GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

Confidential GPU Attestation: H100 to B300 Matrix

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-09-24 · confidential gpu attestation · h100 attestation · h200 attestation · b200 attestation · b300 attestation · nvidia confidential computing · gpu evidence · tee verification · confidential containers

Original article: https://gpusmith.com/articles/en/confidential-gpu-attestation-deployment-matrix

In this article

Executive Summary

As of **24 September 2026**, the answer to “confidential GPU attestation H100 H200 B200 B300” depends on a complete platform, not a GPU name. NVIDIA's current Confidential Containers reference architecture lists **H100** and **H200** for single GPU passthrough, their protected PCIe configurations for multi GPU passthrough, **B200** and **HGX B300** for both, and **RTX PRO 6000 Blackwell Server Edition** for single GPU passthrough. (docs.nvidia.com) Blackwell B300 entered that reference architecture in its **1.1.0** release; the separate self hosted virtual machine reference implementation still does not list B300. (docs.nvidia.com) (docs.nvidia.com) These are distinct support scopes, not contradictory statements about the silicon.

A purchase order should name the GPU form factor and system, CPU trusted execution environment (TEE), confidential computing mode, hypervisor or container path, release, and verifier. NVIDIA's container matrix validates AMD Genoa or Milan with **SEV-SNP**, or Intel Emerald Rapids or Granite Rapids with **TDX**, on specified Ubuntu and kernel versions. (docs.nvidia.com) A GPU configured in confidential mode has not, by that act alone, been cryptographically verified or authorized to receive a model key. The approval point is an attestation result that binds fresh CPU and GPU evidence to an explicit policy, followed by a relying party's controlled release decision. (www.rfc-editor.org) (confidentialcontainers.org)

The cloud examples show why the exact service matters. Google Cloud documents an **A3 High H100 plus Intel TDX** confidential virtual machine, made generally available in **July 2025**, and a **G4 RTX PRO 6000 plus AMD SEV** path. (docs.cloud.google.com) (docs.cloud.google.com) Azure documents an **NCCadsH100v5** path with AMD SEV-SNP and an H100, rather than a blanket promise for every GPU in a provider catalog. (learn.microsoft.com) Cloud tokens can also have narrower meanings: Google says one `cc_mode` claim attests the driver only, while device claims separately identify VBIOS and device identity. (docs.cloud.google.com)

For acceptance, capture the platform inventory, release-specific firmware and driver versions, fresh challenge nonce, raw CPU and GPU evidence, endorsement and certificate chains, reference values, verifier identity and version, per-device result, relying-party policy, key-release decision, and a deliberate rejection test. (docs.amd.com) (confidentialcontainers.org) Refresh this package after firmware, driver, image, policy, or verifier changes and when the result expires. The publicly retrieved material establishes supported

combinations and verification fields, but supplies no comparable, independent H100-versus-H200-versus-B200-versus-B300 attestation-overhead benchmark. Procurement should therefore ask for a workload-specific measurement, with the test method and boundary attached, rather than treat an encryption claim as a throughput result.

Introduction and Background

Confidential GPU computing protects selected workload data and code while in use inside a hardware-backed TEE, extending a CPU confidential virtual machine to an accelerator. The useful buying question is narrower than whether the processor carries a confidential-computing logo: can the proposed configuration produce evidence that an independent party can verify before releasing secrets? The Confidential Computing Consortium expressly includes attestation in its definition; it describes the TEE user as responsible for data and policy placed in that environment. (confidentialcomputing.io) (confidentialcomputing.io) The distinction matters for model weights, prompts, regulated records, and signing keys whose owner does not control every host administrator.

The Consortium defines the category through computation in a hardware-based, attested environment. (confidentialcomputing.io) Its terminology also says data-in-use protection is only one part of application security. (confidentialcomputing.io) Confidential Containers adds a concrete example: encrypting an image alone does not authenticate its contents or guarantee integrity. (confidentialcontainers.org)

This report uses **attestation** in the Remote ATtestation procedureS (RATS) sense. An **attester** creates evidence, a **verifier** appraises it against endorsements and reference values, and a **relying party** applies its own authorization policy to the resulting statement. (www.rfc-editor.org) Evidence can prove a measured hardware and software state within the documented boundary. It does not establish that a model is accurate, that an application is free of flaws, that an operator followed a change process, or that an input is lawful. Those are separate reviews. NIST's platform-firmware guidance is useful for procurement controls beyond runtime evidence; its newer confidential-computing blueprint was still an initial public draft in May 2026. (csrc.nist.gov) (csrc.nist.gov)

GPU Smith describes its role as specifying, integrating, and validating private AI infrastructure against written acceptance criteria. (gpusmith.com) Its published assurance scope includes serialized custody and as-built acceptance records. (gpusmith.com) For an adjacent engineering advisor, the practical perspective is to make the acceptance evidence a contractual deliverable; the GPU and verifier products in the comparison remain their respective vendors' offerings.

Three support layers must be kept separate. A GPU's technical ability to enter confidential mode is one layer; a vendor's validated **reference architecture** is another; a cloud or Kubernetes distribution's supported service is a third. NVIDIA's current Confidential Containers matrix, its R595 trusted-computing notes, Google's confidential VM documentation, and Red Hat's distribution matrix answer different questions. (docs.redhat.com) An auditor should retain the exact document version and access date beside each assertion because “latest” pages can change after commissioning.

H100 and H200: Hopper Paths

Capabilities

In NVIDIA's current container reference architecture, **H100** and **H200** appear as single GPU passthrough entries. Separate protected PCIe (PPCIe) entries cover multi GPU passthrough on Hopper HGX. The mode is topology-sensitive: NVIDIA says Hopper HGX multi GPU passthrough requires `ppcie`, and documents that mode as Hopper-only. (docs.nvidia.com) One should not take an H100 single GPU quote and silently extend it to an eight GPU HGX node. In Hopper PPCIe mode, GPU-to-GPU communications over NVLink or NVSwitch are not encrypted. (NVIDIA R595 release notes) The acceptance record needs the actual board or module, switch topology, and configured mode. NVIDIA states that key rotation is not supported in PPCIe mode.

The CPU side remains part of the claim. The NVIDIA container matrix pairs these GPUs with either an AMD SEV-SNP host or an Intel TDX host from its validated processor families. AMD describes SEV-SNP as adding VM memory-integrity protection; Intel documents measurement registers that can be placed into a TDX attestation quote. (www.amd.com) (www.intel.com) Those CPU reports do not, by themselves, establish the GPU's firmware state. NVIDIA's GPU evidence includes a device certificate chain and report, and its multi GPU collection path also enumerates switches and topology. (docs.nvidia.com)

Intel states that TDX quote generation takes place inside the trust domain. (cc-enabling.trustedservices.intel.com) AMD states that its secure processor signs an SEV-SNP attestation report with a Versioned Chip Endorsement Key. (docs.amd.com) The two mechanisms need separate evidence collection and verification procedures.

Adoption and deployment status

There are concrete cloud H100 paths, each with a different CPU dependency. Google Cloud's A3 High confidential instance uses an H100 and Intel TDX; its one H100 configuration became generally available in **July 2025**. (docs.cloud.google.com) (docs.cloud.google.com) Azure's NCCadsH100v5 uses fourth-generation AMD EPYC with SEV-SNP and an H100 NVL; its size specification lists **94 GB GPU memory** for that one GPU. (learn.microsoft.com) (learn.microsoft.com) These service descriptions do not establish confidential H200, B200, or B300 availability on either cloud. A buyer should require the provider's current machine-type and region record for the exact ordered service.

Microsoft describes a confidential VM presenting an attestation token to a relying party before protected information is released. (learn.microsoft.com) This is a service-level control path, not a substitute for checking the GPU-specific evidence.

Strengths and limitations

Hopper has unusually explicit sample flows: NVIDIA's H100 guide walks through evidence collection, remote appraisal, and a token. An H200 buyer can reuse the *method* of demanding a CPU quote, GPU report, endorsements, and a policy result, but should not copy an H100 certificate or firmware allowlist. Google warns that its Confidential Space `cc_mode` token field concerns the GPU driver only, while separate claims address

the device's VBIOS and identity. (docs.cloud.google.com) (docs.cloud.google.com) Azure likewise distinguishes automatic platform attestation during boot from independent guest attestation, which a workload owner may perform. (learn.microsoft.com)

The operational limitation is policy ownership. A valid device certificate says who made the device, while an acceptance policy says whether that particular measured configuration should receive a model key. RFC 9334 keeps those decisions separate. (www.rfc-editor.org) A deployment that gathers a token but never gates secret release has telemetry, not a completed authorization flow. The acceptance test should withhold the secret when the nonce, VBIOS allowlist, CPU measurement, or per-device result is wrong.

Hopper acceptance sequence

The following is a proposed **fail-closed acceptance sequence** for the chosen H100 or H200 design. Before approving a PPCIe topology, assess exposure to physical access and privileged administrators: NVIDIA says an attacker with physical or logical superuser access can passively capture ciphertext and attempt to break it or the key. It should be performed against the supplier's approved release and repeated after material changes.

- **Freeze the bill of materials:** Record GPU SKU, host CPU, TEE, board form factor, and exact passthrough topology.
- **Match the support scope:** Attach the versioned platform and cloud or distribution support statement.
- **Record configuration:** Capture BIOS, IOMMU, GPU mode, guest image, firmware, VBIOS, and driver state.
- **Issue a challenge:** Generate a fresh nonce under the relying party's control.
- **Collect CPU evidence:** Preserve the original quote, endorsements, and launch measurements.
- **Collect GPU evidence:** Preserve the report, identity certificates, and each expected device identifier.
- **Appraise independently:** Record verifier version, reference set, policy hash, and signed per-device result.
- **Gate the model key:** Release only when the CPU, GPU, and workload policies all pass.
- **Run negative tests:** Change the nonce and then reject a mismatched measurement or missing expected device.
- **Retain and refresh:** Store immutable evidence, then repeat after an update, replacement, or policy change.

B200 and HGX B300: Blackwell Paths

Capabilities

B200 and **HGX B300** occupy single and multi GPU passthrough rows in NVIDIA's current Confidential Containers reference architecture. The B300 row was added with the **August 2026** release **1.1.0**; B200 appeared in the initial **1.0.0** general-availability release. (docs.nvidia.com) (docs.nvidia.com) NVIDIA describes Blackwell multi GPU communication over encrypted NVLink, while its `ppcie` mode remains a Hopper

mechanism. (docs.nvidia.com) This is a material implementation distinction: a procurement request should state whether the design is an HGX node, a single passed-through GPU, or a multi GPU virtual machine, then cite the corresponding validated path.

The Blackwell path still depends on host trust. In the current container stack, all host GPUs must be configured for confidential computing and assigned to one confidential container virtual machine. (docs.nvidia.com) A partial assignment or mixed-mode node therefore falls outside this documented reference architecture. The validated host choices and operating-system requirements are the same matrix categories used above: AMD SEV-SNP or Intel TDX, Ubuntu **25.10 or 26.04**, and a **6.17+** kernel. The documented cluster path also specifies Kubernetes, containerd, Kata Containers, GPU Operator, guest driver, and Trustee versions; an ordinary host GPU-driver installation can prevent Virtual Function I/O (VFIO) binding into the guest.

Adoption and deployment status

Release status should be stated at the layer making the claim. NVIDIA's **1.1.0** Confidential Containers documentation includes HGX B300; its separately published self hosted virtual-machine reference implementation lists H100, H200, B200, and RTX PRO 6000 BSE, but not B300. Red Hat's OpenShift Sandboxed Containers **1.13** matrix identifies a bare-metal H100 or DGX B200 path as generally available on specified OpenShift versions. (docs.redhat.com) That distribution-specific statement cannot be used to label a B300 Kubernetes build generally available, nor to infer a cloud offering.

A support-status distinction remains inside Red Hat documentation: its bare-metal OpenShift platform matrix describes H100 and DGX B200 general availability, while its NVIDIA GPU attestation integration is labeled **Technology Preview**. (docs.redhat.com) The documented Trustee integration forwards GPU evidence to NVIDIA Remote Attestation Service before applying release policy. (docs.redhat.com) A buyer should ask whether the specific attestation workflow, rather than only the container platform, is supported for production.

Strengths and limitations

The R595 trusted-computing notes describe up to **eight Blackwell GPUs** in one confidential virtual machine and specify firmware **1.4.X** for HGX B200 and B300, with respective VBIOS patterns **97.00.E4.00.XX** and **97.10.64.00.XX**. (docs.nvidia.com) These numbers are a dated compatibility snapshot, not a permanent minimum for every future release. NVIDIA directs buyers to its secure-AI compatibility matrix for the live GPU, VBIOS, driver, and mode combinations. (www.nvidia.com) Firmware, driver, and verifier versions belong in the acceptance manifest and an approved update plan.

For multi device approval, a single overall “success” field is insufficient unless the verifier's policy also covers every expected GPU and interconnect component. NVIDIA's claims guide describes detached per-device claims; Confidential Containers' Trustee models each device as a submodule and binds additional device evidence into a primary report. (confidentialcontainers.org) A sensible negative test omits one expected GPU or changes the declared topology, then checks that policy blocks key release. That test is an acceptance requirement proposed here, not a vendor certification claim.

Blackwell acceptance sequence

For B200 and HGX B300, use the same trust-chain roles with a topology inventory that explicitly covers each device. The procedure below is a purchaser acceptance design rather than a claim of vendor certification.

- **Name the architecture:** Specify HGX system, GPU count, virtual-machine path, and software release.
- **Check compatibility:** Archive the exact driver, VBIOS, firmware, and confidential-mode combination.
- **Inventory interconnects:** Record switches and the intended GPU-to-GPU path before evidence collection.
- **Check assignment:** Compare the node's physical accelerator inventory with the confidential VM's assigned devices.
- **Collect composite evidence:** Obtain a fresh CPU report and every expected GPU's attestation report.
- **Verify each member:** Inspect device identity, measurement match, debug state, and signed result separately.
- **Check completeness:** Reject an overall success when an expected GPU or switch is absent from policy.
- **Authorize once:** Bind the accepted result to a named workload and its key-release request.
- **Force a denial:** Deliberately change topology or one device allowlist entry and retain the denial log.
- **Repeat on change:** Re-attest after firmware, driver, image, verifier, or resource-policy updates.

RTX PRO 6000 Blackwell and Adjacent Form Factors

Capabilities

The current NVIDIA Confidential Containers matrix lists **RTX PRO 6000 Blackwell Server Edition** for **single GPU passthrough**. Its absence from the matrix's multi GPU row should be read as “not validated in this reference architecture,” not as a universal statement about hardware capability. NVIDIA's SDK compatibility guidance asks for an **R580 or later** compatible driver for RTX PRO 6000 Blackwell. (docs.nvidia.com) The precise VBIOS and driver pairing still needs the live compatibility matrix and the platform release notes at commissioning.

Adoption and deployment status

Google Cloud's current G4 confidential virtual machine documentation pairs RTX PRO 6000 with **AMD SEV** and describes standard, Spot, flex-start, and reservation provisioning. (docs.cloud.google.com) That is a provider-specific offer, not an endorsement of every RTX PRO configuration. Google also documents the H100 A3 High path with Intel TDX, demonstrating that CPU TEE choice is service-specific even inside one provider. (docs.cloud.google.com) In Confidential Space, Google says only single GPU passthrough is supported. (docs.cloud.google.com)

Strengths and limitations

The RTX PRO option can be useful when a one GPU confidential VM meets the workload's memory and concurrency requirements. However, a product name plus a cloud machine family is not a complete attestation statement. The purchasing team needs to know whether the provider exposes the raw GPU device report, a

provider-issued token, or both; which verifier appraised the report; and which claims are merely driver measurements. Google's documentation separates GPU driver and GPU device attestation and recommends Secure Boot for stronger driver provenance. (docs.cloud.google.com) (docs.cloud.google.com)

A cloud-supplied token can simplify operations, but the customer still owns its relying-party policy. Google says the token assertions must pass the workload identity provider's policy before resource access. (docs.cloud.google.com) In a self hosted system, the equivalent control may live in Trustee's Key Broker Service, whose policy decides which resources are released to which workloads. (confidentialcontainers.org) The same acceptance question applies in both settings: can the workload obtain the sensitive key if the GPU report is missing, stale, mismatched, or omitted from the policy? The expected answer should be proven by a retained negative-test result.

Feature Comparison

Table 1 is a dated **support matrix**, not a claim that every cloud or distribution offers every row. “Single” and “multi” refer to GPU passthrough into the confidential virtual machine under NVIDIA's container reference architecture as read on **24 September 2026**. (docs.nvidia.com)

GPU or topology	Documented confidential path	CPU and software dependency	Acceptance evidence
H100, one GPU	Single GPU passthrough in Google’s A3 High confidential VM. (docs.cloud.google.com)	Intel TDX on A3 High; Azure’s H100 path uses AMD SEV-SNP. (learn.microsoft.com)	Collect CPU quote, GPU device evidence and a verifier result. (docs.cloud.google.com) (www.rfc-editor.org)
H200, one GPU	Single GPU passthrough in NVIDIA’s container matrix. (docs.nvidia.com)	Verify the proposed CPU TEE and release-specific firmware. (www.amd.com) (cc-enabling.trustedservices.intel.com)	Retain measurement, certificate and policy result. (www.rfc-editor.org)
H100/H200 HGX, multi GPU	Hopper protected PCIe mode in the reference path. (docs.nvidia.com)	Inventory every accelerator assigned to the confidential VM.	Check per-device and topology evidence against the requested GPU count. (confidentialcontainers.org)
B200, one or multiple GPUs	Both passthrough modes appear in the initial NVIDIA general-availability matrix. (docs.nvidia.com)	Validate the chosen CPU TEE and release-specific component set.	Retain per-device claims and switch inventory; Red Hat also documents a DGX B200 path. (confidentialcontainers.org) (docs.redhat.com)
HGX B300, one or multiple GPUs	Added to NVIDIA’s container architecture **1.1.0**. (docs.nvidia.com)	Confirm the exact platform release and component set.	Retain release-specific firmware, GPU and CPU evidence, then apply result policy. (www.rfc-editor.org)
RTX PRO 6000 BSE, one GPU	Single GPU path in Google’s G4 confidential VM. (docs.cloud.google.com)	Compatible driver **R580+** in NVIDIA guidance; G4 pairs the GPU with AMD SEV. (docs.nvidia.com)	Distinguish driver-only claim from full device evidence. (docs.cloud.google.com)

The matrix supports a purchasing decision only when the selected row is paired with a specific deployment guide. NVIDIA's Kubernetes reference validates **containerd** and excludes nested virtualization; those constraints do not automatically describe an Azure virtual machine or Google's G4 service. Conversely, Red Hat's published matrix adds distribution-specific conditions to its H100 and DGX B200 paths. (docs.redhat.com) A supplier should identify which support statement governs the proposed bill of materials and provide an escalation path for any deviation.

Attestation chain and evidence package

The chain begins with a **fresh challenge** from the party that will trust the result. The attester returns CPU and GPU evidence that includes measurements and a device identity chain. The verifier checks signatures, certificate status, expected measurements, and policy, then produces an attestation result. The relying party independently decides whether to release the key. (www.rfc-editor.org) NVIDIA describes a Reference Integrity Manifest (RIM) service for signed expected measurements and a remote service that returns a signed result; it also documents customer-managed verification. (docs.nvidia.com) Trustee splits these functions among an Attestation Service, Reference Value Provider, and Key Broker Service. (confidentialcontainers.org)

The Entity Attestation Token specification requires a freshness mechanism and specifies at least **64 bits of entropy** when using its nonce claim. (www.rfc-editor.org) (www.rfc-editor.org) RFC 9334 cautions that a nonce alone cannot establish when an underlying measurement value was generated. (datatracker.ietf.org) Therefore, policy should bind the challenge, evidence, verifier result, and release event in one auditable transaction.

The following **evidence checklist** is intended as a procurement deliverable. Each item should carry a collection time, producing tool and version, and a cryptographic hash in the audit repository.

- **Scope:** service name, region or site, purchase SKU, serial or unique device identifier, GPU count, and intended topology.
- **Host:** CPU family, TEE type, firmware and BIOS settings, hypervisor, host operating system, kernel, and IOMMU setting.
- **Guest:** confidential VM image digest, launch measurement, Secure Boot state, Kata or container runtime version where used.
- **Accelerator:** GPU mode, firmware, VBIOS, driver, switch inventory, and encrypted interconnect configuration.
- **Challenge:** nonce, requesting identity, generation time, and replay-prevention window.
- **Evidence:** raw CPU report, GPU report, per-device claims, and device certificate chains.
- **Endorsements:** manufacturer roots, intermediate certificates, revocation or Online Certificate Status Protocol (OCSP) result, and signed RIMs.
- **Verification:** verifier implementation and version, reference-value set, appraisal-policy hash, signed result, and per-device outcome.
- **Authorization:** relying-party policy, resource identifier, key-release decision, and trace linking result to workload.
- **Negative tests:** changed nonce, rejected measurement, missing GPU, policy mismatch, and unavailable verifier.

- **Lifecycle:** token expiry, re-attestation schedule, update trigger, owner, and retention location.

NVIDIA's current GPU claim set exposes fields for nonce matching, secure boot, debug state, driver and VBIOS, measurement match, and certificate status. (docs.nvidia.com) AMD documents a CPU certificate chain through its Versioned Chip Endorsement Key; the Linux guest API can return a report and certificate data. (docs.amd.com) (cdn.kernel.org) These inputs are different artifacts and should not be collapsed into an unqualified “attested” checkbox. The Linux guest interface documents report retrieval through `SNP_GET_REPORT`; preserving the returned report is stronger evidence than storing only a success string. (cdn.kernel.org)

Performance and Benchmarks

The available support documents describe compatibility and topology rather than a controlled, cross-generation benchmark of model tokens per second, encrypted interconnect bandwidth, or verification latency. The independent studies discussed below also use different test conditions. A buyer should request a benchmark protocol covering both steady-state inference and the operational cost of attestation and re-attestation.

The protocol should fix workload, precision, model size, batch schedule, prompt and output length, host CPU, driver, GPU firmware, confidential mode, interconnect, and key-service placement. It should report the distribution of quote generation time, reference-value retrieval, certificate-status lookup, verifier processing, key release, and first-token time separately. This separation matters because a remote verifier or external revocation endpoint can dominate the startup path even when the GPU's compute path is unchanged. NVIDIA's architecture explicitly distinguishes evidence collection, RIM retrieval, and local or remote appraisal.

For production sizing, repeat the measurement after a cold start, after a driver update, under concurrent pod launches, and when the verifier is unavailable. RFC 9334 expressly recognizes failure when evidence or result appraisal fails or the verifier cannot be reached. (www.rfc-editor.org) The acceptance criterion should specify whether that failure blocks secret release and whether a previously authorized workload may continue for a bounded period. That policy is a customer decision; it cannot be inferred from a GPU data sheet.

Two research preprints illustrate why method matters. An H100 and TDX study reports average time-to-first-token increases of **21.8%** and **27.8%** for two tested models. (arxiv.org) A separate B200 study reports approximately **1% to 3%** throughput overhead on its paired single-host runs. (arxiv.org) These are different workloads and metrics, so their percentages cannot rank H100 against B200 or predict H200 or B300 behavior.

The word **performance** also needs a boundary. A GPU memory encryption statement does not quantify inference slowdown; a token issuance time does not quantify application throughput; and a one GPU cloud shape does not establish multi GPU scaling. Google Cloud's current A3 High release note identifies one H100, while its supported-configurations page says confidential NVIDIA GPU VMs do not support clusters for multi-node workloads. (docs.cloud.google.com) (docs.cloud.google.com) These are deployment limits, not benchmark measurements. The report therefore records no fabricated overhead percentage and treats any future supplier figure as a claim until its test conditions can be reproduced.

Data Analysis and Evidence

The **24 September 2026** comparison yields a useful count without inventing a market statistic. NVIDIA's container matrix has **seven GPU/topology rows**: H100, H200, H100 PPCIe, H200 PPCIe, B200, HGX B300, and RTX PRO 6000 BSE. (docs.nvidia.com) Its host table names **two TEE families**, AMD SEV-SNP and Intel TDX, on **Ubuntu 25.10 or 26.04** with a **6.17+** kernel. Its component matrix specifies **Kubernetes 1.32+**, **containerd 2.3.x**, **Kata 4.0.0**, **GPU Operator 26.3.1+**, and a guest driver **595.58.03**. (docs.nvidia.com) These are versioned validation conditions for one reference architecture, not a universal minimum for all confidential GPU services.

A NIST initial public draft explains that a TEE report can contain initial firmware, bootloader, operating-system, and application measurements, and that a remote verifier checks authenticity before comparing the measurements with expected values. (nvlpubs.nist.gov) (nvlpubs.nist.gov) It is a draft blueprint, not a final certification of any GPU service.

The dates show how quickly status can change. NVIDIA called its container architecture **1.0.0** generally available on **29 April 2026** and added HGX B300 in **1.1.0** on **11 August 2026**. (docs.nvidia.com) Google Cloud's one H100 A3 High shape became generally available on **31 July 2025**. Red Hat's **1.13** matrix places H100 or DGX B200 on a bare-metal generally available path from OpenShift **4.21.24+**, while its Azure H100 integration is marked Technology Preview at **4.22.5+**. (docs.redhat.com) (docs.redhat.com) The status words are attached to different products; they cannot be aggregated into a single “B200 is GA everywhere” claim.

Table 2 translates marketing claims into a testable **claim, evidence, verifier, and refresh trigger** record. The timing below is a proposed audit policy unless a cited source states a fixed token lifetime.

Claim to approve	Evidence to retain	Verifier or decision owner	Expiry or refresh trigger
Correct CPU TEE launched	SEV-SNP report or TDX quote, endorsement chain, guest measurement. (docs.amd.com) (cc-enabling.trustedservices.intel.com)	CPU evidence verifier; relying party checks launch policy. (www.rfc-editor.org)	New boot, image or firmware change, or expired result.
Genuine GPU in approved mode	GPU report, certificate chain, nonce, mode, VBIOS and driver claims.	GPU verifier with signed RIM and revocation check.	New challenge, GPU or driver update, or certificate-status change.
Expected complete topology	Inventory and per-device result for every GPU and switch. (confidentialcontainers.org)	Verifier policy plus purchaser topology allowlist.	Device replacement, mode change, or GPU count change.
Approved workload receives a key	Result token, workload identity, policy hash, release log. (confidentialcontainers.org)	Relying party or Key Broker Service. (confidentialcontainers.org)	Policy or image change; Google Confidential Space token lasts **one hour**. (docs.cloud.google.com)
Failed appraisal blocks release	Deliberately mismatched evidence and denial log. (www.rfc-editor.org)	Independent acceptance tester and resource owner.	Each major release and verifier or policy change.

The table exposes a common quantitative mistake: an attestation token's lifetime is not the lifetime of firmware assurance. Google's Confidential Space token is documented as lasting **one hour**, but that service-specific value is not an expiration rule for every GPU deployment. (docs.cloud.google.com) Nor is a VBIOS version alone proof that the correct model image was launched. The former is a device or driver claim; the latter belongs to the guest and application policy. Trustee documents separate attestation, resource-release, and Kata-agent policies for exactly this kind of separation. (confidentialcontainers.org)

Sample evidence manifest

The following schema is an **illustrative purchaser-owned record**, not an NVIDIA or cloud-provider output format. Values must be populated from actual attestation and deployment evidence, with hashes pointing to immutable stored artifacts.

```
{
 "collected_at_utc": "<timestamp>",
 "platform": {"provider_or_site": "<name>", "sku": "<exact-sku>", "tee": "<SEV-SNP|TDX>"},
 "topology": {"gpu_model": "<exact-model>", "gpu_count": "<integer>", "mode": "<mode>"},
 "versions": {"host_firmware": "<version>", "gpu_vbios": "<version>", "guest_driver": "<version>",
"verifier": "<version>"},
 "challenge": {"nonce": "<random-value>", "requester": "<identity>"},
 "artifact_hashes": {"cpu_report": "<sha256>", "gpu_evidence": "<sha256>", "endorsements": "<sha256>",
"result": "<sha256>"},
 "policy": {"appraisal_hash": "<sha256>", "relying_party_hash": "<sha256>"},
 "decision": {"all_devices_accepted": "<boolean>", "secret_released": "<boolean>", "negative_test_log": "
<sha256>"}
}
```

NVIDIA's `nvattest` command can collect live evidence as JSON with one entry per device, and replaying saved evidence requires its corresponding nonce. (docs.nvidia.com) The manifest therefore preserves the nonce and raw artifact hash together. The purchasing team should retain the policy decision and denial log alongside the successful result.

Implications and Future Directions

The responsibility boundary changes with the deployment model. In a cloud service, the provider controls the physical host and publishes the eligible machine type; the customer controls its workload image, relying-party policy, and interpretation of returned claims. Google Cloud says a verifier may be its own service, Intel Trust Authority, or a customer-built verifier, and its resource access flow tests token assertions against customer policy. (docs.cloud.google.com) (docs.cloud.google.com) Azure describes a platform boot attestation plus an optional independent guest attestation path. (learn.microsoft.com) On premises, the owner must also preserve BIOS settings, firmware inventory, physical custody, verifier operation, reference values, and key-broker availability. AMD's and Intel's CPU evidence formats remain separate from NVIDIA's GPU device report. (docs.amd.com) (cc-enabling.trustedservices.intel.com)

Table 3 is a **RACI** assignment template. **A** means accountable for acceptance, **R** performs the work, **C** is consulted, and **I** receives the record. It is a proposed contract map, not an assertion that every provider offers the same service.

Control	Infrastructure provider or on-prem operator	Verifier operator	Workload owner	Independent acceptor
Platform inventory and supported topology	R	C	A	I
CPU and GPU evidence acquisition	R	R	A	I
Endorsement and reference-value maintenance	C	R	A	I
Attestation and resource-release policy	I	C	A/R	C
Negative tests and update re-attestation	R	R	A	R
Evidence retention and approval record	C	C	A/R	R

The table deliberately makes the **workload owner** accountable for approving the protection it needs. RFC 9334 gives the relying party a distinct policy decision; Confidential Containers similarly separates an Attestation Service from the Key Broker Service that releases resources. (www.rfc-editor.org) (confidentialcontainers.org) A provider may operate those services, but operation does not transfer the customer's obligation to define which CPU, GPU, image, and device states are acceptable. AWS Nitro Enclaves documentation illustrates a distinct, non-GPU pattern in which an external service checks signed enclave measurements against its access policy. (docs.aws.amazon.com) AWS's general shared-responsibility guidance also assigns guest operating-system and application management to the customer, although its Nitro Enclaves attestation documentation should not be mistaken for a confidential GPU support claim. (aws.amazon.com) (docs.aws.amazon.com)

For near-term procurement, a request for proposal should require a **versioned support statement**, a **sample evidence package**, and a **live fail-closed demonstration** before production authorization. After a driver or VBIOS update, compare the new measurement with the intended signed reference value before re-opening key release. NVIDIA's RIM service and GPU claims support this style of check; Red Hat notes that its Trustee reference hashes are supplied from outside Trustee rather than generated by it. (docs.redhat.com) That makes reference-value provenance an explicit owner and change-control item. Future platform releases may widen the matrix; each new row should be evaluated under the same evidence and policy criteria, without carrying forward an old token as proof of a new build.

Vendor questions before approval

These questions convert a general confidential-computing promise into contract language and acceptance artifacts.

- **Supported path:** Which dated matrix row, release, and exception list govern this exact SKU?

- **Verifier:** Who operates it, which version runs, and can the customer inspect its policy and signed result?
- **Trust roots:** Which manufacturer certificates, revocation checks, and reference values are used?
- **Completeness:** How does policy detect a missing accelerator, switch, or device claim?
- **Freshness:** Who creates the challenge, and how is replay rejected when services are unavailable?
- **Secret release:** Which independent party actually denies access after a failed appraisal?
- **Update control:** Which changes require a new baseline and acceptance run before production resumes?

Frequently Asked Questions (FAQs)

Does H100 confidential computing automatically provide attestation?

No. Confidential mode and successful cryptographic appraisal are separate actions. NVIDIA says enabling mode alone does not verify the confidential environment or release secrets. The approval package needs a CPU quote, GPU evidence, endorsements, reference values, and the relying party's decision. (www.rfc-editor.org)

Is H200 attestation identical to H100 attestation?

The RATS roles and evidence categories are comparable, but the exact GPU, firmware, driver, and topology must be on the applicable support matrix. NVIDIA lists each as a separate row and requires protected PCIe for Hopper HGX multi GPU passthrough.

Can B200 or B300 be attested in a multi GPU deployment?

Use the dated matrix above for the container path. The self hosted virtual-machine reference implementation has a narrower list, so the proposed software path must be named.

Who verifies the GPU evidence?

NVIDIA documents local, remote, and customer-managed approaches. Whoever operates the verifier, the relying party still sets the result policy and controls release of a protected key or resource. (www.rfc-editor.org) Trustee's Key Broker Service is one documented implementation pattern. (confidentialcontainers.org)

What should trigger re-attestation?

At minimum: a new workload launch, firmware or VBIOS change, driver or image update, device replacement, policy change, result expiration, or verifier change. This is a recommended acceptance policy derived from the fields and roles above. NVIDIA exposes nonce, measurement, version, and certificate-status checks, while RFC 9334 requires appraisal against current policy. (www.rfc-editor.org)

Conclusion

The defensible decision is an **evidence-backed deployment approval** for the exact configuration described in the dated matrix. Each quote, design, and acceptance test must identify the support scope it relies on.

An acceptable trust chain joins a CPU TEE report, GPU device evidence, signed endorsements and reference measurements, a verifier's per-device result, and a relying-party policy that actually gates the secret. (www.rfc-editor.org) (confidentialcontainers.org) The evidence manifest and negative-test log should travel with the as-built system, then be refreshed after changes. Cloud customers should demand the same chain even when a provider supplies the machine type and token service; on-premises owners must operate more of the chain themselves. Neither arrangement turns memory protection into proof of application correctness, model quality, or end-to-end security. The strongest procurement language states the exact topology, the accepted measurements, who verifies them, when they expire, and what happens when appraisal fails. It also assigns an owner to each artifact: inventory, quote, device report, reference value, signed result, policy, and denial log. That record gives security and procurement teams a repeatable decision rule when components are replaced or software is updated. A proposal that cannot produce the package should remain at evaluation stage until the missing evidence and release conditions are specified.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://docs.nvidia.com/datacenter/cloud-native/confidential-containers/latest/supported-platforms.html
2. https://docs.nvidia.com/datacenter/cloud-native/confidential-containers/latest/release-notes.html
3. https://docs.nvidia.com/enterprise-reference-architectures/deploying-proprietary-models-confidential-compute-self-hosted-vms/latest/reference-implementations.html
4. https://www.rfc-editor.org/rfc/rfc9334.html
5. https://confidentialcontainers.org/docs/attestation/architecture/
6. https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/create-a-confidential-vm-instance-with-gpu?hl=en
7. https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/release-notes
8. https://learn.microsoft.com/en-us/azure/confidential-computing/gpu-options
9. https://docs.cloud.google.com/confidential-computing/confidential-space/docs/reference/token-claims?hl=en
10. https://docs.amd.com/api/khub/documents/Fs8c5rfhxC4nlZfXaZal0Q/content
11. https://confidentialcomputing.io/2023/04/06/why-is-attestation-required-for-confidential-computing/
12. https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/Common-Terminology-for-Confidential-Computing.pdf
13. https://confidentialcontainers.org/docs/features/signed-images/
14. https://csrc.nist.gov/pubs/sp/800/193/final
15. https://csrc.nist.gov/pubs/ir/8320/e/ipd
16. https://gpusmith.com/
17. https://docs.redhat.com/en/documentation/openshift_sandboxed_containers/1.13/html/deploying_confidential_containers_on_bare-metal_servers/cc-discover_metal-cc
18. https://docs.nvidia.com/datacenter/cloud-native/confidential-containers/latest/configure-cc-mode.html
19. https://docs.nvidia.com/595trd1-trusted-computing-solutions-release-notes.pdf
20. https://www.amd.com/en/developer/sev.html

21. https://www.intel.com/content/www/us/en/developer/articles/community/runtime-integrity-measure-and-attest-trust-domain.html
22. https://docs.nvidia.com/attestation/quick-start-guide/latest/attestation-examples/collecting_evidence.html
23. https://cc-enabling.trustedservices.intel.com/intel-tdx-enabling-guide/07/trust_domain_at_runtime/
24. https://docs.amd.com/api/khub/documents/%7EuAtQszeypAVVEk_B91Ojg/content
25. https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/nccadsh100v5-series
26. https://learn.microsoft.com/en-us/azure/confidential-computing/guest-attestation-confidential-vms
27. https://learn.microsoft.com/en-us/azure/confidential-computing/confidential-vm-faq
28. https://docs.redhat.com/en/documentation/openshift_sandboxed_containers/1.13/html/deploying_red_hat_build_of_trustee_for_workloads_running_on_bare-metal_servers/trustee-discover_metal-trustee
29. https://www.nvidia.com/en-us/data-center/solutions/confidential-computing/secure-ai-compatibility-matrix/
30. https://docs.nvidia.com/attestation/attestation-client-tools-sdk/latest/sdk_compatibility.html
31. https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/attestation
32. https://docs.cloud.google.com/confidential-computing/confidential-space/docs/create-grant-access-confidential-resources
33. https://confidentialcontainers.org/docs/attestation/policies/
34. https://docs.nvidia.com/attestation/quick-start-guide/latest/architecture.html
35. https://www.rfc-editor.org/rfc/rfc9711.html
36. https://datatracker.ietf.org/doc/rfc9334/
37. https://docs.nvidia.com/attestation/advanced-documentation/latest/claims-guide/gpu_claims.html
38. https://cdn.kernel.org/doc/html/latest/virt/coco/sev-guest.html
39. https://arxiv.org/abs/2607.19353
40. https://arxiv.org/abs/2608.26575
41. https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/supported-configurations
42. https://nvlpubs.nist.gov/nistpubs/ir/2026/NIST.IR.8320E.ipd.pdf
43. https://docs.nvidia.com/attestation/nv-attestation-sdk-cpp/latest/sdk-cli/command-reference.html
44. https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/attestation-overview
45. https://docs.aws.amazon.com/enclaves/latest/user/set-up-attestation.html
46. https://aws.amazon.com/compliance/shared-responsibility-model/
47. https://docs.redhat.com/en/documentation/openshift_sandboxed_containers/1.13/html/deploying_confidential_containers_on_bare-metal_servers/configure-cc-overview_metal-cc

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com