# Nvidia Phones Home: No Backdoor Required

How the front door works: attestation, unforgeable identity and location verification, documented in plain sight.

**Abstract**

When governments in the East and the West, and the intelligence community at large, press Nvidia on whether its chips contain a backdoor, the question, though reasonable, looks in the wrong place: a hidden backdoor is unnecessary when the front door already does the work. Modern Nvidia data-center GPUs ship with hardware attestation. Each chip carries an unforgeable cryptographic identity, fixed in the silicon at manufacture, and will prove its firmware state on demand, through documented libraries (NVML and the attestation Corelib) and a published protocol (SPDM, the DMTF's Security Protocol and Data Model). That identity is verified, through a signed token, against reference values held by the vendor's Remote Attestation Service. Pair it with location verification, which infers where a chip sits from the time its signals take to reach servers at known places, and you have a phone-home system assembled entirely from parts that are in plain sight. This paper argues that the backdoor debate misreads the situation. Attestation is genuinely not a backdoor: it is read-only, locally initiated and the same mechanism, part for part, that secures laptops, phones and cloud hardware. But it is exactly the plumbing a control system needs, and the policy around it, from the United States Chip Security Act to proposals for a licence before a chip may connect, is turning an identity check into a permission system. We set out the export-control pressure that is driving this, what attestation is at the level of keys, protocols and tokens, how location verification works down to the physics, why it is not a backdoor and why the concern is nonetheless well founded, the precedents in which an identity mechanism became a control one, and what it all means for anyone buying AI infrastructure who would rather not have its continued use depend on a vendor's, or a government's, goodwill.

**Keywords:** GPU attestation; Nvidia; export controls; location verification; device identity; SPDM; DICE; confidential computing; remote attestation; Chip Security Act; digital sovereignty; AI infrastructure; hardware trust; supply chain

## 1. Introduction

Governments in the East and the West, and the intelligence community at large, keep putting Nvidia the same blunt question: do your chips contain a backdoor, some hidden means by which a foreign power could watch them or switch them off? Nvidia's answer has been equally blunt: no, we do not put secret tracking hardware or kill switches in our products. Both statements can be true, and the exchange can still miss what matters, because a hidden backdoor is unnecessary when the front door already does the work.

Modern Nvidia data-center GPUs ship with hardware attestation. Each chip carries a cryptographic identity fixed in the silicon at manufacture, an identity it cannot be made to forget or to misreport, and on request it will prove that identity and report the exact state of the firmware it is running. The mechanism is not concealed. It is described in Nvidia's own developer documentation, exposed through named libraries, built on a published industry protocol, and checked by a verification service whose error codes and token formats are documented in public. Pair that identity with location verification, a technique that works out roughly where a chip is from the time its signals take to reach servers at known places, and the two together make a phone-home system assembled entirely from parts that are in plain sight.

This paper is about that front door. We argue three things. First, that the backdoor debate looks in the wrong place, because the interesting mechanisms are the documented ones, and we can follow them down to the level of keys, protocol messages, signed tokens and the speed of light. Second, that these mechanisms are, individually, benign and well motivated: identity and integrity are the foundation of trustworthy computing, and the people who built them had good reasons. Third, that assembled and made mandatory they become something else, the architecture of a control system, and that the policy now in train is assembling and mandating them.

We proceed in order. We set out the export-control pressure that explains why anyone would build this (Section 2). We open up attestation itself, from the fused device secret through the SPDM handshake to the verifier's signed verdict (Section 3). We take location verification apart to the physics and the statistics (Section 4). We then make the careful case that this is a front door and not a backdoor, and the equally careful case that the worry is still well founded, with precedents for an identity mechanism hardening into a control one (Section 5). Finally we turn to the reader who is not a government but a buyer, and ask what the arrangement means for anyone whose work now rests on this hardware (Section 7).

A word on scope and sourcing. This is a position paper, not a vendor disclosure or a leak, and everything in it is drawn from public material: Nvidia's own developer documentation and standards references, the published export-control rules, the academic literature on geolocation, and reporting current to late November 2026, all listed in the references. Where we read beyond what a source states outright, we say so. We have deliberately kept the technical account at the level of documented mechanisms and public standards, and have avoided any operator's internal

configuration, because the argument does not need them and because the whole thesis is that the troubling capability is assembled from parts already in the open. The dates and legislative states are as reported at the time of writing and will move; the structure of the argument, which is about what the mechanisms make possible, will not.

## 2. Why anyone would build this

### 2.1. The export-control ratchet

For three years the most advanced AI chips have been the subject of a running contest between United States export policy and the market's appetite for compute. The pattern is a ratchet. Washington sets a performance ceiling above which a chip may not be sold to China; Nvidia designs a variant tuned to sit just beneath it; the ceiling is lowered to catch the variant; the cycle repeats. To understand why location verification is now wanted, it helps to see what the ceiling is actually made of, because the metric is as much a part of the story as the chips.

The controlling rule is the United States Export Administration Regulations entry known as ECCN 3A090, introduced in the Bureau of Industry and Security interim rule of October 2022 and tightened in the update of October 2023. The 2022 rule drew its line using two parameters, raw compute and interconnect bandwidth; chip designers promptly discovered that a part could be cut just at the interconnect to clear the bar while keeping most of its training value, and the 2023 rule responded by dropping the interconnect parameter altogether and adding a second test. The first parameter is Total Processing Performance, or TPP, a single number meant to capture how much useful arithmetic a chip can do:

$$\mathrm{TPP} = 2\,R_{\mathrm{MAC}}\,b, \tag{1}$$

where $R_{\mathrm{MAC}}$ is the aggregate multiply-accumulate throughput of the chip in tera-operations per second at the relevant precision, $b$ is the bit length of the operands, and the factor of two counts a multiply-accumulate as two operations. The second parameter, added in 2023, is performance density: TPP divided by the die area in square millimetres. Its purpose is to catch a chip that reaches a permitted TPP only by spreading the same compute across a larger piece of silicon, which is exactly the move a designer makes to evade a pure-throughput cap. Table 1 sets out the thresholds that resulted.

The density cut-off of 5.92 was not arbitrary. The A100, the chip the rule was written to stop, has a performance density of roughly 6.04 by this measure, a hair above the line; the number was chosen to catch it and its close relatives while leaving mainstream graphics and gaming parts alone. Once the line is drawn this finely, the incentive on both sides is obvious. The regulator wants to know not merely what a chip can do but where it physically is, because a part sold lawfully into one jurisdiction retains all its value if it can be quietly carried into another. That is the demand that location verification answers.

### 2.2. The China editions

There is no single name for the chips Nvidia built to live under this ceiling, but they are recognisable by their model numbers, and each tells a small story about where the line stood when it was designed. Table 2 sets out the main ones.

The A800 and H800 are the clearest illustration of the ratchet. They were the A100 and H100 with the chip-to-chip interconnect throttled, which cleared the 2022 rule's interconnect test while preserving almost all of the single-chip training performance that mattered. The 2023 update, by dropping that test and adding the density one, caught them both at a stroke. The H20 that followed was cut harder, on raw compute this time, and had the stranger fate of being banned by Washington, briefly permitted again, and then waved away by Beijing, which discouraged its own firms from buying a chip it regarded as deliberately hobbled. The Blackwell-generation B30A and RTX 6000D were reported through 2025 as the next attempt, the B30A described as a single die of roughly half a full B300, though neither was confirmed to be shipping.

### 2.3. From cut-down chips to no chips

The latest turn removed the cut-down chips from the story altogether. Around the end of 2025 the United States approved export of the H200, an ordinary, full-strength chip of an older generation rather than a China special, on tight terms: units had to route through the United States for inspection and carried a 25% levy. Beijing then declined to approve large-scale imports, and by early November 2026 Reuters reported some 600,000 H200s sitting in Hong Kong, waiting. On 17 November 2026 Nvidia told investors that it was no longer shipping licensed Hopper data-center products to China at all, and that its revenue forecast now assumes no data-center compute sales there. The practical result is that, at the time of writing, Nvidia supplies no AI data-center chips to China.

That is the backdrop against which location verification is being built. When the question a government most wants answered is where a given chip physically sits, and whether a chip sold under one set of conditions has quietly moved to another jurisdiction, a reliable way to establish a chip's identity and whereabouts stops being a monitoring convenience and becomes an instrument of policy. The striking fact, and the subject of the rest of this paper, is that the instrument is already mostly built, and was built, in the first place, for other and better reasons.

### 2.4. The scale that makes it urgent

The reason a government reaches for location verification at all is that the paper controls are visibly leaking, and the leak is large. Smuggling advanced accelerators is not a matter of the odd card in a suitcase; it is an organised trade with the margins to support one. Reporting cited in support of the Chip Security Act put the volume of controlled Nvidia parts diverted into China at more than a billion dollars in a single three-month period, and the routes are the familiar ones of any sanctions-evasion economy: a lawful sale into a permissive third country, a reseller, a re-export that is never declared, and a chip that surfaces in a data center it was never licensed to enter. Bloomberg's reporters traced exactly such a path, following H800 and H100-class servers from a Singapore partner to a Shanghai facility, the kind of documentary trail that exists only after the fact and only when someone goes looking.

**Table 1.** The control metric. The October 2023 United States rule controls an accelerator by two tests at once, aggregate throughput (TPP) and how densely that throughput is packed onto the die. The density test was added so that a chip could not clear the line simply by using more silicon; the chosen cut-off sits just below the performance density of the A100, which is why the first China editions, the A800 and H800, were caught.

| Control tier (ECCN 3A090) | Total Processing Performance (TPP) | Performance density (TPP per mm$^2$ of die area) |
|---|---|---|
| 3A090.a (controlled) | 4800 or above; or 1600 or above when the density test below is also met | 5.92 or above, applied together with a TPP of 1600 or above |
| 3A090.b (grey zone, lighter controls) | between 2400 and 4800 | between 1.6 and 5.92 |

**Table 2.** The China editions. Nvidia has repeatedly built cut-down chips tuned to sit just below the prevailing United States export threshold; each was recognisable by its model number, and most were overtaken by the next tightening.

| Model | What it is | Status |
|---|---|---|
| A800 / H800 | The first China versions of the A100 and H100, cut at the interconnect to clear the 2022 line | Blocked by the October 2023 tightening, which dropped the interconnect test, along with the RTX 4090 |
| H20 (and L20) | A cut-down Hopper part, offered after the A100 and H100 were banned | Banned, briefly resumed in July 2025, then discouraged by China's own regulators |
| RTX 5090 D | A China gaming card shown at CES in January 2025: full gaming performance, reduced AI performance | A consumer product |
| B30A / RTX 6000D | Reported Blackwell-based China variants; the B30A a single die, roughly half a B300 | Reported during 2025, not confirmed as shipping |

That is the gap location verification is meant to close, and seeing the gap clearly also shows the method's natural ceiling. Paper controls fail because a chip, once sold, carries no live account of where it is; a licence and a customs form describe where it was supposed to go, not where it went. A chip that reports its own rough position changes that, but only for chips that are switched on, networked and running the reporting agent, which is to say not for the chips a serious smuggler most wants to hide. The honest framing, which the policy debate often blurs, is that location verification raises the cost and the risk of diversion at scale, turning a quiet re-export into something that must be kept permanently offline or permanently lied about, without ever making diversion impossible. It is a deterrent and an auditing tool, not a seal. The urgency comes from the size of the leak; the limits come from the fact that the leak is run by people who will not volunteer to be measured.

There is a second pressure in the background, pulling the other way, which is that China is not standing still while the controls tighten. Beijing's discouragement of the H20, a chip Washington had grudgingly allowed, was not softness; it was a push towards domestic accelerators, and every month the controls hold is a month that pressure builds. This is the uncomfortable strategic backdrop to the whole enforcement question: a verification regime that works too well accelerates exactly the self-sufficiency it was meant to forestall, while one that works too poorly fails at its stated job. Location verification sits in the middle of that tension, which is part of why its design, and the question of how far it will be pushed towards control, is contested rather than settled.

## 3. The front door: what attestation is

### 3.1. Two claims: identity and integrity

Before location comes identity. Every modern Nvidia datacenter GPU can attest to itself, and attestation is best understood as a device making two separate claims and backing both with cryptography. The first claim is identity: I am this specific, genuine part, made by this manufacturer, and not a counterfeit or a stand-in. The second is integrity: the firmware I am running is the firmware I should be running, unmodified, at the version I report. A verifier who accepts both claims knows it is talking to a real chip in a known-good state. Everything that follows, the keys, the protocol, the libraries and the verification service, exists to make those two claims checkable by someone who was not present when the chip was made and does not trust the machine the chip sits in.

### 3.2. Device identity: fused secrets and the DICE pattern

The identity claim rests on a secret that never leaves the chip. At manufacture, a unique value is fused into the silicon, readable only by the chip's own security logic and exposed through no external interface. From that root the chip derives a key pair whose private half it can use to sign, and whose public half, wrapped in a certificate issued by the manufacturer's certificate authority, it can hand out. Because the root secret is unique per device and physically bound to it, the resulting identity cannot be copied onto another chip, cannot be relabelled, and cannot be forged without the manufacturer's signing key. In effect each chip carries a serial number that is not printed on a label but proven with mathematics.

The approach is not peculiar to Nvidia; it mirrors a well-established industry pattern, codified by the Trusted

Computing Group as DICE, the Device Identifier Composition Engine, and the pattern is worth stating because it explains both the strength and the sensitivity of the design. In DICE, a per-device Unique Device Secret (UDS) is combined, through a one-way function, with a measurement of the first layer of code the device runs, to produce a Compound Device Identifier (CDI). That CDI seeds the next layer, which measures and seeds the next, so that identity is chained to software state at every step. Two kinds of key fall out of this. A long-lived DeviceID key reflects the hardware and its lowest, immutable layer, and serves as the chip's permanent name. Short-lived Alias keys reflect the current firmware and change whenever that firmware changes, so that a signature made with an Alias key testifies not only to which chip signed but to what it was running when it did. A single altered byte of firmware yields a different measurement, a different CDI, and a different Alias identity downstream, which is precisely what makes tampering detectable and, as we note later, precisely what makes the firmware root so consequential.

### 3.3. SPDM: the protocol

Identity and measurements are useless unless a verifier can draw them out of the chip over a defined protocol, and here too the mechanism is a published standard rather than a private channel. Nvidia's GPUs speak SPDM, the Security Protocol and Data Model, specified by the Distributed Management Task Force as DSP0274; Nvidia's implementation uses SPDM version 1.1 to collect measurements. SPDM defines a request-and-response handshake between a requester (the host, or a management controller) and a responder (the device). A typical exchange runs through a recognisable sequence: `GET_VERSION` and `GET_CAPABILITIES` to establish what each side supports; `NEGOTIATE_ALGORITHMS` to agree on the cryptography; `GET_DIGESTS` and `GET_CERTIFICATE` to retrieve the device's certificate chain, which the requester validates up to the manufacturer's root; `CHALLENGE`, in which the requester sends a fresh random number and the device signs it, proving possession of the private identity key; and `GET_MEASUREMENTS`, which returns the signed firmware measurements, again bound to a requester-supplied nonce.

The nonce is the quiet heart of the design. Because the device signs over a random value the verifier has only just generated, a reply captured from an earlier exchange cannot be replayed to impersonate a healthy chip: the old signature will not match the new nonce. The measurements themselves arrive as a set of measurement blocks, each a hash of some component of the firmware or configuration, and the verifier's job is to check both that the signature is genuine and that the hashes equal the values a good chip should report. SPDM thus gives, in one authenticated session, the answer to both of attestation's questions: who the device is, and what it is running.

### 3.4. The libraries: NVML and the attestation Corelib

None of this is secret, and the evidence is gathered through documented software that anyone with access to the machine can run. NVML, the Nvidia Management Library, is the C interface beneath the familiar `nvidia-smi` command; it reports a card's utilisation, memory, temperature, power draw, error counts and running processes, and among its identification fields are the board serial number, the PCI device identifiers and the firmware version. A single `nvidia-smi -q` on a data-center card prints much of this in plain text. NVML ships with every driver, as `libnvidia-ml.so` on Linux and `nvml.dll` on Windows, with wrappers for Python, Rust and Go, so a chip's declared identity is readable today by any program with a login on the host.

The attestation evidence proper is gathered by a more specialised library, the Nvidia attestation Corelib. It is the component that speaks SPDM directly to the GPU firmware. Corelib discovers the GPUs in a machine, opens an SPDM session with each, and retrieves its firmware measurement report, its device certificate chain, its architecture and version, and the nonce-bound measurements that guard against replay. It only collects the evidence; the judgement of whether that evidence is acceptable happens elsewhere. Corelib is published and documented, with a plain C interface, and at the time of writing is supported on Blackwell GPUs on recent Ubuntu, for both x86-64 and aarch64 hosts. Our reading, not stated outright on its introduction page, is that NVML was historically the route to an attestation report but required the chip to be placed in Confidential Computing mode, whereas Corelib is the path for ordinary machines; either way, the capability is moving from a specialist corner into the mainstream toolchain, which is itself part of the story.

### 3.5. Verification: reference values, tokens and revocation

Collecting evidence is only half of attestation; someone must decide whether the evidence is good, and that decision is where the phone-home becomes concrete. The evidence a chip produces, its certificate chain and its signed measurements, is passed to a verifier, which Nvidia provides as the Nvidia Remote Attestation Service, or NRAS, reachable at a documented endpoint (for GPUs, a path of the form `/v4/attest/gpu`). The verifier does three things. It validates the certificate chain back to Nvidia's certificate authority, rejecting evidence whose chain does not check out, with documented error codes such as 4007, `INVALID_CERTIFICATE_CHAIN`. It compares the reported measurements against reference values, the hashes a genuine chip at a given firmware version ought to produce, which Nvidia publishes as signed reference integrity manifests. And it checks that the device certificate has not been revoked, through a separate online certificate status service.

If all three pass, NRAS returns not a bare yes but a signed Entity Attestation Token, an EAT, standardised by the IETF as RFC 9711 and carried as a JSON Web Token. The token is itself a verifiable artefact: it states what was attested and by whom, is signed by the attestation service, and can be handed on to a third party, a cloud customer, an orchestration system, in due course a regulator, who can check it without repeating the whole handshake. This is the feature that matters most for the argument of this paper. Attestation does not merely let a chip prove itself to the machine it sits in; it mints a portable, signed, standards-based claim about that chip's identity and state that can travel anywhere and be trusted by anyone holding the verifier's public key.

The plumbing already produces exactly the kind of token a permission system would consume.

### 3.6. The on-die root and the measurement log

It is worth pausing on where, physically, this all happens, because the location of the trust root is what makes the firmware question in Section 5.2 sharp. The cryptography of attestation does not run on the GPU's thousands of arithmetic cores, which are built for matrix multiplication and nothing like a secure element; it runs on a small on-die security and management processor, a dedicated controller with its own firmware that holds the fused secret, performs the signing, and speaks SPDM on the chip's behalf. Everything a verifier learns about the chip, it learns through that processor, and the integrity of the whole scheme reduces to the integrity of that one small, closed component. This is not a criticism unique to Nvidia; every hardware root of trust has an innermost layer that must simply be believed. It is the reason, though, that an attestation is only ever as trustworthy as the party that wrote and signed the root firmware, and that party is the vendor.

The measurements the processor reports are not a single hash but a structured log. Under SPDM the device returns a set of measurement blocks, each identified by an index and each carrying the hash of one measured thing: a firmware image, a configuration region, a security-version number, a debug-state flag. A verifier reads the whole set and checks it item by item against the values a known-good chip at a declared version should present. Those expected values are themselves distributed as signed documents, reference integrity manifests, so that the verifier is not trusting its own memory of what is correct but a manufacturer-signed statement of it, versioned and revocable. The structure matters for the argument, because it means attestation is fine-grained: it does not merely say good or bad, it says precisely which component is at which version, which is the resolution a policy engine would need to permit some firmware states and refuse others. The measurement log is, in effect, a machine-readable bill of health, and a bill of health is a document that can be made a condition of entry.

A further detail completes the picture. The CHALLENGE and GET_MEASUREMENTS exchanges can be run inside an established SPDM secure session, in which the two parties first perform an authenticated key exchange and then protect the remaining traffic, so that the evidence cannot be read or tampered with in transit by the very host the chip is trying to prove itself to. The design assumes, correctly, that the host operating system may be hostile to the attestation it is relaying. That assumption is prudent engineering for confidential computing, where the whole point is to defend a tenant against the operator. It is the same assumption a control regime would make for the opposite purpose: a channel built to resist tampering by the machine's owner is, by construction, a channel the owner cannot quietly disable or forge.

### 3.7. Confidential computing: the fuller picture

Attestation did not arrive for the sake of export control; it arrived for confidential computing, and understanding that origin keeps the assessment honest. On Hopper and later data-center GPUs, Nvidia supports a confidential computing mode in which the work running on the GPU, its code and its data, is protected from the host it runs on: memory is encrypted, the GPU is placed in a protected state, and the tenant can run a model on infrastructure they do not own without the operator being able to read it. The whole scheme is worthless without attestation, because the tenant must be able to verify, remotely and before sending any secret, that the GPU really is in the protected mode it claims and is running the firmware it claims. The GPU's attestation is typically combined with the attestation of a confidential CPU environment, such as Intel TDX or AMD SEV-SNP, so that the tenant can check the whole machine, processor and accelerator together, before trusting it.

This is a genuinely good use, and it is the reason the identity and measurement machinery exists at the quality it does. It is also why Corelib's arrival, collecting attestation evidence without requiring confidential computing mode, is significant: it takes a facility that was once reserved for tenants running encrypted workloads and makes it a routine thing any operator, or anyone an operator reports to, can invoke against an ordinary chip. The mechanism is unchanged; its reach is not.

The point to hold on to across this whole section is that attestation is a front door. The owner of a GPU, or a cloud customer renting one, can and does use it to make sure the hardware underneath is genuine and unmodified before trusting it with a workload. That is its designed purpose, and a good one. But the same mechanism that lets an owner verify a chip also lets anyone the chip will talk to establish, beyond forgery, exactly which chip they are dealing with, and to carry away a signed token saying so.

### 3.8. The limits of the claim

Precision requires saying plainly what attestation does not do, both because overstatement invites the dismissal that the whole concern is paranoia, and because the exact boundary is where the policy argument actually lives. Attestation, as it stands in the silicon today, is a reporting facility and nothing more. It exposes no means to run code of an outsider's choosing on the GPU, reads none of the tenant's data, and carries no command that could halt a workload or disable a card. A chip that has just produced a flawless attestation will go on computing whatever its owner tells it to, whether or not any verifier liked the result, because the verifier sits at the far end of a wire with no return path to the chip's execution. Everything alarming in this paper is therefore a statement about what could be built on top of attestation, not about a capability latent in it. The gap between the two is a single piece of wiring: a gate that consults the verdict before letting the chip do something it wants to do, network access, model loading, operation at all. The silicon does not contain that gate. The honest worry is not that the gate is hidden in the chip but that the gate is cheap to add outside it, that the verdict it would consume is already minted in a portable form, and that the law is now specifying where the gate should go. Keeping the distinction sharp is what lets one argue against the gate without having to pretend the chip is already a weapon.

## 4. Phone home: location verification

### 4.1. The physics

Knowing which chip is not the same as knowing where it is, and the where is what export policy cares about. The technique for the where is older than the chips, borrowed from two decades of research on locating ordinary Internet hosts, and it rests on a limit nobody can design around: nothing travels faster than light. A chip, or the server holding it, exchanges a signal with a machine at a known location, and the round trip is timed. The measured round-trip time sets a hard ceiling on how far away the chip can be, because no response can come back sooner than the speed of signal propagation allows. Formally, the maximum distance to the chip is

$$d_{\max} = \frac{v_p \left( \tau_{\mathrm{RTT}} - \tau_{\mathrm{proc}} \right)}{2}, \tag{2}$$

where $\tau_{\mathrm{RTT}}$ is the measured round-trip time, $\tau_{\mathrm{proc}}$ is any processing and queuing delay that is not travel time, the division by two converts round trip to one way, and $v_p$ is the propagation speed of the signal. In optical fibre that speed is about two-thirds of the speed of light in vacuum, roughly $2.0 \times 10^8$ metres per second.

A worked example shows how tight, and how loose, this is. Suppose a chip answers a challenge with a round-trip time of one millisecond and negligible processing delay. Then $d_{\max} = (2.0 \times 10^8)(10^{-3})/2 = 10^5$ metres, or about 100 kilometres: the chip is certainly within 100 kilometres of the landmark, and may be anywhere inside that radius. The bound is one-directional in a useful way. Delays can only lengthen a measured round trip, never shorten it below the travel time, so queuing, a slow network interface and indirect routing all make $\tau_{\mathrm{RTT}}$ larger and the inferred $d_{\max}$ more generous. This means the method cannot be fooled into thinking a distant chip is near: extra delay only ever makes it look farther away. It can, however, be very imprecise, because a single measurement yields a disk, not a point.

### 4.2. From one landmark to many

A disk becomes a location when several landmarks are combined. Each landmark at a known position, having measured its own round-trip time to the chip, contributes a distance ceiling and hence a disk within which the chip must lie; the chip must lie in the intersection of all of them. This is multilateration, and it is the core of the technique that the networking literature calls Constraint-Based Geolocation, introduced by Gueye and colleagues in 2004. Their contribution was to treat each latency measurement as a constraint bounding a continuous region, rather than guessing a single point, and to return a confidence region that tells the application how trustworthy the estimate is, which is exactly what an enforcement regime would need before acting on a location.

A concrete case shows the intersection at work. Suppose a chip is probed from three landmarks: one in Amsterdam measures a round-trip time implying a radius of about 120 kilometres, one in Frankfurt about 90 kilometres, and one in Paris about 400 kilometres. Each disk alone is a weak statement, but the region common to all three, the Amsterdam disk trimmed by the Frankfurt one and loosely bounded by the Paris one, narrows the chip to the west of the Netherlands and the nearby border area, enough to say with confidence that it is not in, say, Singapore or Shanghai. Add more landmarks and the common region shrinks further, though never to a point, and always subject to the caveat that an evader can only inflate the radii, never shrink them. For the policy question, which is a question about countries rather than streets, three or four well-placed landmarks already carry most of the weight, and this is why published estimates put a usable global system in the range of a few dozen to a few hundred servers rather than thousands.

The accuracy such methods achieve is modest but real, and it has improved. Constraint-Based Geolocation and its successors, such as the Octant framework, report median errors in the low tens of miles under good conditions: Octant's authors cite a median accuracy of about 22 miles, against roughly 70 miles for the earlier delay-based systems they compared with. These are self-reported results from controlled experiments, not independent benchmarks, and real deployments contend with asymmetric routing, congestion and hosts that are simply far from any landmark, all of which inflate the error. The honest summary is that latency geolocation reliably answers country-level and often region-level questions, and does not reliably answer street-level ones. For export control, country-level is most of what is wanted: the policy question is whether a chip licensed for one jurisdiction is now operating in another, and that is a question a disk a hundred kilometres across can often answer.

### 4.3. Fleet Intelligence

At Nvidia this physics lives in a product called Fleet Intelligence, announced in December 2025 and now generally available: a managed service for watching over a data center's fleet of GPUs, their power draw, their temperature, their health. It infers a GPU's location from two things together, the attestation data the chip provides about itself and the network latency between the chip and Nvidia's servers. Fleet Intelligence spans the Hopper, Blackwell and Vera Rubin generations, though the attestation it leans on is available only on the two newest, Blackwell and Vera Rubin. The two halves of the phone-home system are, at this point, clicked together in a shipping product: identity from attestation, which says beyond forgery which chip is reporting, and position from latency, which says to within a region where that chip sits. Neither half is novel. What is new is their integration, by the vendor, into a service that reports both to the vendor.

### 4.4. Can it be spoofed?

The natural objection to any location scheme is that a determined party will simply lie about where it is, and it is worth working through why that is harder here than it first appears, and where it remains easy. The latency measurement has an asymmetry that the attacker cannot wish away, already visible in Equation (2): delay can be added but not subtracted. An evader can insert artificial delay, relaying the chip's answers through an intermediary so that the round trip looks longer and the chip appears farther from the landmark than it is. What the evader cannot do is make

the signal arrive sooner than the speed of light permits, so a chip cannot be made to look closer to a landmark than it physically is. This one-sidedness is the method's quiet strength. To impersonate a permitted location, an evader must appear nearer to that location's landmarks than its true position allows, and that is the one thing the physics forbids.

The practical attacks therefore work on the measurement, not the physics. Proxying the chip's traffic through a machine inside the permitted region can forge the appearance of being there, provided the proxy is close enough to the landmarks and the added delay on the real leg is hidden within normal network variance; this is the same move a consumer makes with a VPN, and it is the reason a serious verification regime pairs latency with other signals, cross-checks against many landmarks, and, as the Georgetown analysis recommends, retains a small number of physical inspections whose purpose is precisely to make the proxying bet dangerous. The research literature on proof of location has pushed further, proposing protocols in which a device must respond so quickly to unpredictable challenges that relaying is detectable by the delay it necessarily adds, a distance-bounding idea that turns the speed-of-light asymmetry into an active test rather than a passive observation. None of this makes location unspoofable. It makes spoofing a continuing operational burden that must be maintained perfectly and forever, against a verifier that can change its challenges, add landmarks and send inspectors, which is a very different proposition from defeating a controls regime once on paper.

The deeper limit is not spoofing but abstention. Every technique in this section assumes the chip is online and running the agent. The evader's strongest move is not to lie about location but to decline to report at all, and against that, location verification as currently built has no answer, which is exactly why the policy discussion keeps drifting towards making the reporting, and ultimately the permission, a condition of operation rather than an option.

### 4.5. How good, and how evadable

The limits matter, and they cut both ways. The most thorough public assessment, a 2026 study from Georgetown's Center for Security and Emerging Technology, compared ping-based location verification against physical inspection across more than ten million simulated enforcement scenarios and found the ping-based approach the more cost-effective in most of them, while recommending that it be paired with a small number of physical inspections rather than relied on alone. The same body of work estimates that a landmark network of a few dozen to a few hundred servers, figures in the range of 39 to 500 have been cited, would give region-to-city-level accuracy for the highest-end chips that are the priority. IEEE Spectrum's reporting notes the authors' own caution that such a system is costly and best aimed only at the most capable hardware.

Against a determined evader, the weaknesses are structural, not incidental. Location verification as currently built is customer-installed and opt-in: the monitoring runs because the operator chose to run it, so a smuggler simply does not install it, or keeps the chips off the network entirely, and a scheme that depends on the watched party agreeing to be watched is weak against exactly the people it is aimed at. Detection, by the nature of the method, happens only after a chip has moved and checked in from its new home, not before. It places a chip to a country, not to an address. The host-side agent is open source, so its behaviour can be audited, but the GPU firmware it talks to is not, so the behaviour of the chip itself still rests on the vendor's word. These are the reasons that, today, location verification is a monitoring tool and not a control one. The whole weight of the policy question is how stable that distinction is, and it is to that question we now turn.

### 4.6. Beyond latency: combining signals

It would understate the method to leave it at round-trip times, because latency is only the most physical of several signals a verifier can combine, and combination is what turns a loose estimate into a confident one. The attestation exchange itself carries incidental evidence of place. The network path a chip's traffic takes, the sequence of intermediate hops, the autonomous-system numbers it traverses, the address block its host presents, all correlate with geography, and a verifier that already holds a signed, unforgeable identity for the chip can bind these softer hints to a hard name rather than to a spoofable one. Conventional address-based geolocation, the same databases that place a website visitor in a city, is weak on its own and easily defeated by a proxy, but it is a useful cross-check against a latency estimate: when the two agree, confidence rises; when they disagree, the disagreement is itself a flag worth an inspection.

There is a subtler point about identity-bound location that is easy to miss. Ordinary Internet geolocation places an address, and an address can be rented, tunnelled or shared, so the thing it locates is slippery. Attestation-bound location places a specific, certified chip, so the question shifts from where is this address, which is answerable only loosely, to where did this exact, named chip just check in from, which is answerable with the chip's cooperation and, crucially, auditable after the fact through the signed tokens it leaves behind. A fleet operator, or a regulator reading the operator's records, is not trying to pin an anonymous packet to a map; they are watching a known serial number announce itself over time, and a chip that was in one country last month and appears to be in another this month is visible as a change even if neither reading is precise to the kilometre. Movement is often easier to detect than absolute position, and movement across a controlled border is exactly what the policy cares about.

None of this escapes the abstention problem of the previous section: every signal still assumes the chip is online and reporting. But it does mean that the accuracy figures quoted for pure latency are a floor, not a ceiling, and that a mature verification regime would not rely on any single method. It would fuse the physical bound from latency, the topological hints from routing, the address-based cross-check and the historical record of a named chip's check-ins into a composite judgement with a stated confidence, which is precisely the form of evidence an enforcement decision, or a permission decision, would be built to consume.

## 5. Front door, not backdoor

### 5.1. Why it is not a backdoor

It is worth being precise, because the word backdoor is doing a great deal of work in this debate, and precision cuts against alarmism as much as for it. In the technical sense, attestation is not a backdoor. A backdoor is a hidden way in: a means to run code, read data or take control that the owner neither knows about nor authorised. Attestation is none of those. It is documented rather than concealed, specified in public standards and vendor manuals down to the error codes. It is read-only: it proves identity and firmware state and offers no way to execute anything, exfiltrate anything or switch anything off. And it is initiated locally: software on the host must ask the chip for the evidence and then choose to send the result somewhere, so a machine with no agent installed, or no network, reports nothing at all. It is, mechanism for mechanism, the same facility that sits in the TPM of a laptop, the secure enclave of a phone and the trust chip of a games console, and its first purpose is the benign one, set out in Section 3.7, of letting an owner or a cloud tenant check that their hardware is what it claims to be.

### 5.2. Why the concern is well founded

And yet the concern is not misplaced, because the very properties that make attestation safe for an owner make it powerful for a controller. Each chip now has a unique, unforgeable identity: the device certificate is, in effect, a serial number that cannot be filed off or faked, which is exactly what a sanctions regime needs and exactly what a chip's current holder cannot escape. Verification runs through the vendor's servers and ends in a signed, portable token: the act of proving a chip genuine is also a moment at which the chip, or its host, reaches out to Nvidia and a durable attestation of which chip checked in, and from roughly where, is produced and can be handed to anyone. A capability sold as fleet health is, in the same breath, a census. And the firmware is closed: what the chip actually does when asked to attest, and what else it might do, rests on Nvidia's word, because the one link in the chain that cannot be independently audited is the silicon's own code. Each of these is a virtue from one angle and a lever from another, and they are the same features seen twice.

### 5.3. Precedents: when identity becomes control

The worry that an identity mechanism will harden into a control one is not speculative, because it has happened repeatedly with other technologies, and the pattern is always the same: a facility introduced to establish who or what a device is acquires, by the addition of a central list and a check, the power to decide whether it may function. Three examples make the shape plain. Mobile phones carry an IMEI, a unique equipment identity meant simply to name the handset; layered on top of it, the shared equipment-identity registers let carriers block a listed device from the network, turning a serial number into an on-off switch. High-definition video links use HDCP, whose device keys were meant to authenticate a display; the same scheme includes key revocation, so that a device whose keys appear on a revocation list is refused content. Remote attestation of PCs, the direct analogue of what Nvidia's GPUs now do, was introduced so that a service could verify a client's integrity, and is now routinely used to gate access: no acceptable attestation, no entry. Two further examples show the pattern reaching into things people own outright. Apple's Activation Lock ties a device to an account identity so that a stolen phone cannot be reused; the same binding means a device can be rendered inert from afar, and the line between anti-theft and remote disablement is a database entry. Modern tractors and cars authenticate their own parts and phone home for entitlements, so that a manufacturer can, and sometimes does, refuse to let hardware the owner has paid for operate without the vendor's continuing blessing, a practice that drove the right-to-repair backlash. In each case nothing secret was added. An existing, documented identity mechanism was pointed at a list, and the list was given teeth. That is the precise manoeuvre available with GPU attestation, and the policy described next is the list acquiring its teeth.

### 5.4. The policy ratchet and the feedback loop

That last point is where the plumbing meets the law. The United States Chip Security Act, introduced in May 2025 as H.R. 3447 in the House, with a Senate companion, S. 1705, led by Senator Tom Cotton, would make location verification compulsory: within 180 days of enactment the Secretary of Commerce would have to require covered advanced chips to carry location-verification mechanisms, by feasible and appropriate techniques, with a later phase to assess further measures such as anti-tamper. The bill does not mandate a particular technology, but the Center for Security and Emerging Technology analysis discussed above judged the ping-based method the most cost-effective of the plausible options. The House sponsors are a bipartisan group, Representatives Bill Huizenga and Bill Foster among them, alongside the leaders of the House select committee on China. The industry is not uniformly keen: the Semiconductor Industry Association, of which Nvidia is a member, has objected to location mandates as untested and potentially unworkable, even as Representative Foster has said that Nvidia told him privately that location verification is technically feasible, something Nvidia has not confirmed in public.

None of this is law yet. The Chip Security Act passed the House Foreign Affairs Committee in spring 2026; as of mid-November 2026, chip provisions ride as amendments on the Senate's version of the defence authorisation for the coming fiscal year, expected to be taken up in December, after a related measure, the GAIN AI Act, which would require chipmakers to prioritise American customers before exporting, passed the Senate as an amendment but was left out of the House version and looked unlikely to survive to the final text. Comparable language has failed before. The direction, however, is consistent, and the heightened political attention of late 2026 has only sharpened it.

The step that would change the character of the whole system is the one Representative Foster has signalled for a future bill: a licence to operate an advanced GPU at all, checked before the chip is allowed to connect. To see why that is the decisive step, it helps to borrow the language of control systems. Today the phone-home is an open loop: the chip senses and reports where it is, and nothing reaches back to it; the output of the sensing does not act on the sensed thing. It is a dashboard, an instrument that shows a

reading and does nothing with it. A licence checked before a chip may connect closes the loop. It wires the report to a decision, grant or refuse, and the decision back to the chip, as permission to run or a refusal to run. A closed feedback loop from sensing to action, with the authority to withhold action, is exactly what distinguishes a control system from a dashboard, and it would be built not by adding a secret mechanism but by making a published one mandatory and feeding its output into a gate. It would be a kill switch in all but name, assembled entirely from parts that are, individually, already in plain sight.

So the honest reading is double. The plumbing is neutral, and the people who built it built it for defensible reasons: identity and integrity are the bedrock of trustworthy computing, confidential computing genuinely needs them, and an owner who cannot be sure of the hardware beneath a workload is right to want them. But it is, part for part, the plumbing a control system would require: an unforgeable name for every chip, a channel by which each chip announces itself, a signed token that makes the announcement portable, and a firmware root that only the vendor can see into. The backdoor the regulators keep asking about would be redundant. The front door is already there, and the law is reaching for the latch.

### 5.5. Not only one government

It would be a mistake to read this as a story about a single country. The mechanism is owned by a company, and a company is subject to whichever governments can compel it, so the latch, once it exists, can in principle be reached by more than one hand. A control channel built to enforce one state's export policy is, in its bones, a general-purpose control channel; the firmware root that answers to the vendor answers, through the vendor, to the vendor's home jurisdiction today and could answer to others under different law tomorrow. This is not a prediction that it will be misused. It is the observation that a buyer anywhere, including inside allied countries and inside the vendor's own market, is accepting a dependency whose ultimate terms are set outside any contract they signed, by legislatures they do not vote in. The European debate about digital sovereignty has made the same point about cloud and software for a decade; the chip is the same argument one layer down, and a harder layer to re-source. For an organisation in the Netherlands, or anywhere outside the handful of states that make and govern these parts, the relevant question is not whose policy the switch enforces this year but that the switch exists and that its wiring runs abroad.

## 6. The case for the system, taken seriously

It would be a poor paper that described this architecture only as a threat, because the case for building it is real and is made by serious people, and the buyer's decision in the next section is better for having felt its force. So before turning to what the arrangement costs, we set out, as strongly as we can, why one might want it.

The first argument is non-proliferation, and it is the plainest. The most capable AI accelerators are, increasingly, treated as strategic materials, closer to enrichment centrifuges than to laptops, on the view that a sufficiently large concentration of them is an input to capabilities a state may not wish a rival, or a non-state actor, to possess. If one accepts that premise even partially, then knowing where the material is becomes a legitimate interest in the same way that knowing where fissile material is does, and a mechanism that lets a chip prove its location is a less invasive way of knowing than the alternatives. The honest comparison is not between location verification and perfect freedom; it is between location verification and the blunter instruments that its absence invites, total bans on whole categories of chip, mandatory physical inspection regimes, or the collapse of exports into a handful of trusted operators. A targeted, auditable, mostly-software check may be the least bad enforcement tool on the menu, and the Georgetown analysis reaches broadly that conclusion on cost-effectiveness grounds.

The second argument is that verifiability is better done in the open than in the dark. A regime of secret tracking hardware, of the kind the backdoor question imagines, would be unauditable by construction and corrosive of all trust in the supply chain. A regime built on published protocols, documented libraries, open host agents and standard tokens is one that researchers, customers and other governments can inspect, test and criticise, as this paper does. If some form of compute governance is coming, and the political momentum suggests it is, then its being assembled from parts that are in plain sight is a feature and not only a danger: the same transparency that lets us describe the control system lets everyone else hold it to account. A mechanism you can read is a mechanism you can constrain.

The third argument is narrower and more technical. Attestation, considered purely as what it was built for, genuinely raises the security floor. Confidential computing lets an organisation run sensitive work on infrastructure it does not own without exposing that work to the operator, which is a real enabler for regulated industries, for cross-border data arrangements and for anyone renting rather than owning their compute. Verifiable firmware integrity defends against a class of supply-chain and tampering attacks that are otherwise very hard to detect. A buyer who reflexively disabled every attestation feature to escape the control concern would be giving up protections they actually want. The mechanism's dual-use character cuts both ways: the same properties that worry us are, in their intended setting, exactly the properties that make the hardware more trustworthy, not less.

None of this dissolves the concern of Section 5, and we do not pretend it does. The reply to the non-proliferation argument is that strategic-material regimes come with treaties, inspectorates and appeal, whereas a vendor-operated switch comes with terms of service; the reply to the transparency argument is that an open mechanism is no less effective as control for being readable, and readability is cold comfort to whoever is on the wrong side of the gate; the reply to the security-floor argument is that one can want attestation for integrity while refusing it as permission, and that the whole fight is over which of those two it becomes. The point of stating the case for the system is not to concede the argument but to locate it precisely. The disagreement is not about whether the mechanisms are useful, they are, nor about whether they are open, they are, but about whether the output of a check the owner cannot refuse should ever

be wired to a gate the owner cannot open. That is a question of governance, not of engineering, and it is the buyer, not only the regulator, who should have a view on it.

## 7. What it means for the buyer

### 7.1. Dependence has a shape

The China story is the sharp edge of a duller, broader fact, and it is the broader fact that should concern anyone building on this hardware. Compute has become infrastructure, in the way that electricity and the network are infrastructure, and the mechanisms now being standardised decide who holds the switches. A buyer of AI hardware is not only buying silicon; they are buying into a chain of trust whose root is held by the vendor and, increasingly, overseen by the vendor's government. For most owners most of the time this is invisible and harmless. It becomes visible the moment interests diverge: a change of export policy, a sanctions listing, a commercial dispute, a decision taken in a capital on the other side of the world. The lesson of the precedents in Section 5.3 is that the distance between a monitoring facility and a control one is a single administrative decision, and that the decision is not the buyer's to make.

### 7.2. The questions to ask

The lesson is not that attestation is sinister; it is that dependence has a shape, and it is worth seeing the shape clearly before committing to it. An organisation that runs on infrastructure it cannot inspect, cannot relabel and cannot operate without a distant party's continued permission has taken on a particular kind of risk, and should at least price it. The questions are practical ones, and they can be asked of any vendor at purchase. Which parts of the stack phone home, to whom, and can that reporting be disabled without disabling the hardware? What signed artefacts about our chips leave our premises, and who can hold them? What stops working, and on whose decision, if the relationship sours or the law changes? Where is the firmware root, who can see into it, and what is the plan if that trust is withdrawn? What is the second source, if there is one, and how quickly could we move to it? These are the same sovereignty questions that apply to a cloud contract, a software licence or an offshore dependency, asked one layer lower, of the chip itself.

To make the questions less abstract, consider a mid-sized organisation deciding how to run a sensitive model, with three options on the table. It can rent capacity from a hyperscaler, where the hardware, its attestation and its reporting are entirely the provider's, and the organisation's leverage is a contract and an exit clause. It can buy its own accelerators and run them in a colocation facility, owning the silicon but inheriting whatever phone-home behaviour the firmware carries and whatever the law comes to require of it. Or it can build on a smaller, more inspectable stack, accepting less raw performance in exchange for knowing where the switches are. There is no universally right answer among these; a bank with strict data-residency duties, a startup optimising for cost, and a public body wary of foreign leverage will weigh them differently. The discipline is to make the choice with the dependency in view rather than discover it later. The failure mode is not choosing the hyperscaler; it is choosing it without having asked what stops, and on whose word, if the political weather changes. A dependency priced in at purchase is a managed risk. The same dependency discovered during an incident is a crisis.

### 7.3. Nashua's position

We take the view, across the infrastructure we build and run, that control you cannot inspect is control you do not really have, and that the right default for anything load-bearing is to know exactly where the switches are and whose hand can reach them. That view does not resolve into a slogan about avoiding a vendor; the leading AI accelerators come from a small number of suppliers, and attestation, in its benign form, is something we actively want, because verifiable identity and integrity are how an operator earns a tenant's trust. It resolves instead into a discipline: to map the phone-home paths in anything we deploy, to prefer arrangements where the reporting is ours to switch off, to keep the sovereignty questions above on the purchase checklist rather than the post-mortem, and to be candid with the people who rely on us about which dependencies are inspectable and which rest on a distant party's goodwill. Attestation is a good mechanism, built for good reasons. It is also a standing reminder that in computing the front door and the back door are often the same door, and that the only durable protection is to know whose hand is on the latch.

## 8. Conclusion

The regulators pressing Nvidia are right to be suspicious and are looking in the wrong place. There is probably no hidden backdoor in an Nvidia GPU, and there does not need to be. The chip already carries an identity it cannot deny and a means to prove it, fused into the silicon, drawn out over a published protocol, checked against reference values, and returned as a signed, portable token, all of it documented. The industry is now standardising the second half, a way to tell where each chip sits, under the pressure of export control, and wiring the two together in shipping products. None of the pieces is a backdoor. Assembled, mandated and closed into a feedback loop, they are the architecture of remote control, in the open. The useful response is not alarm but clarity: to read the documentation that is already public, to understand that identity, attestation and location verification are one system seen from three sides, and to decide, deliberately and in advance, how much of one's own infrastructure should depend on a door whose latch is held by someone else.

The deeper lesson reaches past Nvidia and past this moment in export policy. As computing concentrates into a few kinds of hardware made by a few firms under the eye of a few governments, the question of who can see into, name and ultimately gate a machine becomes a question about autonomy itself, for companies and for countries. Attestation did not create that question; it made it legible, by giving every chip a voice with which to announce exactly what it is and, soon, where. We would rather that voice be understood than feared, because understanding is the precondition of control, and the alternative to understanding is not safety but surprise. The mechanisms are open. The choice of what to build on top of them is still, for now, open too, and that is precisely why it is worth arguing about in public while the choice remains to be made.

## Appendix A. A worked attestation exchange

To make the front door concrete, here is the sequence by which a verifier establishes, from nothing, that it is talking to a specific genuine Nvidia GPU in a known-good firmware state. The steps correspond to the mechanisms in Section 3; the names follow the SPDM specification and Nvidia's documented tooling. The reader who wants the one-paragraph version may read the bold lead-ins alone.

**1. Enumeration.** A collector on the host, the attestation Corelib, discovers the GPUs present and, for each, prepares to open an authenticated channel to the on-die security processor that holds the chip's identity. Nothing is trusted yet; the host is treated as potentially hostile to the evidence it is about to relay.

**2. Capability negotiation.** Over SPDM, the verifier and the device agree on versions, supported capabilities and cryptographic algorithms (`GET_VERSION`, `GET_CAPABILITIES`, `NEGOTIATE_ALGORITHMS`). This fixes the rules of the rest of the conversation and prevents a downgrade to weaker cryptography.

**3. Certificate retrieval.** The verifier requests the device's certificate chain (`GET_DIGESTS`, `GET_CERTIFICATE`) and validates it upward, link by link, to Nvidia's certificate authority. A chain that does not resolve to the expected root is rejected outright; this is the check behind an error such as 4007, `INVALID_CERTIFICATE_CHAIN`.

**4. Proof of possession.** The verifier sends a fresh random nonce in a `CHALLENGE`, and the device signs it with the private identity key that never leaves the silicon. A valid signature proves the device holds the key the certificate names, so it is the specific part the certificate describes, not a replay or a clone.

**5. Measurement collection.** The verifier requests the firmware measurements (`GET_MEASUREMENTS`), again bound to a nonce. The device returns its measurement log: a set of blocks, each the hash of one firmware or configuration component, signed as a whole. Run inside a secure session, this traffic is protected from the host relaying it.

**6. Reference comparison.** The verifier compares each reported measurement against the manufacturer-signed reference integrity manifest for the declared firmware version. Identity has now been joined to integrity: the verifier knows which chip, and in what state.

**7. Verdict and token.** If certificate, signature, freshness and measurements all check out, the verification service (NRAS) issues a signed Entity Attestation Token, an RFC 9711 artefact carried as a JSON Web Token, stating what was attested. A separate revocation check confirms the device certificate is still valid.

**8. Onward use.** The token is portable. Whoever holds the verifier's public key can trust it without repeating the exchange, which is what lets an attestation taken on one machine become a credential presented to another, to a cloud control plane, to an orchestration system, in principle to a regulator. Step eight is where an identity check becomes, with one policy decision, a permission.

## Appendix B. A short glossary of the stack

**Attestation.** A device proving, cryptographically, both its identity and the integrity of the firmware it is running, in a form a remote party can verify.

**Device certificate.** A certificate, issued by the manufacturer's certificate authority, binding a chip's public identity key to the specific part; the private half is fused into the silicon and never exported.

**DICE (Device Identifier Composition Engine).** The Trusted Computing Group pattern in which a per-device Unique Device Secret (UDS) is combined with a measurement of each code layer to derive a Compound Device Identifier (CDI), chaining identity to software state.

**SPDM (Security Protocol and Data Model).** The DMTF standard, DSP0274, defining the request-and-response handshake by which a verifier authenticates a hardware component and collects its measurements. Nvidia uses version 1.1 for measurement collection.

**Nonce.** A fresh random number the verifier supplies so that the device must sign something new, defeating replay of a captured earlier response.

**Measurement block.** One entry in a device's signed measurement log, the hash of a single firmware or configuration component, checked against an expected reference value.

**RIM (reference integrity manifest).** A manufacturer-signed document stating the measurements a genuine chip at a given firmware version should report, against which a verifier compares the evidence.

**Secure session.** An authenticated, encrypted SPDM channel, set up by a key exchange, that protects attestation traffic from the host relaying it, on the assumption that the host itself may be hostile to the evidence.

**Security processor.** The small, dedicated on-die controller that holds a GPU's fused secret, performs its signing and speaks the attestation protocol; the innermost trust root, and the one component that cannot be audited from outside.

**Performance density.** TPP divided by die area in square millimetres, added to the United States export tests in 2023 so that a chip could not clear the throughput cap merely by spreading the same compute across more silicon.

**NVML (Nvidia Management Library).** The documented C library beneath `nvidia-smi`, exposing a card's identity and telemetry (serial number, device identifiers, firmware version, utilisation, power, temperature).

**Attestation Corelib.** Nvidia's library for collecting a GPU's attestation evidence over SPDM without requiring confidential computing mode; it gathers evidence but does not judge it.

**NRAS (Nvidia Remote Attestation Service).** The verification service that validates a GPU's evidence against reference values and certificate status and returns a signed verdict.

**EAT (Entity Attestation Token).** The IETF standard, RFC 9711, for a signed, portable token conveying an attestation result; here carried as a JSON Web Token.

**Confidential computing.** A mode in which a GPU protects a tenant's code and data from the host operating it, with attestation as the means by which the tenant verifies the protection before trusting the machine.

**TPP (Total Processing Performance).** The aggregate-throughput measure at the centre of United States export thresholds, defined in Equation (1); paired since 2023 with a performance-density test.

**Location verification.** Inferring where a chip physically is from the time its signals take to reach servers at known positions, bounded by the speed of light as in Equation (2); latency-based and accurate to roughly country level.

## Appendix C. A timeline of the controls

The mechanisms and the policy have moved in step. The sequence below gathers the dates used through the paper into one line of sight; each entry is drawn from the reporting cited in the references, and dates are as reported at the time of writing.

**October 2022.** The United States Bureau of Industry and Security issues its first advanced-computing rule, drawing the export line with two parameters, raw compute and interconnect bandwidth. Nvidia responds with the A800 and H800, cut at the interconnect to clear the bar.

**October 2023.** The rule is updated: the interconnect parameter is dropped, a performance-density test is added, and ECCN 3A090 takes its current two-test shape (Table 1). The A800, H800 and the consumer RTX 4090 are caught.

**2024 onward.** Nvidia offers the further cut-down H20 and L20. China's own regulators later discourage domestic firms from buying the H20, pressing instead towards local accelerators.

**January 2025.** The RTX 5090 D, a China gaming card with full gaming and reduced AI performance, is shown at CES.

**May 2025.** The Chip Security Act is introduced as H.R. 3447, with a Senate companion, requiring location verification on covered chips within 180 days of enactment.

**Mid 2025.** Blackwell-generation China variants, the B30A and RTX 6000D, are reported but not confirmed as shipping. The H20 is briefly permitted again, then discouraged by Beijing.

**October 2025.** The GAIN AI Act, requiring chipmakers to prioritise American customers, passes the Senate as an amendment to the defence authorisation; it is later left out of the House version.

**December 2025.** Nvidia announces Fleet Intelligence, the fleet-management service that infers GPU location from attestation data and network latency; location-verification work is reported in the same period.

**Late 2025 to early 2026.** The United States approves export of the full-strength H200 on tight terms, routed through the United States for inspection and carrying a levy.

**Spring 2026.** The Chip Security Act passes the House Foreign Affairs Committee.

**November 2026.** Reuters reports roughly 600,000 H200s stranded in Hong Kong awaiting Chinese import approval. On 17 November, Nvidia tells investors it is no longer shipping licensed Hopper data-center products to China, and chip-security provisions are reported riding as amendments on the Senate defence authorisation, expected to be taken up in December.

## Appendix D. Common objections, briefly answered

Several reasonable objections recur whenever this argument is made in public, and answering them in one place sharpens the paper's claims by marking their edges.

**Is this not just a TPM, which nobody frets about?** Mechanically, yes, and that is rather the point. The worry is not that GPU attestation is a novel or sinister technology; it is that the same technology, at the scale and strategic weight of frontier accelerators and under the specific export-control pressure described here, is being pointed towards an end, mandatory permission to operate, that we do not accept for the TPM in a laptop. The mechanism is ordinary. The proposed use is not.

**Can the owner not simply turn attestation off?** Today, on an ordinary machine, largely yes: nothing is reported if no agent runs and nothing is wired to a gate. That is exactly why the paper distinguishes the current open loop from the proposed closed one. The whole thrust of the policy discussion is to remove the ability to decline, by making the reporting, and in the stronger proposals the permission, a condition of operation rather than a setting.

**Is a consumer gaming card affected?** Not in the same way. The attestation and location machinery discussed here is a data-center GPU facility; the China gaming variants in Table 2 are cut on AI performance precisely to fall outside the controlled tier. The concern is about frontier training and inference hardware, not the card in a home PC.

**Can Nvidia already switch my chip off from afar?** On the public evidence, no. Attestation is read-only and carries no command path back to the silicon, as Section 3.8 sets out. The argument is about what becomes cheap to build once a mandatory, portable proof of identity and location exists, not about a disable command latent in the chip today.

**If the firmware is closed, how do we know it does only what is documented?** We do not, independently, and that is a stated limitation rather than an accusation. The security processor's firmware is the one link in the chain that cannot be audited from outside, so its behaviour rests on the vendor's word and the vendor's jurisdiction. That irreducible trust is a reason for buyers to weigh the dependency, not evidence that it is being abused.

**Does location verification find my chip at its desk?** No. The method is accurate to roughly country level, sometimes regional, as the physics in Section 4.1 dictates. It is built to answer whether a chip has crossed a border it should not have, not to place it in a building.

**Why build any of this if a smuggler can stay offline?** Because most chips, most of the time, are online and doing useful work, and a regime that forces serious evaders into permanent disconnection or permanent deception has already raised the cost of diversion considerably, even if it cannot stop the most determined. The tool is a deterrent and an audit, not a seal, and it should be judged as one.

## References

[1] Reuters, via bdnews24 (18 November 2026). *Nvidia stops shipping US-approved AI chips to China.* https://bdnews24.com/business/pyentjtvr9

[2] Fox Business (2026). *Nvidia halts China AI chip shipments*

*despite US approval*.
https://www.foxbusiness.com/fox-news-tech/nvidia-halts-china-ai-chip-shipments-despite-us-approval

[3] Tom's Hardware (2026). *US eases Nvidia export restrictions: H200 cleared for China under tight controls*. https://www.tomshardware.com/tech-industry/semiconductors/us-eases-nvidia-export-restrictions-h200-cleared-for-china-under-tight-controls

[4] DataCenterDynamics (2026). *Almost 600,000 Nvidia H200s stranded in Hong Kong as Chinese government stalls on import approvals*. https://www.datacenterdynamics.com/en/news/almost-600000-nvidia-h200s-stranded-in-hong-kong-as-chinese-govt-stalls-on-import-approvals-report/

[5] Bureau of Industry and Security, US Department of Commerce (25 October 2023). *Implementation of additional export controls: certain advanced computing items*. Federal Register. https://www.federalregister.gov/documents/2023/10/25/2023-23055/

[6] Center for Security and Emerging Technology, Georgetown University (2023). *The BIS 2023 export control update: an explainer*. https://cset.georgetown.edu/article/bis-2023-update-explainer/

[7] TrendForce (2025). *GeForce RTX 5090 D for China*. https://www.trendforce.com/news/?p=31706

[8] Notebookcheck (2025). *Nvidia readies new Blackwell-based accelerators tailored for China under US export limits*. https://www.notebookcheck.net/Nvidia-readies-new-Blackwell-based-accelerators-tailored-for-China-under-US-export-limits.1091993.0.html

[9] Finimize, reporting Bloomberg (2025). *Nvidia AI chips were tracked to a Shanghai data center*. https://finimize.com/content/bloomberg-nvidia-ai-chips-were-tracked-to-a-shanghai-data-center

[10] Windows Central (2025). *Report: AI chip trackers. Nvidia responds*. https://www.windowscentral.com/hardware/nvidia/report-ai-chip-trackers-nvidia-responds

[11] Reuters, via AOL (December 2025). *Exclusive: Nvidia builds location-verification technology*. https://www.aol.com/articles/exclusive-nvidia-builds-location-verification-044501961.html

[12] Outlook Business (2025). *Nvidia tests optional software to verify where its AI chips are running amid smuggling claims*. https://www.outlookbusiness.com/start-up/deeptech/nvidia-tests-optional-software-to-verify-where-its-ai-chips-are-running-amid-smuggling-claims

[13] Tom's Hardware (2026). *Nvidia details new software that enables location tracking for AI GPUs*. https://www.tomshardware.com/pc-components/gpus/nvidia-details-new-software-that-enables-location-tracking-for-ai-gpus-opt-in-remote-data-center-gpu-fleet-management-includes-power-usage-and-thermal-monitoring

[14] NVIDIA developer blog (2025). *Introducing NVIDIA Fleet Intelligence for real-time GPU fleet visibility and optimization*. https://developer.nvidia.com/blog/introducing-nvidia-fleet-intelligence-for-real-time-gpu-fleet-visibility-and-optimization/

[15] MIT Technology Review (12 November 2026). *Nvidia chip tracking*. https://www.technologyreview.com/2026/11/12/1148483/nvidia-chip-tracking

[16] E. Feldgoise, J. Miller and D. Dohmen, Center for Security and Emerging Technology (2026). *Tracking AI chips: assessing location verification as an export control enforcement tool*. https://cset.georgetown.edu/publication/tracking-ai-chips/

[17] IEEE Spectrum (2026). *Find my AI chips: the US plans to track GPUs abroad*. https://spectrum.ieee.org/find-my-ai-chip

[18] B. Gueye, A. Ziviani, M. Crovella and S. Fdida (2004). *Constraint-based geolocation of Internet hosts*. ACM Internet Measurement Conference. https://www.cs.bu.edu/fac/crovella/paper-archive/imc04-geolocation-full.pdf

[19] B. Wong, I. Stoyanov and E. G. Sirer (2007). *Octant: a comprehensive framework for the geolocalization of Internet hosts*. USENIX NSDI. https://cs.uwaterloo.ca/~bernard/octant-nsdi.pdf

[20] CSIS (2025). *The architecture of AI leadership: enforcement, innovation and global trust*. https://www.csis.org/analysis/architecture-ai-leadership-enforcement-innovation-and-global-trust

[21] 119th Congress (2025). *Chip Security Act, H.R. 3447*. https://www.govinfo.gov/content/pkg/BILLS-119hr3447ih/html/BILLS-119hr3447ih.htm

[22] CSIS (2025). *The GAIN AI Act will undermine the global competitiveness of US AI chipmakers*. https://www.csis.org/blogs/perspectives-innovation/gain-ai-act-will-undermine-global-competitiveness-us-ai-chipmakers

[23] Nextgov/FCW (October 2025). *AI export control bill passes Senate as NDAA amendment*. https://www.nextgov.com/artificial-intelligence/2025/10/ai-export-control-bill-passes-senate-ndaa-amendment/408762/

[24] CNBC (17 November 2026). *Jensen Huang scandal fuels bipartisan push to block AI chip sales to China*. https://www.cnbc.com/2026/11/17/congress-ai-china-jensen-huang-chips.html

[25] Distributed Management Task Force (2025). *Security Protocol and Data Model (SPDM) specification, DSP0274*. https://www.dmtf.org/dsp/DSP0274

[26] Trusted Computing Group. *Device Identifier Composition Engine (DICE)*. https://trustedcomputinggroup.org/what-is-a-device-identifier-composition-engine-dice/

[27] L. Lundblade, G. Mandyam, J. O'Donoghue and C. Wallace (April 2025). *The Entity Attestation Token (EAT)*. IETF RFC 9711. https://www.rfc-editor.org/info/rfc9711

[28] NVIDIA (2026). *NVIDIA attestation documentation*. https://docs.nvidia.com/attestation/

[29] NVIDIA (2026). *NVIDIA Remote Attestation Service (NRAS): introduction*. https://docs.nvidia.com/attestation/cloud-services/latest/nras/nras_introduction.html

[30] NVIDIA (2023). *NVIDIA Hopper confidential computing: whitepaper*. https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/HCC-Whitepaper-v1.0.pdf

[31] NVIDIA (2026). *NVIDIA Management Library (NVML)*. https://developer.nvidia.com/management-library-nvml

[32] NVIDIA (2026). *NVML API reference*. https://docs.nvidia.com/deploy/nvml-api/latest/nvml-api-reference.html

[33] NVIDIA (2026). *Attestation Corelib: introduction*. https://docs.nvidia.com/attestation/corelib/latest/introduction.html

[34] Office of Rep. Bill Foster (15 May 2025). *Foster, Huizenga, Moolenaar and Krishnamoorthi introduce bill to stop smuggling of advanced AI chips*. https://foster.house.gov/media/press-releases/foster-huizenga-moolenaar-krishnamoorthi-introduce-bill-stop-smuggling

[35] The Register (15 May 2025). *Congress wants location tracking built into high-end GPUs*. https://www.theregister.com/2025/05/15/gpu_tracking_house/

[36] FDD Action (25 March 2026). *H.R. 3447, the Chip Security Act*. https://www.fddaction.org/endorsements/2026/03/25/h-r-3447-the-chip-security-act/

[37] 119th Congress (2025). *S.Amdt. 3505 to S. 2296 (the GAIN AI Act)*. https://www.congress.gov/amendment/119th-congress/senate-amendment/3505/text

[38] Nextgov/FCW (December 2025). *Bill prioritizing American customers for AI chips not expected to make it into final NDAA*. https://www.nextgov.com/policy/2025/12/bill-prioritizing-american-customers-ai-chips-not-expected-make-it-final-ndaa-sources-say/409920/

[39] ASF and GeoComply (2025). *Leveraging existing location verification technology to adapt export controls to the AI age*. https://www.geocomply.com/wp-content/uploads/ASF-and-GeoComply-Leveraging-Existing-Location-Verification-Technology-to-Adapt-Export-Controls-to-the-Artificial-Intelligence-Age.pdf