Google

November 2025

Google Private AI Compute:

Extending On-Device Privacy with the Power of the Cloud

From Google Platforms and Devices,
Google DeepMind, and Google Cloud.

We believe privacy is the foundation of personalized AI assistance.

The most effective AI doesn't just answer a user's query; it anticipates needs and understands the rich personal context behind them, such as their current task, their ambient environment, or their upcoming commitments. This requires processing of complex and often sensitive data, such as audio context, end-to-end encrypted messages, or device screen content.

Historically, the architecture for securely processing and caring for this type of data was on-device processing. The advancements to more proactive and personal AI experiences require a level of deep reasoning and advanced capabilities that demands the computational power of larger models, like the latest Gemini models, hosted in the cloud.

Private AI Compute in the cloud is our latest privacy infrastructure, built to deliver the speed and power from the cloud while extending the **user security and privacy assurances of on-device processing.** When a user interacts with Google AI experiences running on Private AI Compute, the system is engineered to ensure that their personal information, unique insights, and how they use them are **private only to the user.**

Private AI Compute is built on years of Google's existing investments in security and privacy and is engineered to deliver the following requirements.

User data processed by Private AI Compute is not available to anyone other than the user, including Google

Information is processed within a system that is designed to protect sensitive user data, which is referred to in this document as the protected execution environment, based on a hardware root of trust. No user data leaves the protected execution environment without explicit user intent or action.

Hyper-scalable private inference

To bring everyone the best of our AI, Google's cloud infrastructure is highly optimized to produce exceptional serving efficiency, capacity, and computational power. This design enables Google's most advanced AI models to be served at rates of millions of user queries per second. Private AI Compute uses this end-to-end globally distributed and optimized model serving infrastructure to deliver private inference on the latest Gemini models, powered by Google's own custom Tensor Processing Units (TPUs).

External verifiability

For this initial release, an external auditor has validated that our system design meets strict privacy and security guidelines. We view this as a foundational step, with a roadmap designed to introduce further transparency and attestation mechanisms that will enable deeper independent verifiability over time.

> **The following sections detail how we deliver on these requirements.**

Table of contents

01

Introducing our design approach

Private AI Compute builds on our deep investment in delivering AI responsibly and years of Google's innovations in privacy infrastructure and advanced security.

Private AI Compute builds on our investments in responsible AI, privacy, and security with foundations such as Private Compute Core to protect your most sensitive mobile device data; Federated Learning and Differential Privacy, which enable us to improve our products without revealing sensitive information; and end-to-end encrypted backup. On these foundations, combined with our existing Google Cloud security architecture, we designed Private AI Compute with the following additional multi-layered protections to ensure that user data is not available to anyone other than the user.

Create a protected execution environment across both CPU and TPU workloads

For CPU and TPU workloads (referred in this document as trusted nodes) we use an AMD-based hardware Trusted Execution Environment (TEE) to encrypt and isolate memory and processing from the host. For TPU workloads such as serving LLM workloads, we use a hardened TPU platform that delivers privacy and security properties comparable to a typical TEE, by using our TPU hardware (Google's Titanium Intelligence Enclave).

Starting with the sixth-generation of Google Cloud TPU (Trillium), Google Cloud's Titanium Hardware Security Architecture expanded to TPU hardware to meet Private AI Compute's needs. Only attested workloads run on the trusted nodes, administrative access into the workloads is not possible, and nodes are hardened against physical exfiltration of data.

Establish encrypted communication channels between trusted nodes

As the Private AI Compute serving infrastructure involves a multi-node distributed system, the system performs peer-to-peer attestation and encryption between the trusted nodes to ensure user data is decrypted and processed solely within a protected environment.

Attest trusted nodes

Attesting trusted nodes ensures their integrity as part of establishing encrypted communication channels such as Noise and Application Layer Transport Security (ALTS), to ensure that user data is shielded from broader Google infrastructure. The initiation of these encrypted connections uses bi-directional attestation: each workload requests and cryptographically validates the workload credentials of the other, ensuring mutual trust within the protected execution environment. Workload credentials are provisioned only upon successful validation of the node's attestation against internal reference values. Failure of validation prevents connection establishment, thus safeguarding user data from untrusted components.

In the Private AI Compute architecture, the client establishes a Noise encryption connection with a frontend server and establishes mutual attestation of auditable binaries over that connection. Subsequently, the frontend server establishes an ALTS encryption channel with other services in the scalable inference pipeline and with model servers running on the hardened TPU platform. These services and servers are provisioned only upon successful attestation validation. The overall trust in the Private AI Compute system is established by transitive trust between these servers.

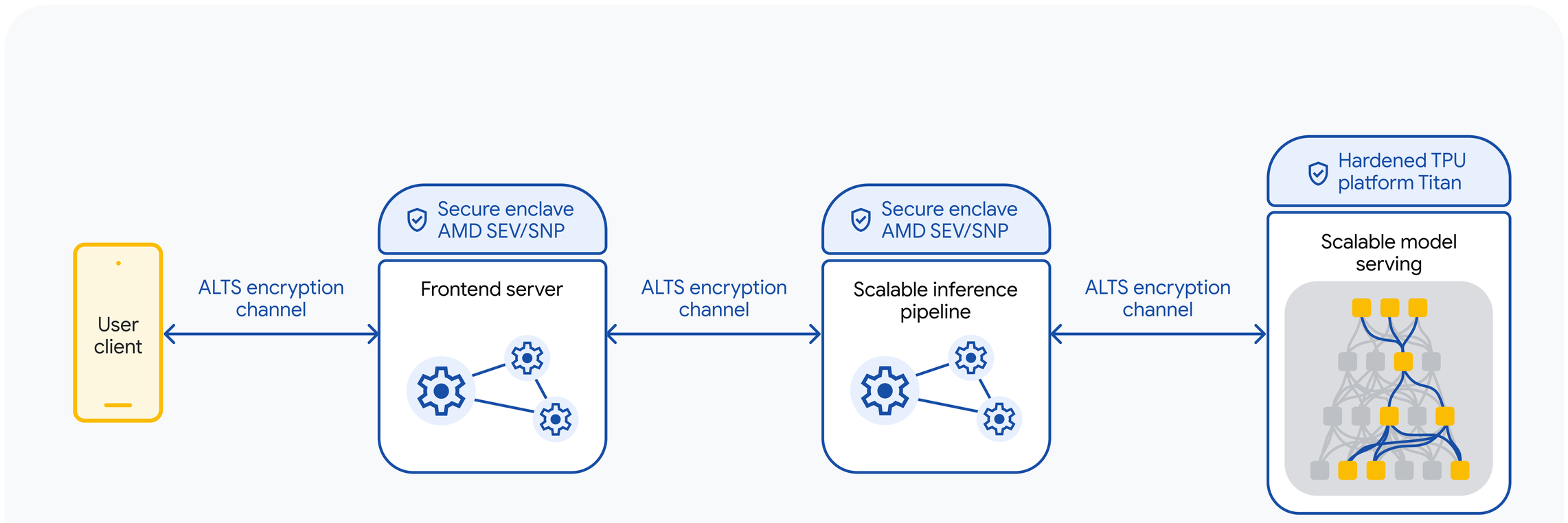

Figure 1: Simplified diagram to illustrate the Private AI Compute chain of trust.

Reduce the trusted computing base of AI services

Private AI Compute minimizes the number of components and entities that must be trusted for data confidentiality, to only the essential components required to maintain system integrity and protect sensitive user data.

Ensure private telemetry and analytics

Deployments of Private AI Compute that require analytics and aggregate insights make use of privacy-enhancing technologies (PETs), such as confidential federated analytics, to ensure that only anonymous statistics (e.g. differentially private aggregates) are visible to Google. With confidential federated analytics, unaggregated data is processed by open source software protected by hardware TEEs, currently Oak on AMD SEV-SNP with a hardware root of trust. The privacy properties of confidential federated analytics are transparent and can be verified by external parties.

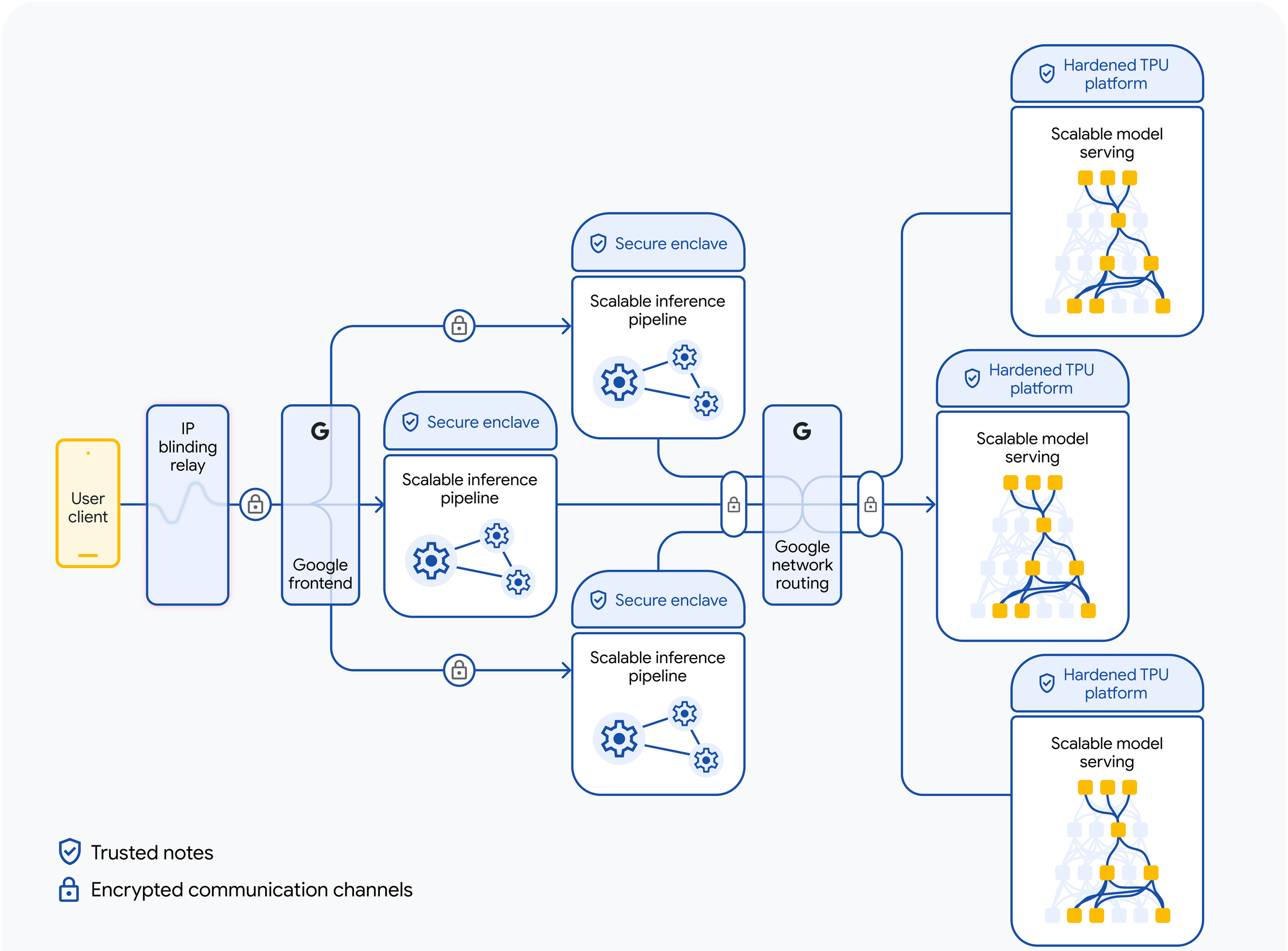

Figure 2: The system architecture of Private AI Compute.

02

Protections against privacy and security threats to user data access

In addition to the above privacy engineering measures, we know that strong privacy requires strong security to maintain the integrity guarantees of the system, including malicious modification of its privacy properties.

Private AI Compute is built on the foundation of our existing leading security infrastructure, including encryption for client-server communications; and binary authorization ensuring only signed, authorized code and validated configurations are running across our software supply chain. Binary Authorization for Borg, for example, provides similar security guarantees to the industry standard SLSA framework. Robust audit trail logging provides the foundation for all these measures, ensuring clear identification and comprehensive review of any security-relevant events or deviations.

As the sections below detail, Private AI Compute is implemented on top of this robust foundation and goes further – to defend against security risks such as insider access, service exploitation, and data exposure. Private AI Compute isolates user data in Virtual Machines (VMs) to contain compromise in case of a bug, hardens our systems against physical exfiltration with memory encryption and input/output memory management unit (IOMMU) protections, and uses hardware and software attestation to ensure system integrity.

Protections for privileged access misuse

Insider threats can happen when privileged access is abused to gain unauthorized access to data. For example, an insider could gain unauthorized control over critical system functions, make unauthorized changes to exfiltrate user data; or the insider can identify and target specific users or data flows and try to redirect and manipulate them, using knowledge such as client IP addresses or user credentials.

Mitigations

Private AI Compute is designed to implement the following protections against any single or coordinated insider access.

- **Ephemeral by design.** The system is designed so that inputs, model inferences, and computations are only kept as long as needed to fulfill the user's query. Attackers cannot access past data. User data is processed in a protected execution environment at the time of inference request and discarded when the user session is completed.
- **No privileged access to user data.** Private AI Compute is designed so that administrative access to user data is not possible. This achieves the highest level of controls for production services, preventing human access to the data even in "break glass" emergency scenarios. We do this with:
  - **TEE on CPU platform:** The CPU workloads are hosted in confidential virtual machines, a hardware-isolated, memory-encrypted environment shielded from the underlying infrastructure. This ensures user data protection during processing because the workload in a guest virtual machine is protected from the host.
  - **Zero shell access on hardened TPU platform:** Arbitrary execution (e.g. shell access) is not possible on nodes running in the Private AI Compute system, including our hardened TPU platform.
  - **Physical security:** Google's best-in-class, multi-layered physical security ensures malicious attackers cannot access data centers globally.
  - **Run key services on a state-of-the-art confidential computing platform** based on AMD's hardware Trusted Execution Environment. Frontend services run in a confidential virtual machine: a hardware-isolated, memory-encrypted environment shielded from the underlying infrastructure. This ensures user data protection during processing because the workload in a guest virtual machine is protected from the host and the code is verified via attestation.
- **Verify server authenticity to prevent eavesdropping and tampering.** Before any data exchange occurs, clients establish trust with a Private AI Compute endpoint through validating the endpoint server's identity. This relies on remote attestation and cryptographic key validation, along with a secure session protocol, to confirm that the endpoint server is genuine, unmodified, and actively protected by the referenced security measures.
- **Non-targetability with IP protection.** Even if an attacker were to defeat all other defenses in place, they would still need to distinguish a specific user's data from the immense volume of total network traffic to intercept it. We make this computationally infeasible.
  - **IP-blinding relay:** We use IP blinding relays which are operated by third-parties to tunnel all traffic bound for the Private AI Compute system. This removes the ability for an attacker to link your IP address–or any other network identifying information–to your specific query, and thereby removes the ability to distinguish one user's traffic from another.

  - **Isolation of authentication and authorization:** We isolate the system's authentication and authorization from inference using Anonymous Tokens. Device authentication and rate limiting are handled by a separate server, which exchanges device credentials for an anonymous, cacheable token to manage usage for the entire system. This token is presented to the Private AI Compute server to authenticate the user without revealing their identity. A user's identity never travels with their data on the inference path.

- **Ensure detection of cybersecurity incidents.** To protect against emerging threats, detection analytics leverage relevant security signals, enriched by frontline threat intelligence and expertise from Google's Threat Analysis Group and Mandiant teams. This ensures a deep understanding of evolving adversary techniques. By integrating Gemini's advanced reasoning, the analysis is accelerated to proactively detect and mitigate incidents based on real-world attack patterns.

Protections against data exposure from unintended errors or vulnerability exploits

Unintended errors in computing systems can lead to data exposure, and these errors can include software bugs, hardware malfunctions, misconfigurations, or even unexpected interactions within complex systems. This can lead to unintended logging, sharing user data with monitoring tools, and inadvertently granting human access.

Additionally, potential attackers can exploit flaws in the hardware components, as well as vulnerabilities in the operating system, hypervisor, and application code to bypass security mechanisms, and grant unauthorized access to system memory and thus user data.

Mitigations

In order to deliver robust protection against both unintended errors and vulnerability exploits, Private AI Compute is designed with the following measures.

- **Carefully control the input and output from Private AI Compute system jobs.** Granular egress policies have been put in place to prevent sensitive data from leaving the system across diverse channels such as monitoring, data logs, or core dumps. These policies cover each data point according to its sensitivity, with a default of no egress to ensure user data cannot be accessed.

- **Safeguard service integrity through VM isolation and runtime protection:** Private AI Compute services run on confidential computing and hardened TPU platforms with hardware roots of trust. Confidential computing uses the virtual machine layer to isolate services and dependencies to ensure a single instance compromise can be contained and promptly remediated, preventing adverse effects on other services.

- **Robust to continuous availability and updates.** Private AI Compute leverages Google's best-in-class secure infrastructure, which pushes out frequent updates to address security and reliability issues, such as Distributed Denial of Service (DDoS) prevention and compliance. Private AI Compute is engineered to preserve privacy commitments actively across updates that address security vulnerabilities, optimize performance, and deliver new or improved features.

03

External verifiability

We believe it is critical that users can trust in the claims we make about our software.

This means ensuring these claims can be verified for the software that is used to process their data, and be verified as the same software that third parties audit.

This initial release of Private AI Compute ensures that the software processing user data can be both externally audited and cryptographically attested. The following sections describe the assurances available through third-party expert audits, system-level attestation, and published ledger of production binary digests.

Third-party privacy and security audits

In order to ensure that the privacy of the system is verifiable by others outside of Google, we enable third-party privacy and security audits of our system binaries and sourcecode. Expert review provided a detailed analysis of and recommendations for our system. We are committed to continuously improving our security posture. A summary of audit findings is available here.

Attestation: establishing a verifiable binary identity for our compliant servers

Google's production infrastructure ensures that all software and hardware on the machine is authorized for the Private AI Compute system. Private AI Compute application binaries are only scheduled on machines that are running intended privacy-preserving software. Its critical components are available for third party audit and digests are published to a ledger.

Private AI Compute enables internal-facing system-level attestation from first mutable code to the OS, anchored in a hardware root of trust based on Titan, and using a Google internal keystore and authentication service to authenticate between the servers. Upon successful attestation using factory-provisioned keys bound to Titan, Google provisions a secret key to server instances that adhere to the stated assurances. Clients will only communicate with server instances that can demonstrate control over this key. This allows clients to have a cryptographic anchor for the server binary identity and maintain an encrypted, authenticated channel between the client and the protected execution environment.

04

Roadmap of Verifiability

We believe transparency and verifiability are key pillars of trust.

To this end, we have made the underlying binaries for platforms and features like Android, Private Compute Core, confidential federated analytics, and our ML suite available for public inspection and review. As a first step with Private AI Compute, we offer transparency for our deployed binaries by publishing cryptographic digests (e.g. SHA2-256) of those binaries in a published ledger. Digests of application binaries used by Private AI Compute servers are published in the ledger before serving traffic. Within the protected execution environment, system identity is verified as authorized to handle private workloads before any user data is processed. Additionally, users can see when Private AI Compute is being used on their Pixel devices – Private AI Compute requests are visible in their Settings Network Logs.

Building on this foundation, future releases of Private AI Compute plan to include:

- **External inspection of remote attestation verification.** To facilitate verification and mitigate the potential threat of undiscoverable modification of the system, we will be enabling experts to inspect remote attestation evidence from the client and server, and confirm that it only includes binaries that were specifically endorsed by Google for the purpose of Private AI Compute and are inspectable by third parties as described above.
- **Continued third-party audits** and expanded support for code and binary inspectability.
- **Expanded security research programs.** Our Vulnerability Rewards Program will be expanded to specifically include Private AI Compute. We will invite privacy and security experts to investigate our system and reward them for discovering potential vulnerabilities.

We look forward to sharing more information about Private AI Compute with you, including additional capabilities that will run on this privacy-preserving infrastructure.