White Paper

Cybersecurity
Intel® Trust Authority

Seamless Attestation with Intel® Trust Authority

Seamless Attestation of Intel® TDX and NVIDIA H100 Trust Execution Environments (TEEs) with Intel® Trust Authority

Authors

Raghu Yeluri
Intel Fellow, Lead Architect, Confidential Computing Services, Intel

Eunjung Yoon
Cloud Software Architect, Security Researcher, Intel

Michael O'Connor
Senior Director, Confidential Computing, NVIDIA

Karthik Jayaraman
Senior Software Engineer, NVIDIA

Table of Contents

Executive Summary

AI is now the most significant workload in data centers and the cloud. It is being embedded into other workloads, used for standalone deployments, and distributed across hybrid clouds and the edge. Many of the demanding AI workloads require hardware acceleration with a GPU. Today, AI is already transforming a variety of segments like finance, manufacturing, advertising, and healthcare. Many AI models are considered priceless intellectual property – companies spend millions of dollars building them, and the parameters and model weights are closely guarded secrets. Even knowing what some of the parameters are in a competitor's model is valuable intelligence. Furthermore, the data sets used to train these models are also considered highly confidential and can create a competitive advantage. As a result, data and model owners are looking for ways to protect these, not just at-rest and in-transit, but in-use as well.

Confidential Computing is an industry movement to protect sensitive data and code while in use by executing inside a hardware-hardened, attested Trusted Execution Environment (TEE) where code and data can be accessed only by authorized users and software. For AI workloads, this would include the model parameters, and weights, and the training or inferencing data. Learn more about confidential computing at the Confidential Computing Consortium.

Attestation and Trust

Attestation is an essential process in Confidential Computing where a stakeholder is provided a cryptographic confirmation of the state of a Confidential Computing environment. It asserts that the TEE instantiated is genuine, conforms to their security policies, and is configured exactly as expected. The frequency of attestation is determined by policy and can happen at launch time and periodically during runtime of the TEE. Attestation is critical to establish trust in the computing platform you are about to use with your highly sensitive data.

Intel and NVIDIA deliver Confidential Computing technologies that establish independent TEEs on the CPU and GPU, respectively. For a customer, this presents an attestation challenge, requiring attestation from two different services to gather the evidence needed to verify the trustworthiness of the CPU and GPU TEE's.

Through this collaboration, Intel and NVIDIA are providing a unified attestation solution for customers to verify the trustworthiness of the CPU and GPU TEEs for Confidential Computing based on Intel® Xeon® processors with Intel® Trust Domain Extensions (Intel TDX) and NVIDIA Tensorcore H100 GPUs. Intel TDX is an architecture extension in the Intel Xeon family of processors that enable hardware-based TEEs. Intel TDX is designed to isolate VMs from the virtual-machine manager (VMM)/hypervisor and any other software outside the Trust Domain, thus helping protect the TD from a broad range of software attacks.

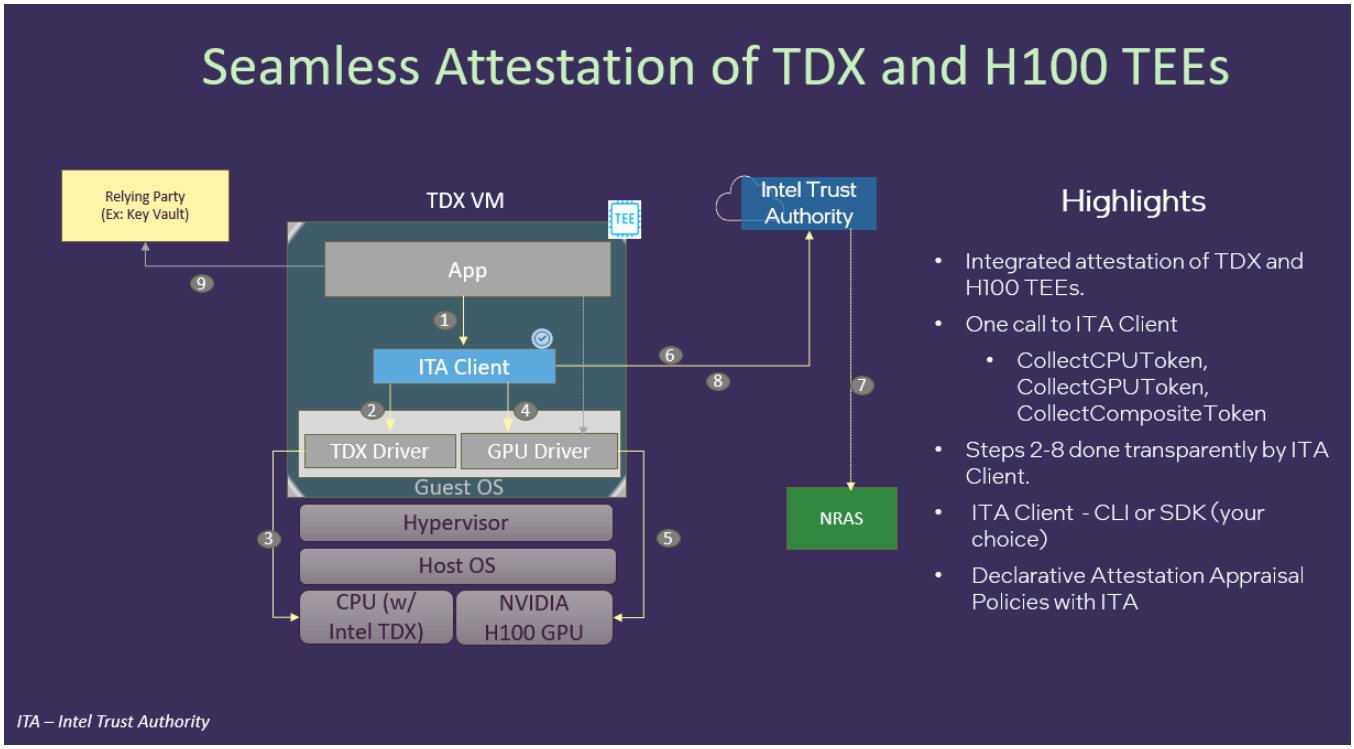

Figure 1. Architecture: Intel Trust Authority + NVIDIA NRAS

TEEs hosted on Intel processors can receive attestation services using several methods. The hosting Cloud Service Provider may offer an in-house attestation service, certain ISVs offer their own, or customers can build a private service. This paper will focus on CPU attestation via Intel's just released cloud-based independent trust service, Intel Trust Authority. TEEs hosted on NVIDIA GPU receive attestation via NVIDIA's Remote Attestation Service (NRAS).

At the Confidential Computing Summit, NVIDIA and Intel shared a unified attestation architecture (Figure 1).

Confidential Computing users will have two different options for attesting the CPUs and the GPUs as shown in Figure 1. Users either make separate calls to the Intel Trust Authority Client for full platform attestation and receive a single combined token or make separate calls to the Trust Authority Client for CPU attestation and GPU attestation, receiving tokens for each. In both options, Intel Trust Authority SaaS verifies the CPU attestation evidence and NVIDIA NRAS verifies the GPU attestation evidence.

Intel Trust Authority
An Independent Trust Service for CPU TEEs

Intel Trust Authority is an operator-independent, trustworthiness verification service delivered as a SaaS, and client component referred to as Trust Authority Client. Intel Trust Authority provides comprehensive remote attestation capabilities for Intel CPU-based TEEs (Intel SGX and Intel TDX), GPU-based TEEs, and non-Intel CPU-based TEEs (AMD-SEV/SNP in preview) with plans for additional confidential computing devices as they become available. Intel Trust Authority provides trustworthiness verification irrespective of where the TEEs are - public, private, or edge clouds. Intel Trust Authority is aligned to IETF-RATS architecture and supports both the passport model and background check model of attestation. Intel Trust Authority has a rich policy framework supporting very granular customer-defined appraisal policies for both CPU and GPU TEEs.

NVIDIA NRAS – Remote Attestation Service for GPU-Based TEEs

The NVIDIA Remote Attestation Service (NRAS) is the SaaS service provided by NVIDIA for verifying the attestation reports of NVIDIA GPUs. NRAS's capabilities include:

- Accept a GPU attestation report (evidence) as input from an attestor.
- Call the RIM service to fetch RIM Bundles (Golden Measurements) for comparison against the evidence.
- Call the NVIDIA OCSP service to check the revocation status of the device and RIM certificate chains.
- Compare the evidence with the measurements from the RIM bundle.
- Return attestation results as a signed EAT.

NRAS creates the signed Entity Attestation Token (EAT), based on JSON Web Token (JWT). The Relying Party or users can call NRAS using the NVIDIA Attestation SDK or by calling the NRAS APIs directly.

In the rest of this whitepaper, we walk through the next level of details of how this unified attestation works, with Intel TDX TEEs on the CPU and NVIDIA H100-based TEE on the GPU, and the workflows for each of the options with the Intel Trust Authority and NVIDIA NRAS. The key objective of this design is to encapsulate and simplify the work applications must do to incorporate attestation into their workflows. Applications simply make API calls into the Trust Authority Client – such as collectCPUToken(), collectGPUToken(), or collectCompositeToken() to trigger the attestation flows. All the complexity of fetching the TEE evidence as a signed report from the TEE hardware, sending that evidence to the Attestation services, and fetching the signed attestation tokens is done behind the scenes by the services behind the Trust Authority Client APIs. In the case of collectCompositeToken(), the Attestation token will be a composite signed EAT Token, with distinct individual CPU and GPU Attestation Tokens contained in it.

Intel Trust Authority Client

The Intel Trust Authority Client is an essential component of Intel Trust Authority services, which encapsulates and abstracts the workflow required to gather the attestation evidence from the TEEs and securely deliver to the Trust Authority SaaS, prior to launching workloads inside any TEE. Trust Authority Client has a very extensible model and is integral to building Confidential Computing solutions. The Trust Authority Client initiates attestation and can fetch the attestation both from the CPU and the GPU, and the signed token and certificates from the remote attestation services, including from Intel Trust Authority and NRAS.

For TDX and NVIDIA H100 GPU attestation, we provide the seamless integration of the Trust Authority Client, Intel Trust Authority, and NRAS for the GPU attestation.

Figure 2 shows the higher-level attestation flow.

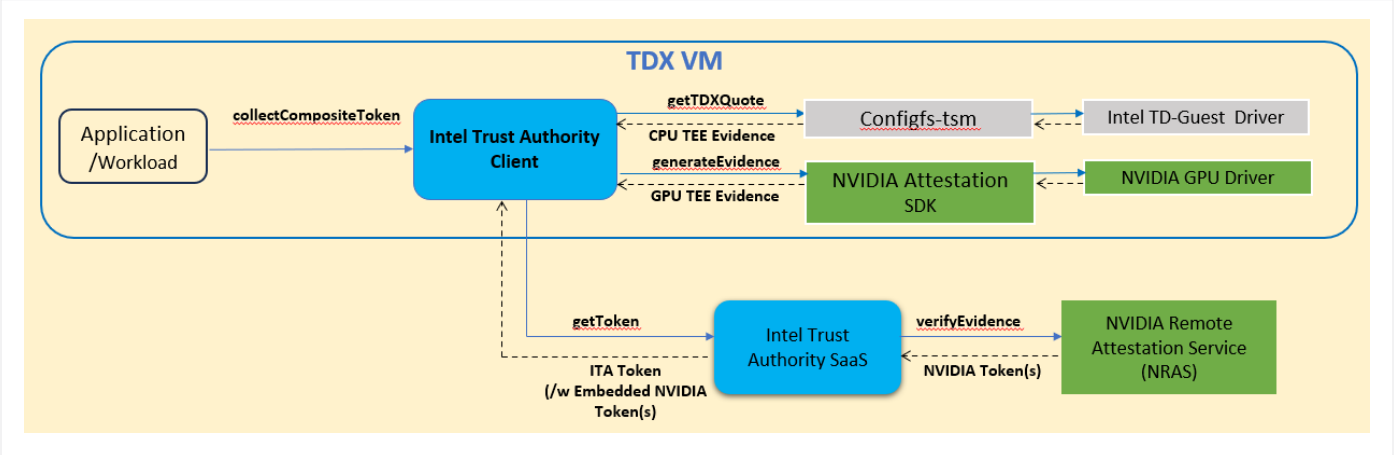

Figure 2. Higher-level flow for Trust Authority Client → Intel Trust Authority → NRAS

The Trust Authority Client collects the attestation evidence from Intel TDX and H100 GPU and calls Intel Trust Authority which in turn calls out to NRAS to get the NVIDIA signed token with the verified evidence.

- Application invokes the collectCompositeToken API in the Trust Authority Client.
- The Trust Authority Client gets a signed nonce from Intel Trust Authority.
- The Trust Authority Client requests the TDX Quote from the TDX guest driver in the CPU TEE (TDX Confidential VM)
- The Trust Authority Client requests the GPU Evidence from the NVIDIA GPU driver in the CPU TEE (TDX Confidential VM) using the NVIDIA Attestation SDK with the nonce.
- The Trust Authority Client receives the attestation evidence from the TDX guest driver and the GPU driver.

(* Intel Trust Authority generated nonce is passed to the request to the GPU driver for the SPDM measurement request to the GPU.)

- The Trust Authority Client calls Intel Trust Authority to request the signed EAT token passing the TDX and GPU attestation evidence, and any composite attestation policy for the TDX and GPUs. The attestation policy also can be pre-defined in the SaaS by the application owner.
- Intel Trust Authority calls out to NRAS using the NVIDIA SDK, sending the evidence for verification, and collects the NVIDIA signed token from NRAS.
- Intel Trust Authority verifies the NVIDIA token(s) and generates a composite Intel Trust Authority signed Token with embedded NVIDIA Token(s) to the Trust Authority Client.
- The Relying Party can get the signed token and the token signing certificate from Intel Trust Authority and verify the token with the certificate.

Example Use Case: Confidential Training

Before any models are available for inferencing, they must be created and trained over a significant amount of data. For most scenarios, model training requires large amounts of computation, memory, and storage. A cloud infrastructure is well suited for this, but requires strong security guarantees – at rest, in transit, and in use. Figure 3 shows a reference architecture for confidential training.

- The TEEs include both the Intel TDX CPU TEEs and H100 GPU TEEs. The Key broker and distribution services must process attestation reports for both the CPUs and GPUs, before they release the keys to the TEEs.
- The Key broker services will release the model and data decryption keys directly into the Trust Domain (TD, which is the Intel TDX TEE), once the attestation of the Intel TDX CPU and the attestation of the H100 GPU are both verified with Intel Trust Authority.
- If the Key Broker services do not release the decryption keys, the application will exit.

This architecture will verify and provide proof to the model builder and the data owner that the model (parameters, weights, checkpoint data, etc.) and the training data are not visible outside the TEEs.

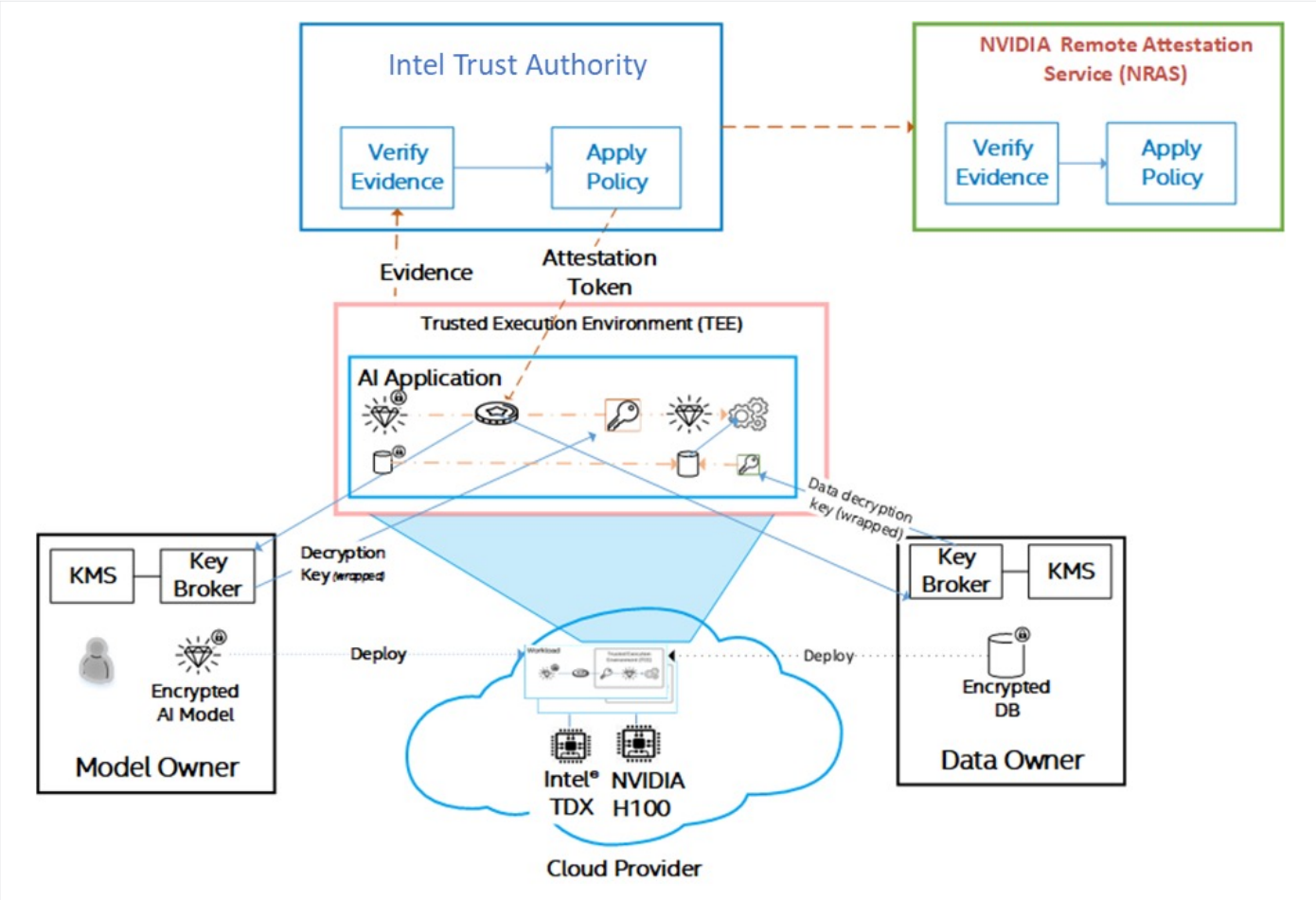

Figure 3. Reference Architecture for Confidential Training

Availability

Intel and NVIDIA have actively collaborated to bring innovative solutions to market. As of September 2024, the Intel Trust Authority has incorporated attestation support for the NVIDIA H100 GPU in its General Availability (GA) release. This support is compatible with CUDA 12.5 and the r550 version of the NVIDIA driver.

Learn more

Learn more about Intel Trust Authority on Intel.com.

1224/MM/TM/PDF ♻PleaseRecycle 363887-001US