インテル® TDX と NVIDIA* H100 の TEE をインテル® Trust Authority でシームレスに認証

この記事は、インテルのサイトで公開されている「Seamless Attestation of Intel® TDX and NVIDIA H100 TEEs with Intel® Trust Authority」の日本語参考訳です。原文は更新される可能性があります。原文と翻訳文の内容が異なる場合は原文を優先してください。

AI は現在、データセンターとクラウドにおける最も重要なワークロードです。ほかのワークロードに埋め込まれたり、スタンドアロン展開に使用されたり、ハイブリッド・クラウドとエッジに分散されています。要求の厳しい AI ワークロードの多くは、GPU によるハードウェア・アクセラレーションを必要とします。AI はすでに金融、製造、広告、ヘルスケアなどのさまざまな分野を変革しつつあります。多くの AI モデルは貴重な知的財産と見なされ、企業はそれらの構築に多額の投資を費やし、パラメーターやモデルの重みなどの情報は厳重に保護されています。競合他社のモデルのパラメーターの一部を知るだけでも、貴重な情報となります。さらに、これらのモデルのトレーニングに使用されるデータセットもまた機密性が高く、競争上の優位性を生み出す可能性があります。そのため、データやモデルの所有者は、保存時や転送時だけでなく利用時にも保護する方法を模索しています。

コンフィデンシャル・コンピューティング (英語) は、ハードウェアで強化され、認証されたトラステッド・エグゼキュ―ション環境 (TEE) 内で実行することで、許可されたユーザーとソフトウェアのみがコードとデータにアクセスできるようにし、使用中の機密データとコードを保護する業界の取り組みです。AI ワークロードでは、モデルのパラメーター、重み、トレーニング、推論などのデータを保護できます。コンフィデンシャル・コンピューティングに関する詳細は、Confidential Computing Consortium (英語) を参照してください。

認証と信頼

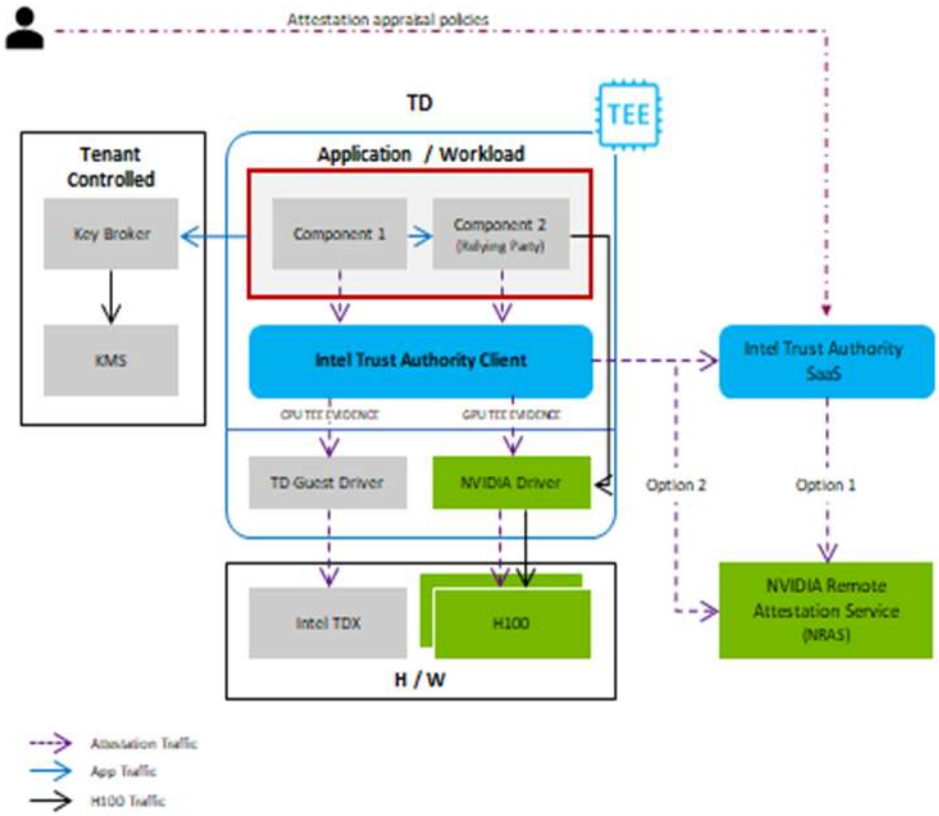
認証は、コンフィデンシャル・コンピューティングにおいて重要なプロセスであり、TEE 環境の状態について暗号による確認を提供します。インスタンス化された TEE が本物であり、セキュリティー・ポリシーに準拠し、期待どおりに正確に構成されているか検証します。認証の頻度はポリシーによって決定され、TEE の起動時およびランタイムに定期的に実行されます。認証は、機密性の高いデータを託す計算プラットフォームの信頼性を確保する上で重要です。

インテルと NVIDIA は、それぞれ CPU と GPU 上の TEE を認証する独自のコンフィデンシャル・コンピューティング・テクノロジーを提供しています。ユーザーが CPU と GPU それぞれの TEE の信頼性を検証するには、2 つの異なるサービスからの認証が必要になります。

ユーザーの負担を軽減するため、インテルと NVIDIA は協力して、インテル® トラスト・ドメイン・エクステンションズ (インテル® TDX) と NVIDIA Tensor Core H100 GPU を搭載したインテル® Xeon® プロセッサー・ベースのコンフィデンシャル・コンピューティング向けに、CPU と GPU の TEE の信頼性を検証する統合認証ソリューションを提供するため取り組んでいます。インテル® TDX は、ハードウェア・ベースの TEE を可能にするインテル® Xeon® プロセッサー・ファミリーのアーキテクチャー拡張機能で、仮想マシン・マネージャー (VMM)/ハイパーバイザーとトラストドメイン (TD) 外のソフトウェアから VM を分離するように設計されており、幅広いソフトウェアの攻撃から TD を保護します。

インテル® プロセッサーでホストされている TEE は、いくつかの方法で認証サービスを受け付けています。ホスティング・クラウド・サービス・プロバイダーが組織内の認証サービスを提供したり、特定の ISV が独自のサービスを提供したり、ユーザーがプライベート・サービスを構築することができます。この記事では、昨年インテルがリリースしたクラウドベースの独自のトラストサービスである、インテル® Trust Authority を介した CPU 構成証明に注目します。NVIDIA* GPU でホストされている TEE は、NVIDIA の Remote Attestation Service (NRAS) を介して認証を受け付けています。

コンフィデンシャル・コンピューティング・サミットにおいて、NVIDIA とインテルは、以下の図に示す統一認証アーキテクチャーを発表しました。



- インテル® Trust Authority クライアントは、CPU と GPU の TEE から構成証明を収集し、インテル® Trust Authority SaaS を呼び出してそれらを検証します。
- クライアントには、NVIDIA* ドライバーを介して GPU 構成証明を収集し、SaaS を介して (オプション 1)、または直接 (オプション 2) NRAS を呼び出すワークフローが含まれています。
- ローコード: 1 回の呼び出しですべての構成証明を収集して検証します。
- 最小限のアプリケーション変更でコンフィデンシャル・コンピューティング認証を追加できます。

上記の図に示すように、ユーザーは、CPU と GPU を認証する 2 つの異なるオプションを利用できます。1 つは、インテル® Trust Authority クライアントを個別に呼び出して完全なプラットフォーム認証を行い、単一の結合されたトークンを受け取ります。2 つ目は、インテル® Trust Authority クライアントを個別に呼び出して CPU 構成証明と GPU 構成証明を行い、それぞれのトークンを受け取ります。どちらのオプションでも、インテル® Trust Authority SaaS は CPU 構成証明を検証し、NVIDIA* NRAS は GPU 構成証明を検証します。

インテル® Trust Authority: CPU TEE 向けの独自のトラストサービス

インテル® Trust Authority は、SaaS として提供される、オペレーターに依存しない信頼性検証サービスであり、クライアント・コンポーネントはインテル® Trust Authority クライアントと呼ばれます。インテル® Trust Authority は、インテル® CPU の TEE (インテル® SGX およびインテル® TDX) に包括的なリモート認証機能を提供します (インテル以外の CPU TEE、GPU TEE、およびその他の秘密計算デバイスは順次サポート予定)。インテル® Trust Authority は、TEE がパブリッククラウド、プライベート・クラウド、またはエッジクラウドのどこにあるかに関係なく、信頼性の検証を提供します。インテル® Trust Authority は IETF-RATS アーキテクチャーに準拠しており、認証のパスポートモデルとバックグラウンド・チェック・モデルの両方をサポートしています。インテル® Trust Authority の豊富なポリシー・フレームワークは、CPU と GPU の TEE に対する詳細な顧客定義の評価ポリシーをサポートします。

NVIDIA* NRAS: NVIDIA* GPU ベースの TEE 向けリモート認証サービス

NVIDIA Remote Attestation Service (NRAS) は、NVIDIA* GPU の認証レポートを検証する、NVIDIA が提供する SaaS サービスです。 次のような機能を提供します。

- 入力として GPU 認証レポート (構成証明) を受け付けます。
- 構成証明と比較するため、RIM サービスを呼び出して RIM バンドル (ゴールデン測定) を取得します。
- NVIDIA* OCSP サービスを呼び出して、構成証明署名と RIM 署名を検証します。
- 構成証明を RIM バンドルの証明書チェーンと比較します。
- 認証結果を署名付きエンティティー認証トークン (EAT) として返します。

NRAS は、JSON Web トークン (JWT) を基に署名付き EAT を作成します。認証利用者やユーザーは、NVIDIA* Attestation SDK を使用するか、NRAS API を直接呼び出して NRAS を呼び出すことができます。

以降のセクションでは、CPU 上のインテル® TDX TEE と GPU 上の NVIDIA* H100 ベースの TEE でこの統合認証がどのように動作するか、そしてインテル® Trust Authority SaaS と NVIDIA* NRAS での各オプションのワークフローについて詳しく説明します。この設計の主な目的は、アプリケーション・ワークフローに認証を組み込むために必要な作業をカプセル化して簡素化することです。アプリケーションがインテル® Trust Authority クライアントに対して collectCPUToken()、collectGPUToken()、collectCompositeToken() などの API 呼び出しを行うだけで、認証フローがトリガーされます。TEE ハードウェアから TEE 構成証明を署名付きレポートとして取得し、それを認証サービスに送信し、署名付き認証トークンを取得するという複雑な作業はすべて、インテル® Trust Authority クライアント API の背後にあるサービスによってバックグラウンドで行われます。collectCompositeToken() を使用すると、インテル® Trust Authority 認証トークンは、個別の CPU および GPU 認証トークンを含む複合署名付き EAT トークンになります。

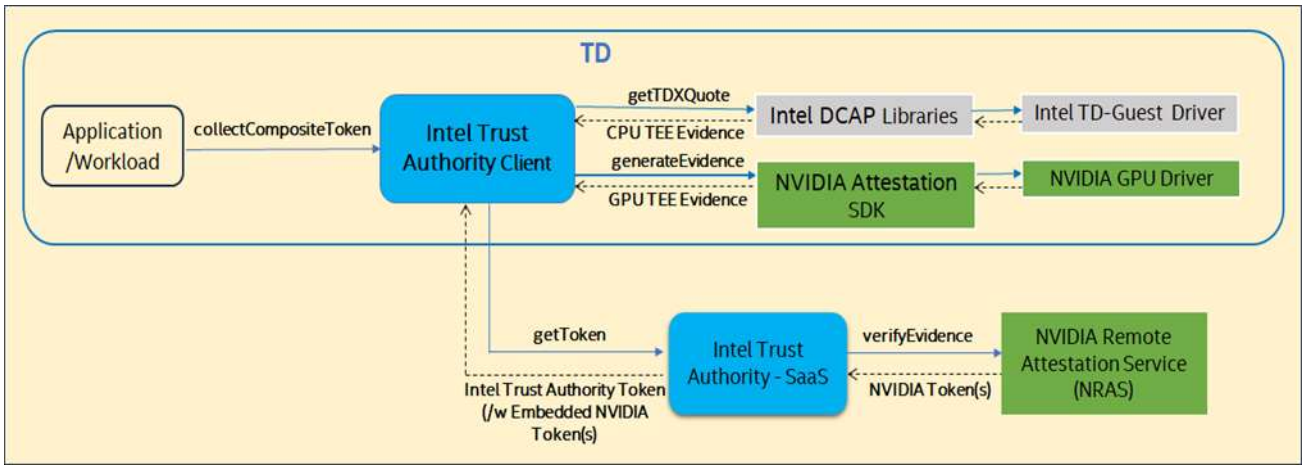
インテル® Trust Authority クライアント

インテル® Trust Authority クライアントは、インテル® Trust Authority サービスの重要なコンポーネントであり、TEE 内でワークロードを起動する前に、TEE から構成証明を収集し、インテル® Trust Authority SaaS に安全に送信するために必要なワークフローをカプセル化および抽象化します。インテル® Trust Authority クライアントのモデルは非常に拡張性が高く、コンフィデンシャル・コンピューティング・ソリューションの構築に不可欠です。インテル® Trust Authority クライアントは認証を開始し、CPU と GPU の両方から構成証明を取得し、インテル® Trust Authority SaaS や NVIDIA* NRAS などのリモート認証サービスから署名付きトークンと証明書を取得できます。

NVIDIA* H100 GPU 認証では、インテル® Trust Authority クライアントと SaaS を NVIDIA* NRAS と統合して GPU 認証を行う 2 つのオプションを提供します。

オプション 1: (インテル® Trust Authority クライアント → SaaS → NRAS)

次の図は、オプション 1 の上位レベルのフローを示しています。

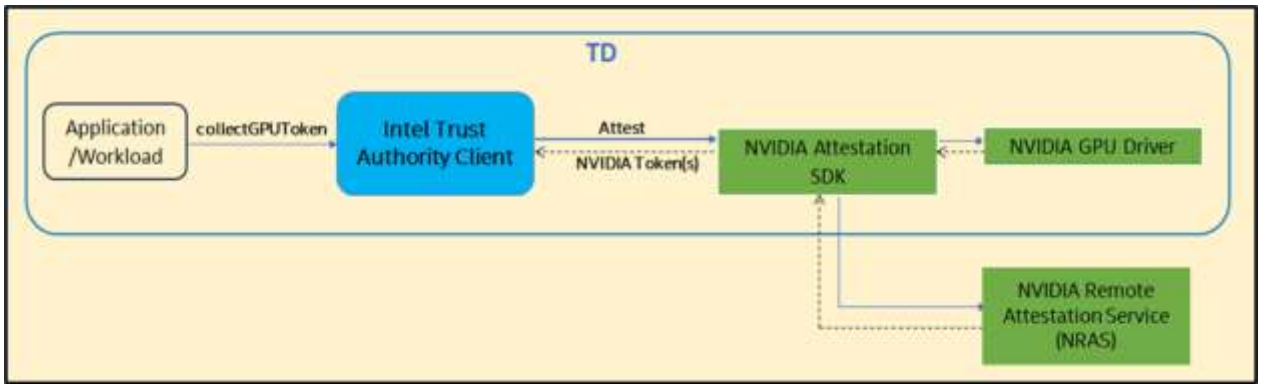


このオプションでは、インテル® Trust Authority クライアントがインテル® TDX および NVIDIA* H100 GPU から構成証明を収集し、インテル® Trust Authority SaaS を呼び出し、インテル® Trust Authority SaaS が NRAS を呼び出して、検証済みの構成証明と NVIDIA の署名付きトークンを取得します。

1. アプリケーションは、インテル® Trust Authority クライアントで collectCompositeToken() API を呼び出します。
2. インテル® Trust Authority クライアントは、インテル® Trust Authority SaaS から署名付きノンス (nonce) を取得します。
3. インテル® Trust Authority クライアントは、ノンスと NVIDIA* Attestation SDK を使用して、CPU TEE (つまり、インテル® TDX CVM) の NVIDIA* GPU ドライバーから認証レポートを要求します。
4. インテル® Trust Authority クライアントは、GPU ドライバーから構成証明を受け取ります。(* インテル® Trust Authority によって生成されたノンスは、GPU への SPDM 測定リクエストのため GPU ドライバーに渡されます。)
5. クライアントは SaaS を呼び出して、GPU 構成証明を渡す署名付き EAT トークンと GPU 認証ポリシーを要求します。GPU 認証ポリシーは、アプリケーションの所有者によって SaaS 内で事前定義されている場合もあります。
6. インテル® Trust Authority SaaS は、NVIDIA* SDK を使用して NRAS を呼び出し、検証用の構成証明を送信し、NRAS から NVIDIA の署名付きトークンを収集します。
7. インテル® Trust Authority SaaS は、NVIDIA のトークンを検証し、NVIDIA のトークンが埋め込まれたインテル® Trust Authority の署名付き複合トークンをインテル® Trust Authority クライアントに生成します。
8. 認証利用者は、インテル® Trust Authority SaaS から署名付きトークンとトークン署名証明書を取得し、証明書を使用してトークンを検証できます。

オプション 2: (インテル® Trust Authority クライアント → NRAS)

次の図は、オプション 2 の上位レベルのフローを示しています。



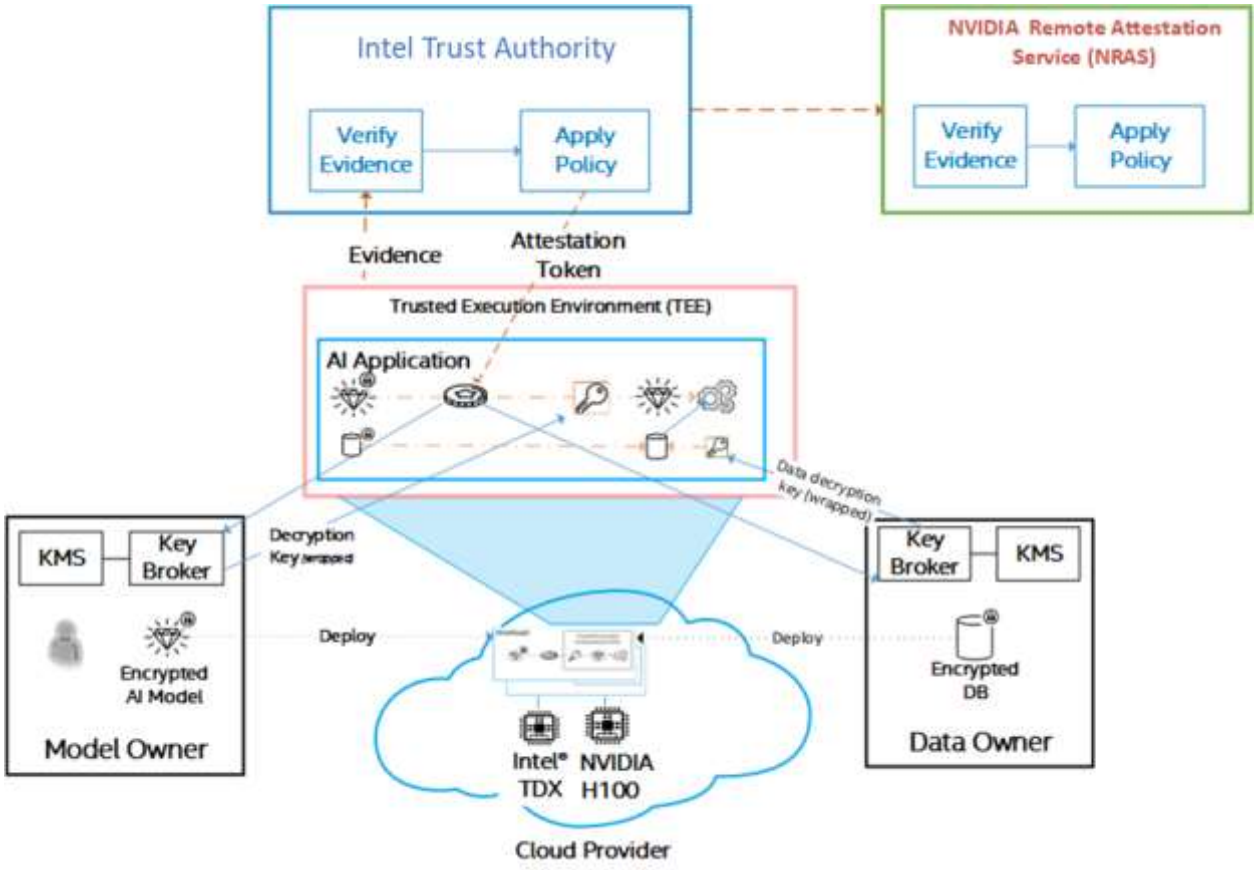
このオプションでは、インテル® Trust Authority クライアントは、インテル® Trust Authority SaaS を呼び出さずに、NRAS を直接呼び出して GPU 認証を行います。NVIDIA* H100 GPU TEE のリモート認証は、インテル® Trust Authority SaaS を介さずに、NRAS によって直接実行されます。

1. アプリケーションは、インテル® Trust Authority クライアントで collectGPUToken() API を呼び出します。
2. インテル® Trust Authority クライアントは、SaaS から署名付きノンスを取得します。
3. インテル® Trust Authority クライアントは、ノンスと NVIDIA* Attestation SDK を使用して、CPU TEE (つまり、インテル® TDX CVM) の NVIDIA* GPU ドライバーに認証レポートを要求します。
4. インテル® Trust Authority クライアントは、GPU ドライバーから構成証明を受け取ります。
5. インテル® Trust Authority クライアントは、NVIDIA* SDK を使用して NRAS を呼び出し、NVIDIA* SDK は検証用の構成証明を送信し、NRAS から NVIDIA の署名付きトークンを収集します。
6. インテル® Trust Authority クライアントは、NRAS からトークン署名証明書を取得し、その証明書を使用してトークンを検証します。

オプション 1 のほうが簡単で、1 回の API 呼び出しで環境の安全性を判断できるため、オプション 1 を**推奨**します。オプション 2 は、複雑な作業を厭わず、各ステップを自分で管理したいユーザー向けに提供されています。

使用例: 機密性の高いトレーニング

モデルを推論に使用できるようにするには、モデルを作成し、大量のデータでトレーニングする必要があります。ほとんどのシナリオでは、モデルのトレーニングには大量の計算、メモリー、ストレージが必要です。クラウド・インフラストラクチャーはこれに適していますが、保存時、転送時、使用時の強力なセキュリティーが保証されていなければなりません。次の図は、機密性の高いトレーニングのリファレンス・アーキテクチャーを示しています。



- インテル® TDX CPU TEE と NVIDIA* H100 GPU TEE の両方が含まれます。鍵ブローカーおよび配布サービスは、鍵を TEE にリリースする前に、CPU と GPU の両方の認証レポートを処理する必要があります。
- インテル® TDX CPU の認証と NVIDIA* H100 GPU の認証が両方ともインテル® Trust Authority で検証されると、鍵ブローカーサービスはモデルとデータの復号鍵を TD (インテル® TDX TEE) に直接リリースします。
- 鍵ブローカーサービスが復号鍵をリリースしない場合、アプリケーションは終了します。

このアーキテクチャーは、モデル (パラメーター、重み、チェックポイント・データなど) とトレーニング・データが TEE の外部には表示されないことをモデル構築者とデータ所有者に保証し、証明します。

提供時期

インテルと NVIDIA は、これらのソリューションを市場に投入するため積極的に取り組んでいます。インテル® Trust Authority は、2024 年前半に NVIDIA* H100 GPU TEE の認証サポートを組み込む予定です。NVIDIA* H100 GPU TEE のサポートは、CUDA* 12.2 Update 1 の早期アクセス機能、および 2023 年 7 月にリリースされた NVIDIA* ドライバーの R535.86.10 で利用できます。NRAS もまた、現在早期アクセスで利用できます。