selective Delay Manipulation	To place an asset's geolocation in a specific, incorrect region (spoofing).	Consistently delaying or hastening packets for certain landmarks.	Requires knowledge of landmark locations. Feasible for dedicated actors but more challenging than universal delay adjustments.

			Strategy effectiveness depends on acquiring assets within "striking distance" of the desired location and a detailed understanding of the overall solution.
Malicious Landmark Take-over	To directly manipulate geolocation data by controlling the source of RTT measurements.	Gaining control over one or more landmarks to report false timing measurements.	Allows for spoofing without relying on man-in-the-middle network manipulation. Effectiveness scales with the number of landmarks controlled. Particularly risky if landmarks are physically located within adversary-controlled regions.

Adversaries can artificially increase the average delay measurements between the asset and the landmarks by deliberately routing traffic in a circuitous way or by otherwise extending the length of the traffic's route in some way. Gill et al., 2010 demonstrate attacks that can move the presumed geolocation up to 1,000km, with a 74% chance of avoiding detection, with better odds for shorter distances, via artificial RTT increases.

If this increase is performed universally for all communication between landmarks and the assets, this has the effect of increasing the uncertainty of the resulting estimated geolocation. In other words, the asset could plausibly be in a wider region than it would otherwise appear if a delay had not been introduced.[42]

In cases where an adversary is attempting to positively attest to the chip being in a specific location where it is not located, the chip would likely have to actually be within several hundred km[43] of the attested location for this spoofing to be reliably covert. So, for example, an adversary with a chip in Kaliningrad that is supposed to be in Gdansk (~100km away) could quite easily expand the estimated geolocation to cover both locations.

[42] Since the upper-bounds for each radius around a given landmark have increased.
[43] This number can be further examined if necessary, but for now using the figures from Gill seems reasonable.

This technique would also work if an adversary is attempting to negatively attest to the chip not being in a restricted or forbidden region. Returning to our example, if the chip is disallowed from being in Russia, expanding the estimated geolocation by several hundred kilometers means the chip could plausibly also be located in Latvia, Lithuania, Estonia, Belarus, Ukraine, Sweden, or Poland.

Gondree & Peterson, 2013 and others state that these kinds of attacks do not apply to their models since they are *verifying* that a certain asset is in a given location. It doesn't appear to matter to them whether other locations also fall within the same geolocation so long as the given location does as well.

A reasonable way to overcome this adversarial strategy is to set hard limits for communication speeds. In other words, if the asset takes longer than *x* milliseconds to respond to the landmark, it cannot be ruled out that it is in a restricted location.

A reasonable way to overcome this adversarial strategy is to set hard-limits for communication speeds. In other words, if the asset takes longer than x milliseconds to respond to the landmark, it cannot be ruled out that it is in a restricted location.

If this increase is instead performed selectively for communication between landmarks and the assets, it could actually place the geolocation in a completely different region than it would otherwise appear. This strategy requires some knowledge of landmark locations on the part of the adversary, as they would need to consistently delay packets headed for certain landmarks for different lengths of time in order to maintain a consistent attested location. This type of delay seems much harder than universally increasing delay but is feasible for a dedicated nation-state actor to do. However, it doesn't seem necessary to engage in this type of delay since universally increasing (and decreasing, as later shown) RTTs achieves the adversary's goals almost as well.

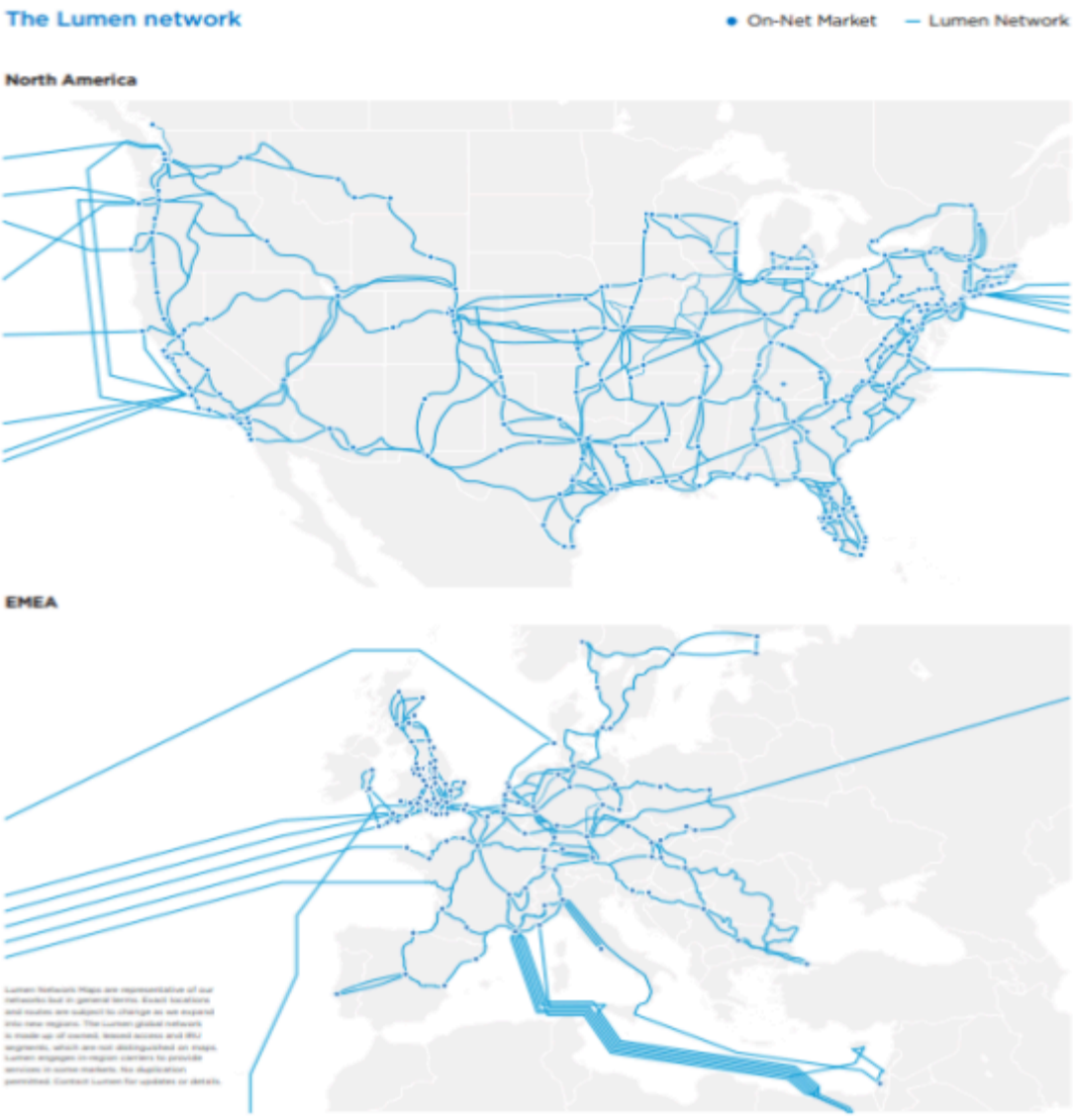
The exception would be if the adversary is unable to acquire large quantities of chips that are registered as being in nearby (~several hundred km) locations. For example, an adversary in South Africa that intends to deploy chips that are registered in Canada would have to try to selectively increase delays in order to genuinely spoof the location of the chip. Universally increasing delays to such an extent that the chip could plausibly be in either Canada or South Africa would mean that the confidence in the geolocation is incredibly low, and this would be highly suspicious and noticeable.

Alternatively, adversaries can artificially decrease the average delay measurements between the asset and the landmarks, utilizing dark fiber and other private high-speed interconnect or peering options. Gondree & Peterson, 2013 put it this way (emphasis added): *"This assumption—that remote sites are not connected by a private network, of significantly better quality than the Internet—is necessary for delay-based IP*

geolocation (and our work); we acknowledge, however, that providers renting dark fiber may undermine such an assumption."

Leasing or renting dedicated bandwidth in a dark fiber cable would probably not pose a considerable challenge for covertly or overtly adversarial actors. Dark fiber is widely available globally, and leasing a single "strand"—an individual optical fiber within a larger cable—would probably cost somewhere between $100-$1,500 per mile per month, considering the low amount of traffic involved[44]. Returning to our Kaliningrad, Russia to Gdansk, Poland example from earlier, leasing a strand of dark fiber would result in monthly costs of approximately $7,800-$117,000. These prices are gross estimates, and costs can obviously vary depending on geographic region and overall dark fiber availability—but it's also worth noting that many of these potential adversaries probably already have access to some dark fiber infrastructure and, therefore, adding this traffic to an existing leased line would be a negligible expense.

[44] https://availabilitydigest.com/public_articles/1308/dark_fiber.pdf, https://nesdarkfiber.com/pricing/



Lumen leasible dark fiber network map, https://www.lumen.com/en-us/networking/dark-fiber.html

In the available literature, the assumption that adversaries will not use dark fiber or similar options is implicitly assumed, often without even noting that there are plausible ways an adversary could shorten the amount of time communication would typically take. Due to this assumption, there are no available calculations for how much of a decrease an adversary could achieve or how that might affect the overall geolocation.

The effect of universally decreasing RTT measurements probably varies depending on the exact algorithm and technique utilized in order to calculate the resulting geolocation and also depends heavily on the distribution of the landmarks.

If the landmarks are uniformly distributed around the asset (so that the resulting geolocation would have placed the asset in the center of the area), decreasing the measured RTTs would increase confidence in

the resulting geolocation and reduce the possible region the asset could plausibly be located in. Theoretically, this is bad for an actor attempting to spoof a location via measurement attacks.

However, non-uniformly distributed landmark networks seem much more realistic in practice[45], and in this case, decreasing the measured RTT would seem to shift the resulting geolocation[46], likely away from the actual location of the asset, towards the largest mass of landmarks. In addition, decreasing the measured RTT would likely increase confidence in this new, false geolocation—a significant boon for an attacker attempting to spoof the true location of the asset.

Selectively decreasing RTT measurements to some landmarks but not to others could achieve similar results to selectively increasing measurements, as previously discussed. Limitations and requirements on the part of the adversary would likely be similar as well.

Adversaries can also artificially increase or decrease the reported average delay measurements between the asset and the landmarks via malicious takeovers of the landmarks themselves. As mentioned, these landmarks are simply servers capable of certain activities that are located in known locations and are not immune in any way to cyber attacks. If an adversary gains control of one or more landmarks, they could plausibly cause the landmark to report deliberately incorrect information regarding the timing measurements, as discussed by Jiang et al., 2021.

This reporting of incorrect information could allow an attacker to selectively manipulate the resulting geolocation without needing to deal with actually delaying or hastening packets going through networks. Possibly, this method could lead to a more reliably spoofed location since there are fewer additional inconsistent variables that the adversary would need to take into account (e.g., congestion, routing delays, etc).

The effectiveness of this strategy depends heavily on the overall amount of landmarks being used, the number of landmarks that have been taken over, and the artificial increase or decrease in measurement that the adversary introduces. Jiang et al., 2021 show that an adversary can create a ~700km geolocation

[45] Especially if the coverage is attempting to be global or for entire continents rather than a small bounded geographic region. In all surveyed papers that attempt to geolocate an asset with internet landmarks, the resulting geolocation did not place the asset in the center but rather closer to an edge or border.
[46] Since the upper-bounds for each radius around a given landmark have decreased.

shift by manipulating ⅓ of the landmarks used.

Fig. 9. Attestation with remote data access.

A malicious server in Qingdao provides false timing information, successfully shifting the attested area of the asset from Qingdao to the Korean peninsula, ~700KM away. Jiang et al., 2021

One risk factor that could make landmarks more susceptible to malicious take-over is being within "striking distance" of the adversary. That is to say, if the People's Liberation Army of China (PLA) is attempting to spoof chip locations, and many of the landmarks are located within China[47], these landmarks would probably be accessible by PLA agents and actors, who could launch cyber-physical attacks. In effect, this proximity would lead to a much larger attack surface for adversaries, who would be able to utilize techniques such as Direct Memory Access (DMA) attacks, Cold Boot attacks, etc., and, therefore, significantly increase the likelihood of malicious landmark take-over.

Proposals for Mitigation

The possibility of universal and selective delay-attacks and the potential use of dark fiber, malicious take-over of landmarks, and similar options seem to imply that even cutting-edge delay-based methods will not prove sufficient for reliably and accurately geolocating assets that are in the possession of covertly or openly adversarial actors in 100% of cases.

It's possible that there are alternative techniques that can deal with both delay-attacks and dark fiber, either algorithmically or via some other method. One possible technique is to require a certain level of confidence in the estimated geolocation, or else this geolocation is discarded. This strategy might work for some increase-delay based attacks but not for decrease-delay attacks. Another issue with this approach is that it seems like it would be very difficult to ascertain the reason that the confidence in an estimated

[47] Which—security considerations aside—seems desirable, since it would lead to more granular geolocations and increased confidence.

geolocation is poor[48]—global network infrastructure and conditions result in an imperfect mapping of RTTs to distances, and these sorts of relationships are very hard to predict ahead of time[49].

Another possible strategy would be to have many landmarks located very close to where a chip is supposed to be. This approach would probably defeat most of the universal spoofing techniques described[50]; however, it has some limitations.. This method would probably not reliably prevent selective increases or decreases, which lead to genuine shifts in the geolocated region, as previously discussed. This method would also increase the risk of malicious landmark take-over, as previously discussed. And finally, this strategy would probably significantly increase costs associated with setting up and maintaining the landmark network since many additional landmarks would need to be deployed in geographically diverse locations around areas of interest—i.e., not just in a data center or two, but in multiple diverse locations within each city of interest.

Despite these security concerns, there could still be significant value in adding location verification features to AI chips, granting that they are likely to be circumventable by motivated and sophisticated adversaries. These features could also enable more actionable steps on the part of regulators, such as remote shut-down and selective throttling, as mentioned earlier.

Despite these security concerns, there could still be significant value in adding location verification features to AI chips, granting that they are likely to be circumventable by motivated and sophisticated adversaries.

Some adversaries might err in their calculations and provide plainly false locations for the chips, or they could have implementation issues that give the game away, and so on. In this context, we would argue that something is better than nothing—even an insecure feature could provide some amount of value (although a secure version is obviously much more desirable) by occasionally catching cheaters. A decent analogy might be your typical email spam filter: broadly successful at catching and stopping spam emails that conform to boiler-plate templates and can be easily identified, even if it is unable to stop more sophisticated actors. Despite their clear imperfections, we still rely on spam filters to keep inboxes secure from the majority of low-effort spam and phishing attempts.

[48] I.e., if this low confidence is the result of intentional spoofing or not.
[49] For example, the average RTT between Seoul and Shanghai is ~356ms, similar to the RTT between Seoul and Bratislava (~349ms)—even though Bratislava is ~9 times further away from Seoul than Shanghai is.
[50] Since it increases the chances that the asset will truly be located in the middle of the geolocation, in which case increasing or decreasing the overall region would still leave the asset within the geolocation.

Imperfectly secured location verification methods would be sufficient deterrence for minimally adversarial actors, who are very averse to getting caught attempting manipulation. This method would probably be sufficient deterrence for covertly adversarial actors, who would have a difficult time making sure their geolocations make sense 100% of the time. Two factors which could significantly increase the likelihood of detecting manipulation would be:

1. **Centralized chip registries**, which would allow regulators to positively verify that chips are where they are supposed to be. This strategy would probably lead to adversaries needing to not only acquire large quantities of chips but also have those chips registered in geographically adjacent locations to the true location of the chips, otherwise, spoofing becomes extremely difficult or perhaps impossible, as discussed above.

2. Deploying and maintaining **large numbers of landmarks close to areas of interest**, which would not only make spoofing via a universal increase or decrease of RTT measurements more difficult to achieve but would also increase the granularity of the resulting geolocations. This method would enable more fine-grained verification, perhaps even at the data center level, as discussed earlier.

Openly adversarial actors, who are indifferent to being caught, would nevertheless be impacted by this feature for two main reasons:

1. It would likely **increase the costs associated with setting up and maintaining an illicit AI cluster** since they would need to build spoofing mechanisms and procedures, and constantly operate and troubleshoot these in order to avoid regulators taking action against their smuggling networks or even the chips in their possession by denying operating licenses, if such a requirement is implemented.

2. Occasionally, they might get **caught attempting to spoof the locations of the chips**, leading to regulators taking action against them.

Although there are inherent challenges in thwarting sophisticated adversarial tactics like timing delay attacks and landmark takeovers, implementing delay-based location verification features for global geolocation into AI chips would be a valuable, albeit not foolproof, tool for regulatory enforcement and deterrence against manipulation. Even an initially imperfect solution can serve a crucial role in deterring low-effort malicious activities and providing actionable regulatory controls.

Delay-based methods for The Anti-Smuggling Geolocation Problem

Setting aside the more general and complex problem of accurately geolocating assets we know nothing about ahead of time, there is a narrower version of the proposed delay-based solution that is specifically relevant to anti-smuggling use-cases, where we are simply attempting to verify that an asset is not in a restricted location. **This method could be an invaluable tool for both verifying and enforcing compliance with current export controls[51], which attempt to limit certain actor's ability to acquire advanced AI chips.** For more on how likely export control circumvention of this kind is and what anti-smuggling methods are available to BIS and other governmental actors presently, see Grunewald & Aird, 2023.

As discussed earlier, this approach is more similar to Gondree & Peterson, 2013 and others, which is less susceptible to adversarial manipulation techniques and implies both simpler algorithms and a more straightforward landmark coverage deployment.

To achieve sufficient global landmark coverage, a number of landmarks would need to be deployed in specific locations. For North America, South America, Africa, Australia, and Western Europe, several landmarks each would probably suffice due to their distance from restricted countries (China, Russia, etc.). For East Asia, parts of Southeast Asia, and Eastern Europe, a single landmark could cover a major city/metropolitan area or subregion. These regions would require more liberal deployment of landmarks than the rest of the world due to their proximity to restricted countries, as demonstrated in the below case study on Japan and China.

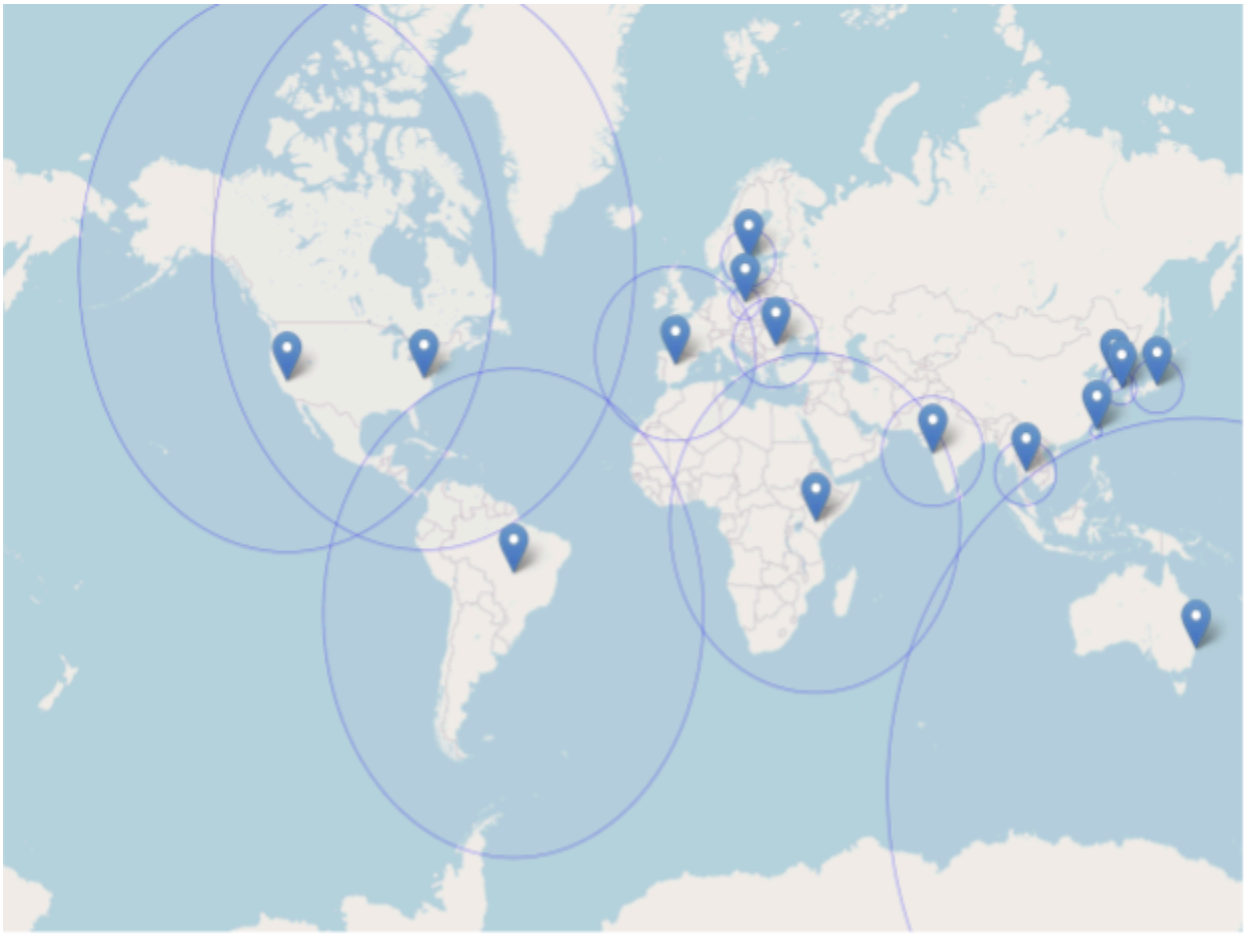
Roughly, a given geographic location (*L*) can be said to be "in-coverage" if the minimal distance between *L* and a restricted location (*R)* is larger than *D*. Or, $min(||L - R||) > D$.

Taking this equation into account and surveying potential landmark locations for ping-measurement latency information[52] could allow for strategic placement of landmarks, which would reduce the overall number of landmarks that need to be deployed. Additional considerations, such as whether or not a given area is likely to host compute clusters[53], could further narrow down the overall amount of deployed landmarks. Initial investigations into how many landmarks might be necessary for achieving a global coverage scheme seem to show that several dozen landmarks at most would probably suffice for the purposes discussed in this paper.

[51] "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections.", Supplementary information section D.2, 88 Fed. Reg. 73458, October 25, 2023, https://www.federalregister.gov/d/2023-23055/p-350

[52] Which, as mentioned, can vary in unintuitive ways.

[53] Based on existing data centers, infrastructure, etc.



Example landmark deployment map with 15 landmarks servers. Made with http://openmaptiles.org/ & https://carto.com/

However, adversaries interested in overcoming landmark-based positioning could potentially utilize dark fiber (as discussed earlier) in order to speed up ping latency between a given asset and a landmark to make the asset appear closer to the landmark than it actually is. Whenever possible, landmarks should be placed in a way that minimizes additional latency[54] and in locations that limit the overall effectiveness of such manipulation methods.

It is at least plausible that more sophisticated and determined adversaries could also utilize faster out-of-band communication methods to "beat" the speed of light in fiber optic cables[55]. Doing so would require a large amount of effort on the part of the adversary (much more than simply utilizing dark fiber) and might not be feasible due to the delay introduced by medium-switching[56]. Although faster candidate out-of-band communication methods do exist, either in theory or practice, it's possible that tools such as traceroute, ISP logs, and other passive methods (see Templeton & Levitt, 2003) could be used in order to

[54] I.e., with direct access to data center fiber connections with as few routing components as possible.
[55] Via high-speed laser satellite connections, for example.
[56] Fiber-to-laser/radio when leaving the source data center and then laser/radio-to-fiber again upon arriving at the destination landmark data center.

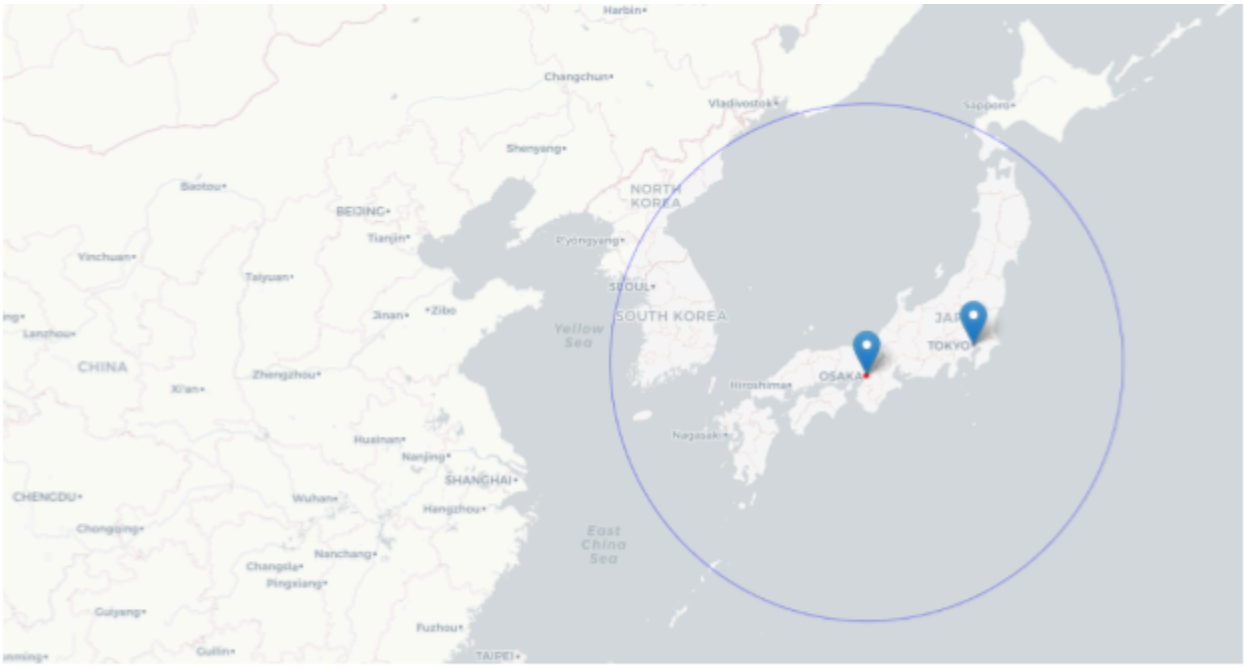
determine what approximate logical/physical route the communication is taking between the asset and the landmark. Coordinating with IXPs and other peering providers could also help determine the exact path the packet took and whether or not spoofing might have occurred (Lichtblau et al., 2017). Therefore, it seems unlikely that minimally adversarial actors would be able to take advantage of these methods without significantly increasing the risk of being caught. However, as mentioned above, if this concern proves to be more serious in the future, the delay-based geolocation calculations could rely on the speed of light in a vacuum instead of the speed of light in fiber optics cables.

Case Study: Japan and China

To illustrate how location verification might be helpful for real-world use-cases, including verifying compliance with export controls and verification that a chip location matches its listed registration location, let's look at two hypothetical examples:

1. A Japanese company purchases a controlled AI chip and deploys it to a data center in Central Tokyo.
    a. Upon deployment, the chip establishes secure communication with the various landmarks.
    b. The closest landmark is determined to be in Osaka[57], with an average data center-to-data center one-way ping time of 4.665ms[58].
    c. The upper-bounds possible distance from the chip to the landmark is 933km, since D = 4.665*200,000.
    d. Plotted on a map (the blue circle represented the possible geolocation):

[57] Assume for the purposes of this exercise that we do not have a landmark in Tokyo.
[58] Measurements taken from https://wondernetwork.com/pings/tokyo.

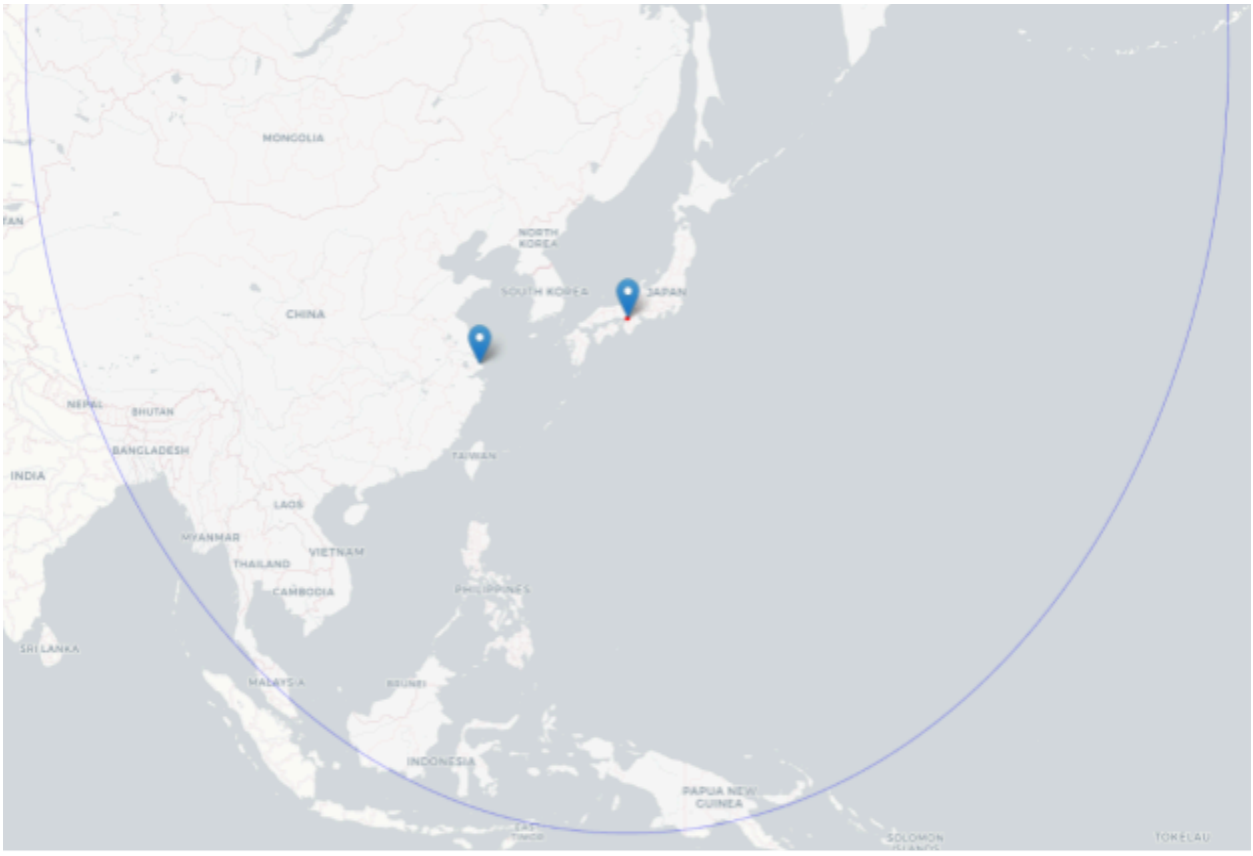


Made with http://openmaptiles.org/ & https://carto.com/

Therefore, according to our previous calculations and thanks to the delay-based geolocation technique, it is implausible that this chip is located in China[59].

2. A Japanese company purchases a controlled AI chip and then resells it to a restricted Chinese company, which deploys it to a data center in Shanghai.
    a. Upon deployment, the chip establishes secure communication with the various landmarks.
    b. The closest landmark is determined to be in Osaka, with an average data center-to-data center one-way ping time of 24.24ms[60].
    c. The upper-bounds possible distance from the chip to the landmark is 4,848KM since D = 24.24*200,000.
    d. Plotted on a map (the blue circle represented the possible geolocation):

[59] North Korea is a theoretically plausible location for this particular example, however, given how disconnected North Korea is from the global internet infrastructure, it seems extremely unlikely that a chip located in North Korea would be able to ping Osaka in under 5ms. This communication would also have additional strange features that could be spotted, such as suspicious source IP, traceroutes, etc.
[60] Measurements taken from https://wondernetwork.com/pings/shanghai.



Made with http://openmaptiles.org/ & https://carto.com/

Therefore, according to our previous calculations and thanks to the delay-based geolocation technique, it cannot be ruled out that this chip is in China, and further investigation into the location and use of the chip is warranted.

As this case study demonstrates, location verification could provide immediate value in helping regulators understand how far away the chip could maximally be from a landmark located inside a non-restricted country based on the recorded ping time. A smaller ping time suggests closer proximity to the landmark in question, while a greater ping time suggests the traffic had a longer distance to travel in order to reach its destination. If regulators cannot confidently determine that the chip is in a permitted location, they can then take action, as elaborated upon earlier.

Proposed Solution Requirements

This section describes in broad terms some of the key features and prerequisites involved in setting up and operating a delay-based geolocation scheme as described earlier. The two main areas of consideration are on-chip additions and landmark servers. The intent is to gesture at the sort of work that needs to be done

before a secure, global solution can be deployed, without diving too deeply into any single requirement in particular.

On-Chip Additions:

1. **Hardware/Firmware**: Unique secret keys would need to be provisioned per chip[61], and these would need to be stored securely via TPM or a similar solution. A limited interface with the rest of the chip for cryptographic challenge/response would also be required. Existing AI chips from companies such as Nvidia and AMD are already equipped with security modules that could likely be used to implement this functionality via a firmware update.[62]

2. **Software:** An application for handling the actual communication with the landmarks would need to be added to the chips, or existing applications would need to be enhanced with this additional functionality.

3. **Precise Timekeeping:** The chip would need to have access to the current time, whether via some internal mechanism or networked time protocols (NTP, etc.). If such information is already available for other purposes, no additional functionality is required. It should be noted that precise local timekeeping carries its own set of security and technical concerns[63]. If implementing an accurate timekeeping system within the asset itself proves too difficult or risky, it should be possible to rely only on the landmark's timekeeping capabilities—independently of the asset.[64]

4. **Landmarks:**

   A secure network of global landmarks would need to be set up, consisting of:

   a. **Servers**: Since the required functionality is minimal, these could be standard off-the-shelf servers.

   b. **Application:** For cryptographic authentication of chips via predetermined mechanism/protocol; for distance calculation for each chip closest to that particular landmark; and for coordination with other landmarks or with a central back-end server.

   c. **Connectivity:** Low-latency internet access. Direct data center deployment is highly recommended in order to minimize congestion, routing, and other variables that could interfere with distance calculations.

   d. **Security:** The servers could be configuration-hardened so that they are only capable of running the specific programs necessary for functioning as landmarks in order to minimize

[61] So that landmarks could securely verify their identities remotely.
[62] Nvidia GPUs have GPU system processors, code named "Peregrine," that support various cryptographic operations (Sijstermans & Xie, 2020). Nvidia's flagship H100 GPUs also have trusted execution environments (TEE) that could likely be used to implement location attestation (Nvidia, 2023). AMD GPUs likely also include some form of platform security processor that could be used to implement location attestation.
[63] See NTP's sordid history, for example.
[64] However, a region-locking feature enforced by the chip itself could require reliable on-chip timekeeping to measure how much time has passed since the last time the chip was able to verify its location.

their attack surface. Additional security mechanisms, both physical (caged racks) and virtual (malware protection, etc.), should be utilized to reduce the likelihood of malicious manipulation or takeover.

The initial purchase and installation of the landmarks is likely the most expensive and complicated part of the undertaking, but considering that the total amount of necessary landmarks necessary in order to achieve sufficient global coverage is likely to be several dozen, at most—even this step seems like it would not be exorbitantly expensive for a government agency[65].

The initial purchase and installation of the landmarks is likely the most expensive and complicated part of the undertaking, but [...] even this step seems like it would not be exorbitantly expensive for a government agency.

Future Research Directions

This paper only gestures at what a future location verification scheme might look like in practice, and there remain several significant open questions and issues that might need to be addressed before such a mechanism can be implemented in practice.

Who, specifically, is responsible for verifying chip locations post-deployment? Possible candidates include the chip manufacturers themselves and governmental agencies with anti-smuggling mandates such as BIS. What kind of oversight would such a scheme require in order to ensure that location verification schemes are not misused in ways that impinge upon personal or civil liberties? Are there any legal—American or international—hurdles that might be relevant to verifying chip locations? Can we learn anything from previous attempts to track illicit or restricted materials globally, such as nuclear materials or firearms?

Further, the cryptographic portion of the solution is only broadly described and requires a more detailed design that takes into account the adversarial models previously discussed. Various other crucial details, including the nature and location of keys and the protocols used for mutual or unilateral authentication, also need to be determined.

[65] Some extremely tentative back-of-the-napkin calculations indicate costs of ~$1,500-$3,000 per server for the initial set-up, and an additional $2,000-$6,500 annually per server in datacenter fees and maintenance costs.

As mentioned above, the exact number and location of the landmark servers would also need to be decided based on the identity of the actors who are currently restricted from obtaining advanced AI chips, international traffic patterns and trends, security considerations, and the overall costs of the program. Additionally, deploying landmarks within countries that might be considered likely to try to maliciously manipulate them is suggested as a potential security concern, but it is possible that sufficiently secure servers or deployment practices could be designed in such a way that this is no longer an issue.

Finally, several smaller technical questions still remain. These questions could potentially be developed further: Could GPS-resistant receivers or encrypted GPS signals be an alternative solution, and what would it take for such technologies to be considered a satisfactory alternative? Are there detection methods that could be developed that make the use of dark-fiber by adversaries detectable to such an extent that it makes spoofing untenable?

Despite these open questions, our findings clearly indicate that it seems both feasible and relatively cheap to implement pure-software delay-based solutions on chips in the near future and that these solutions could actively deter smuggling attempts by some actors and unlock new tools for regulators to enforce existing and future export controls on AI chips.

Acknowledgements

We are grateful to the following people for discussion and input: Michael Aird, Abra Ganz, Erich Grunewald, Lennart Heim, Timothy Fist, Gabriel Kulp, David Manheim, James Petries, Konstantin Pilz. Mistakes and opinions are our own. We are also grateful to Maya Deutchman for copyediting