Trusted Computing Future: Emerging Use Cases and Solutions

Authors:

Ramki Krishnan (ramki.krishnan@intel.com),
Ned Smith (ned.smith@intel.com),
Diego Lopez (diego.r.lopez@telefonica.com),
Ghada Arfaoui (ghada.arfaoui@orange.com)

Introduction

This paper describes data residency requirements from the perspective of an AI inferencing use case. It defines a trusted confidential computing architecture that brings together host-affinity, geo-location-affinity and secure storage.

Protection of data has long been an interesting and challenging pursuit of corporate IT, users, and governments. Data should be protected throughout all its management lifecycle (i.e., creation, storage, process, usage, and destruction) whenever it is "at rest", "in transit", or "in use" [1]. More recently protection of data (during all its lifecycle) includes protection within a physical geo-location.

Data sovereignty laws define a legal framework for managing data protection requirements. These laws are especially important in industries that handle sensitive customer data, such as those governed by the Payment Card Industry Data Security Standard (PCI DSS) [12], or in environments where privacy regulations like the European GDPR [13] apply. For example, EU e-health data should not leave the EU.

Data residency describes how to protect data within a defined geo-location. Data residency has two main aspects: first, host affinity which refers to the need for data to be tied to specific computing environments or hosts where it is stored and/or processed; and second, geo-location affinity which refers to the need for data processing that remains in a defined geographic region.

In a typical data residency use case such as Confidential AI inferencing, AI models are stored encrypted by workload secrets that are stored in a cloud-based key vault. The secrets unlock the AI model in a Confidential Computing (CC) environment. The AI workload is distributed across several CC environments in various edge locations. Confidential AI depends on an industry standard architecture for interoperability and trust. The paper proposes the Trusted Confidential Computing (TCC) architecture which is in an industry standard architecture for addressing the above challenges.

[1] https://www.securitystudio.com/blog/data-at-rest-vs-data-in-transit

Use Cases

Data workflow extends the data lifecycle to include processes that augment, transform, or otherwise add value to data. AI inferencing workflow is one such use case. Figure 1 illustrates an example deployment model for confidential AI inferencing.

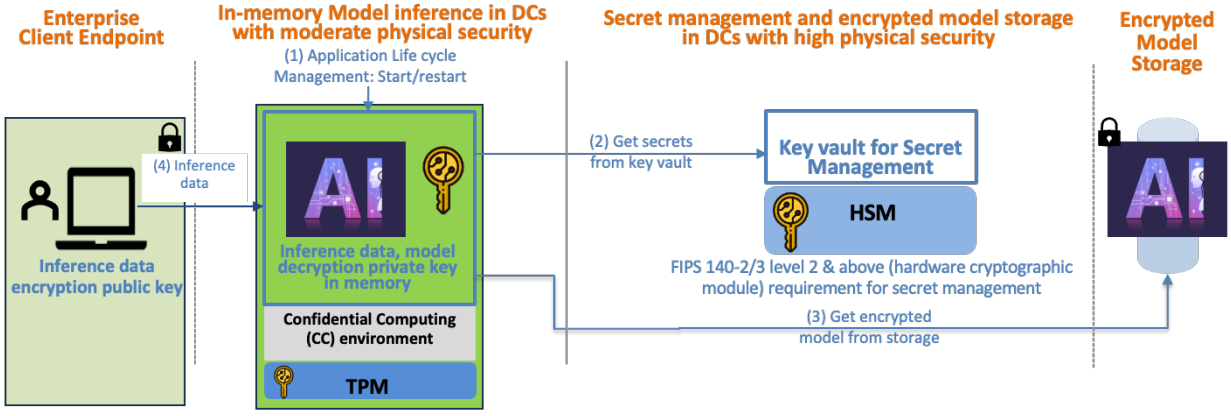


Figure 1: Example deployment model for confidential AI inferencing

In this deployment model:

- The AI model decryption key is securely stored in a Hardware Security Module (HSM)-based key vault located in data centers (DCs) with high physical security.
- The encrypted AI model is also stored in DCs with high physical security.
- The AI model inferencing application runs within a Confidential Computing (CC) environment where the model is processed in memory. This application is distributed across edge-located DCs with moderate physical security.

Upon starting or restarting (Step 1), the AI inferencing application follows these steps:

- Fetches the model decryption key from the HSM-based key vault (Step 2).
- Retrieves the encrypted AI model from storage and decrypts it using the decryption key (Step 3).
- The application is then ready to serve enterprise end users. End users encrypt their data with the inference data encryption public key (Step 4).

Data Residency Requirements

Location-affinity

The location-affinity requirement mandates that sensitive data to be processed only within specific geographic regions, such as a particular city or country. This aligns with EU GDPR legislation which governs the processing of any kind of personal data. The German BDSG [14], a national transposition of EU GDPR, imposes similar requirements.

Other countries; such as China, India, and Russia; have local data-localization mandates requiring that sensitive data — such as financial or personal information — must not leave the country.

Host-affinity

The host-affinity requirement ensures that AI or other sensitive data is processed exclusively on dedicated hosts that meet the highest security standards. For example, a host can be uniquely identified using a Trusted Platform Module (TPM) public key certificate or a Confidential Compute (CC) host ID. This approach aligns with the Payment Card Industry Data Security Standard (PCI DSS), which mandates that all payment processing – such as AI inferencing for transaction analytics – must be conducted on a set of dedicated, secure hosts.

Generalized data residency

In general, data residency requirements can be distilled as follows:

a) Data should be **stored** within a specific geographic location and / or dedicated hardware.

b) Data should be **processed** within a specific geographic location and / or dedicated hardware.

c) Data should **transit** within a specific geographic location and / or though dedicated hardware.

d) The **software** handling Data should be within a specific geographic location and / or dedicated hardware.

Data Residency Solution for Addressing Various Attacks

Sophisticated attacks can be mounted to deceive the requirements expressed in the above use cases. For example:

- **Physical Attacks**, such as:
    - movement of a host or a critical host component by an internal attacker (e.g., a malicious data center administrator).
- **Data in Transit Attacks,** such as:
    - compromise DC API gateway to redirect traffic to an unsafe geo-location;
    - compromise endpoint DNS cache to redirect traffic to a less secure server host/DC.
- **SW Infrastructure Attacks**, such as:
    - workload orchestrator compromise redirects workloads to less secure hosting node.
- **Old Hardware Issues**, such as:
    - old hardware, or hardware without latest firmware patch, returns inaccurate geo-location.
- **Old Software Issues**, such as:

- unpatched software returns inaccurate geo-location.

Combining location-affinity with host-affinity implies the enforcement of data being processed only within a specific geographic region and on specific secure hosting environments. This combined use case, integrating both approaches, can address the sophisticated attacks mentioned above providing a general reference for addressing requirements related to data residency.

Figures 2 and 3 highlight typical data lifecycle attacks and data residency solutions for servers and clients, respectively.

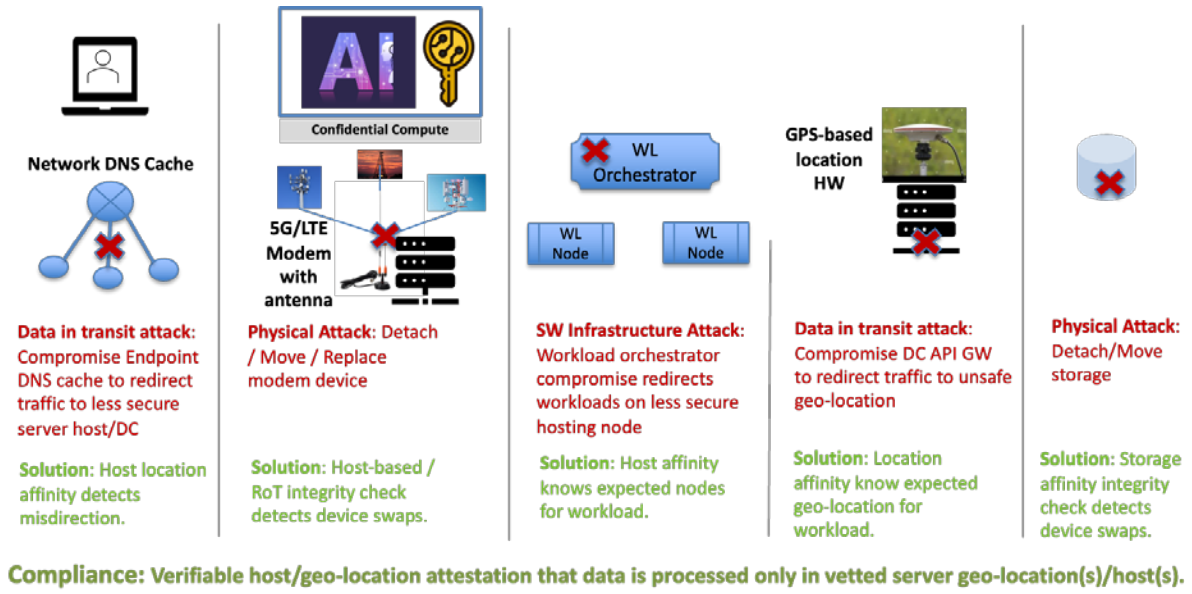


Figure 2: Data residency – security and compliance value for servers

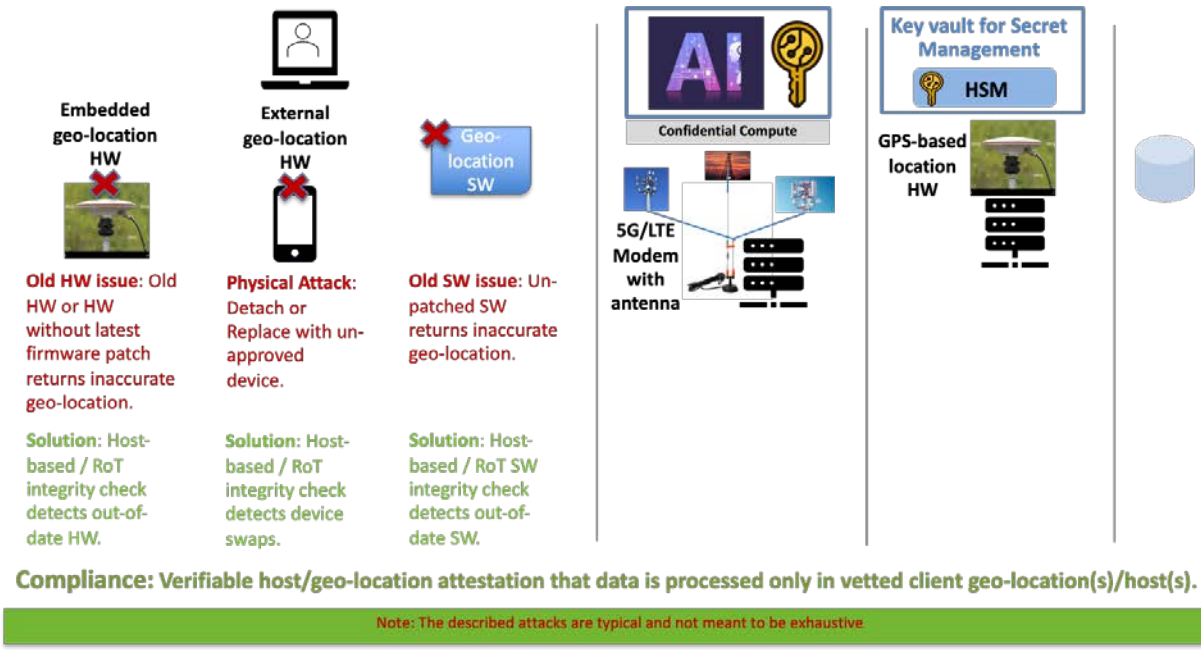


Figure 3: Data residency - security and compliance value for clients

In the context of verifiability for the location-affinity requirement, this means demonstrating that the inference DCs are in specific geographic regions. In the context of verifiability for the host-affinity requirement, this means proving that the HSM adapter is attached only to specific secure hosts. In the context of verifiability for the location-host affinity verifiability requirement, this means proving that both the key vault storage and inference processing occur only on specific, secure hosts in specific geographic regions. In the context of these requirements, sophisticated attacks (e.g., physical attacks such as movement of host or key host component, or compromise of the Data Center (DC) API gateway to redirect traffic to an unsafe geo-location) are depicted.

Key management lifecycle may mirror data lifecycle. If the model needs to be highly available, keys protecting the model may also need high availability. For example, HSM keys may be replicated across cooperating trusted peer HSMs; or a key may be partitioned into shares, distributed across peer HSMs, and reconstructed in the event of key corruption.

Various Challenges in Realizing Data Residency

This section overviews the various challenges in end-to-end trusted computing, and data residency host-affinity.

End-to-end Trusted Computing Challenges

- An endpoint is able to cryptographically prove that a host component (physical TPM, location sensor, etc.) is used only by an application running on the same host.
- An endpoint is able to cryptographically prove that the communication between host component (location sensor, etc.) and application cannot be tampered.
- Endpoints need to use geo-location sensors which can provide a cryptographically provable accurate geo-location[2].
- An endpoint has assurance from the manufacturer that the geo-location sensor HW cannot be spoofed.

Host Affinity Challenges

A variety of secure storage technologies may be used to securely store, access, and use secrets within a CC protected key vault – for example, Hardware Security Module (HSM), Trusted Computing Group (TCG) Trusted Platform Module (TPM2.0) [2] and DICE Protection Environment (DPE) [3]. A host-native secure secret store for keys is essential to establish host-affinity on Confidential Computing (CC) nodes. Secure storage that is available to CC environments would help mitigate some physical attacks, such as moving a HSM adapter between hosts.

[2]Endpoints typically use source IP address for geo-location which is inaccurate since Enterprises can move their IANA assigned IP addresses anywhere in the world.

However, a physical TPM as host-native key storage provider on the motherboard has challenges:

- Sharing the host TPM with CC workloads brings the host (Firmware/Hypervisor/OS) into the CC trust boundary, potentially expanding the attack surface.
- A single physical TPM lacks multi-tenant scalability for effective key management.

Also, TPM lacks a flexible interface for connected partitioned HSM usages:

- RFC7512 exposes HSM as web-aware PKCS11 [19] interface.
- CXL [9], SPDM [10] and TLS [11] transports have emerging attestation evidence conveyance capabilities that extend to a wide variety of trusted endpoints (memory, sensors, io, CC nodes).

Trusted Confidential Computing (TCC) Architecture

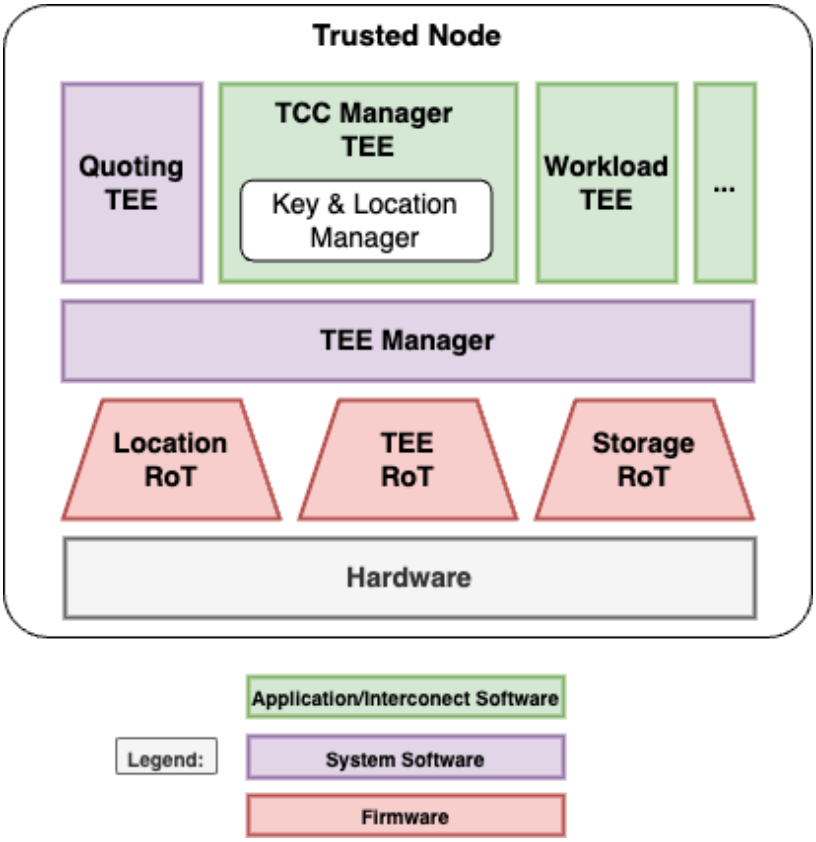


Figure 4: Trusted Confidential Computing architecture

Trusted Confidential Computing architecture depicted in *Figure 4* combines several trusted computing building block technologies to form a trusted computing node that is a platform upon which data residency use cases can be realized in cloud and edge environments. The trusted node consists of trusted hardware containing root of trust for the triad of trusted computing technology consisting of: trusted workload execution, trusted node location,

and trusted workload storage. Trusted nodes attest node components to ensure node integrity meets use-case requirements. Trustworthiness metrics may then become a condition of interoperation between peer trusted nodes.

Confidential Computing Environments

A variety of vendors are offering confidential computing technology such as [4], [5], [6], [7]. Confidential computing technology isolates Trusted Execution Environment (TEE) execution from peer TEEs as well as privileged system software including TEE Managers and Quoting TEE (Figure *4).* TEE isolation is typically implemented in hardware. Vendors of TEE technology work to ensure it is trustworthy.

The TEE Root of Trust (RoT) *(*Figure *4)* ensures the TCC manager, TEE manager, and other system components are measured as part of system initialization. System measurements are used by peer TEEs and other workload nodes to assess the operational integrity of a TCC node. The integrity assessment (a.k.a. attestation) is prerequisite to the TCC manager performing trusted operations.

Trusted Geolocation

The Location RoT (Figure *4*) ensures network endpoint integrity such as infrastructure processing units (IPU) and network interface cards (NIC) are measured as part of IO subsystem initialization or upon "hot-plug" attachment to the TCC node. The TCC Manager performs IO subsystem integrity checks as a prerequisite to binding IO devices to workload TEEs.

The IO subsystem may expose a variety of location sensors (e.g., GPS [15], 5/6G [16], WiFi [17], Bluetooth [18]) to the TCC Manager. Location may depend on locally connected processors and sensors as well as remotely connected nodes that beacon or triangulate location information. Before location information is utilized, the TCC Manager collects attestation evidence from the location sensing infrastructure to assess trustworthiness of location data.

There is potential for interposer attacks on location information. As an example, a location beacon broadcasts at higher power than the authentic beacon source. In response, location providers require authentication to cell towers and other sources of location information. Verification of location credentials is an important part of the TCC Manager duties that assures location information is trustworthy.

Secure Storage

The Trusted Computing Group (TCG) has helped shape trusted computing by publishing specifications for Trusted Platform Module (TPM) [2] and DICE Protection Environment (DPE) [3] that ensures a safe place to store and use cryptographic keys and other secrets.

The Storage RoT (*Figure 4)* ensures the secure storage subsystem components are measured as part of its initialization process so that the TCC Manager can assess security risks before allocating secure storage resources and using its capabilities. The TCC Manager performs storage subsystem integrity checks as a prerequisite to binding storage resources to workload TEEs.

Trusted Node Interconnect

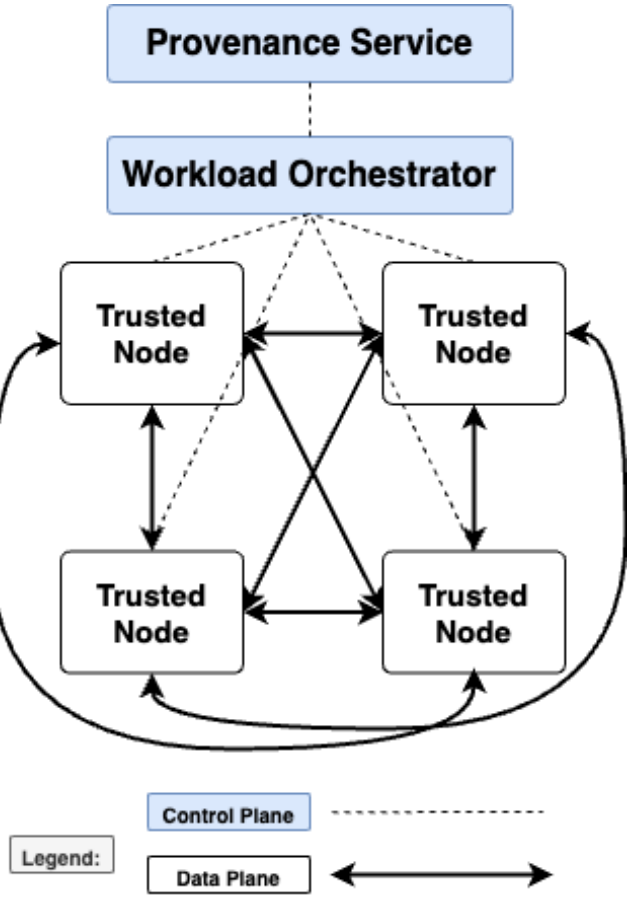


Figure 5: Trusted Node Interconnect

Remote attestation is used to dynamically construct a distributed workflow consisting of multiple nodes that cooperate to achieve a workload objective. The mesh could be autonomous or orchestrated depending on deployment context. The TCC Manager plays a critical role in assessing security risks of the IO and storage subsystems, but itself depends on peer nodes who assess the security risk of the TCC Manager. In a distributed workload mesh (*Figure 5*) consisting of multiple nodes, each node assesses the integrity of its peer nodes before becoming a member of the mesh. Similarly, each node requesting membership in the mesh must present its trustworthiness evidence. The TCC Manager in each node evaluates the security risk for each peer node and likewise presents attestation information to peer nodes to be assessed by them. A workload orchestrator ensures that workload nodes can be distributed and coordinated following a trusted lifecycle. A provenance service manages workloads, identities, TEEs, and platforms across their entire lifecycle.

Two important properties of a trusted mesh / network are:

1. Any node in the trusted network / mesh should not answer any type of request from a non-authorized node;
2. For ensuring privacy, nodes should not share unnecessary attestation information except for applying selective disclosure policies, a Trusted node must prove that it is in a *valid* state but doesn't share many of the details of its integrity assessment.

Conclusion

Data Residency End-to-end Value

Figure 6, below, depicts a production deployment for *Figure 1*, with focus on the communication between Enterprise client application (GenAI Chatbot) and Enterprise SaaS application (GenAI Inferencing). It adds geo-location to multi-factor-authentication (MFA) between workloads. It also describes the interaction with other man-in-the-middle components egress Secure Access Service Edge (SASE) (Data Loss Prevention, Web Application Firewall, etc.) and Ingress API Gateway (DDoS prevention, Load Balancing, etc.)

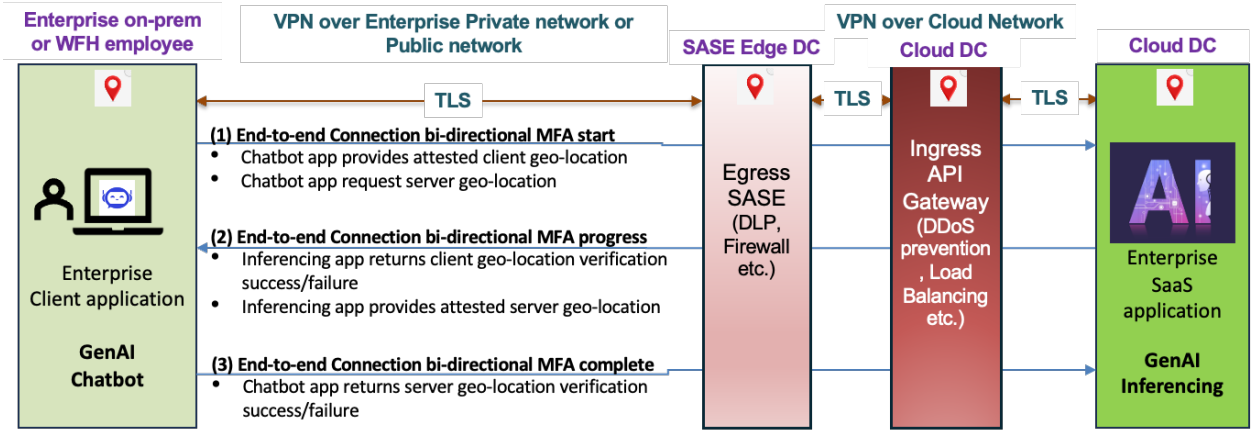


Figure 6: Data Residency End-to-end value

End-to-end solution summary

Deliver precise, third-party verifiable client/server application geo-location as part of bi-directional MFA between:

- Enterprise client application and Enterprise SaaS application connection (as depicted in **Figure 6**);
- Enterprise client application and Egress SASE application (TLS connection);
- Egress SASE application and Ingress API gateway application (TLS connection);
- Ingress API gateway application and Enterprise SaaS application (TLS connection).

End-to-end value

Addresses the data residency requirements and mitigates sophisticated threats while driving compliance.

The use cases for trusted computing in cloud and edge environments depend on data residency context amid increasing pressure to distribute processing across multiple network nodes. Location and host affinity is needed to ensure distributed workloads are processed on trustworthy nodes that reside within acceptable geographies. Nevertheless, workload lifecycles anticipate flexible distribution and deployment on mesh architectures.

Trusted computing technology needs to expand from traditional TCG roots of trust for storage, reporting (e.g., TPM, DPE). and measurement (e.g., DICE [8]) to trusted execution environments that securely connect to general purpose, high capacity, partitionable secret storage; and secure location sensing. This triad of trusted computing technology requires trust management in the form of attestation appraisal and risk assessments overseen by a TCC Manager, Workload Orchestrator, and Provenance Service.

Trusted computing standards are required to ensure a complex ecosystem of technology providers can supply various ingredients that not only interoperate but also function as a cohesive trusted foundation for distributed workloads.

Acknowledgements

Would like to thank Thomas Hardjono and James Borden for their detailed review and feedback.

References

[1] “What’s in a name? Data residency vs Data Sovereignty vs Local Legal Mandates”, Mark Boost, December 2, 2024. Available online: https://www.civo.com/blog/data-residency-vs-data-sovereignty-vs-local-legal-mandates
[2] “Trusted Platform Module Library Part 1: Architecture", Family "2.0”, Level 00, Revision 01.83, January 25, 2024. Available online: https://trustedcomputinggroup.org/wp-content/uploads/TPM-2.0-1.83-Part-1-Architecture.pdf
[3] “DICE Protection Environment”, Version 1.0, October 25, 2024. Available online: https://trustedcomputinggroup.org/wp-content/uploads/DICE-Protection-Environment-Version-1.0_pub.pdf
[4] “Design and Verification of the Arm Confidential Compute Architecture”, Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation, July 11-13, 2022. Available online: https://www.usenix.org/system/files/osdi22-li.pdf
[5] “AMD SEV-SNP: Strengthening VM Isolation with Integrity Protection and More”, AMD White Paper, January, 2020. Available online: https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/solution-briefs/amd-secure-encrypted-virtualization-solution-brief.pdf
[6] “Intel Trust Domain Extensions”, Intel White Paper. Available online: https://www.intel.com/content/dam/develop/external/us/en/documents/tdx-whitepaper-final9-17.pdf
[7] "Confidential compute on NVIDIA Hopper H100", Nvidia White Paper, Version 1.0, July 25, 2023. Available online: https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/HCC-Whitepaper-v1.0.pdf
[8] "Hardware Requirements for a Device Identifier Composition Engine", Version 1.0, Revision 0.91, August 8, 2024. Available online: https://trustedcomputinggroup.org/wp-content/uploads/Hardware-Requirements-for-a-Device-Identifier-Composition-Engine-Version-1.0-Revision-0.91_pub.pdf
[9] "Compute Express Link (CXL) Specification", Revision 3.2, Version 1.0, October 2, 2024. Available online: https://computeexpresslink.org/cxl-specification/

[10] "Security Protocol and Data Model (SPDM) Specification", DSP0274, Version 1.3.2, August 21, 2024. Available online: https://www.dmtf.org/sites/default/files/standards/documents/DSP0274_1.3.2.pdf
[11] "The Transport Layer Security (TLS) Protocol Version 1.3", RFC8446, August, 2018. Available online: https://www.rfc-editor.org/rfc/rfc8446
[12] "Payment Card Industry Data Security Standard (PCI DSS)", Requirements and Testing Procedures, Version 4.0.1, June, 2024. Available online: https://docs-prv.pcisecuritystandards.org/PCI%20DSS/Standard/PCI-DSS-v4_0_1.pdf
[13] General Data Protection Regulation (GDPR), Official Journal of the European Union, L 119/1, May 4, 2016. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
[14] Federal Data Protection Act of 30 June 2017 (Federal Law Gazette I p. 2097), as last amended by Article 10 of the Act of 23 June 2021 (Federal Law Gazette I, p. 1858; 2022 I p. 1045). Available online: https://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html
[15] "Global Positioning System Standard Positioning Service Performance Standard", 5th Edition, April, 2020. Available online: https://www.gps.gov/technical/ps/2020-SPS-performance-standard.pdf
[16] "Framework and overall objectives of the future development of IMT for 2030 and beyond", International Telecommunication Union Radiocommunication Sector, ITU-R, M.2160-0, November, 2023. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2160-0-202311-I!!PDF-E.pdf
[17] "IEEE Standard for Information Technology--Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," in IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016) , vol., no., pp.1-4379, 26 Feb. 2021. Available online: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9363693
[18] Bluetooth Core Specification, Version 6.0, August 27, 2024. Available online: https://www.bluetooth.com/specifications/specs/core60-html/
[19] PKCS #11 Cryptographic Token Interface Base Specification Version 2.40, OASIS Standard, 14 April, 2015. Available online: http://docs.oasis-open.org/pkcs11/pkcs11-base/v2.40/os/pkcs11-base-v2.40-os.doc