Confidential Neural Computing:

Generative AI workloads in a Trusted Execution Environment

Joe Woodworth - Google Research

Agenda

Private Gen AI: Motivation & Risks

Core Technical Components

Ongoing explorations

Private Gen AI

Motivation & Risks

Generative AI

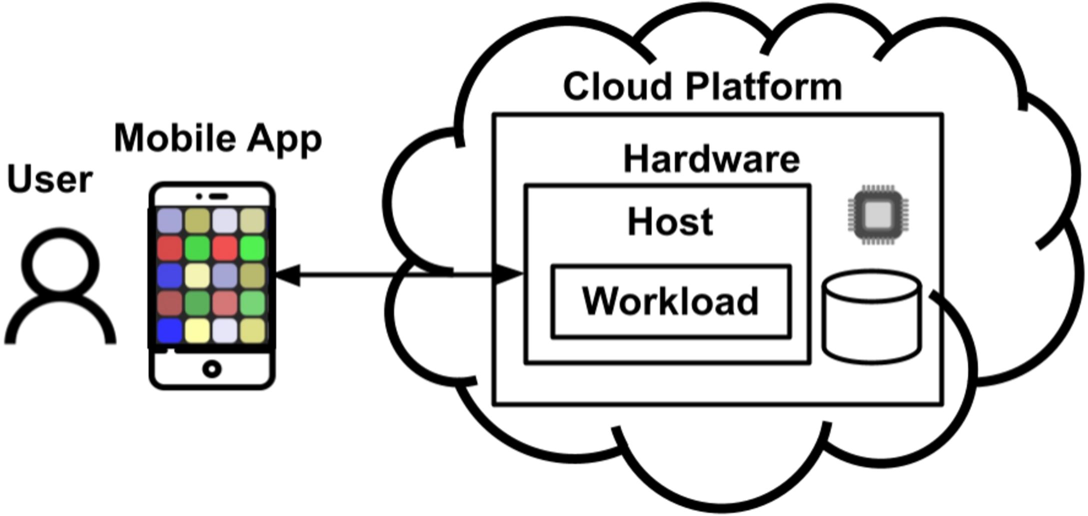
- Generative AI models are growing more & more capable
- Increased demand to integrate these models into products in ***personalized*** ways
- Personalized gen AI processes user data for inference & training
  - Potential dependency on sensitive & ambient data
  - Gen AI based applications could use e.g. screen content, camera, microphone, chat messages, etc.

Computational Scale

- Today's top Generative AI models are ***LARGE***
- Inference workloads
  - often require low latency + high throughput
- Training workloads
  - long running, large datasets, resource intensive

- Running large scale workloads on device is not always feasible
  - Some workloads must be run on a remote server

Privacy risks



Core Technical Components

Terminology

Confidentiality

- information is not made available or disclosed to unauthorized individuals, entities, or processes

Privacy

- an individual or group can control their information or data, and share it selectively

Transparency

- the implementation & execution of a process is visible to & verifiable by individuals or groups

Data Protection

Data is exposed to risk in all states

Data at Rest
- encrypted storage, access controls

Data in Transit
- network protocols, secure communication channels

Data in Use
- confidential computing

Trusted Execution Environment

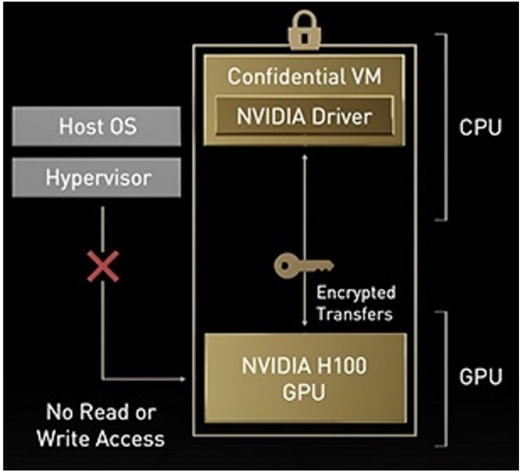
- Hardware-based, secure, isolated area within a device's processor
- Protects ***data in use***
- TEEs help protect against vulnerabilities or malicious code in the Cloud Platform
    - Confidentiality - data in the TEE cannot be accessed from outside the TEE, even by the OS
    - Integrity - code in the TEE cannot be tampered with & runs only as intended

Confidential CPU computing

- AMD SEV SNP
  - Full encryption of a virtual machine's memory with a unique key
  - Protects against snooping from the hypervisor or other VMs on the same host
- Intel TDX
  - Uses isolated virtual machines called Trust Domains (TDs)
  - Uses new CPU instructions & memory management to enforce isolation & attestation of TDs

Confidential Accelerators

- **H100 GPU** supports Confidential Compute Mode
  - even during processing, data is inaccessible from the host CPU, operating system, hypervisor
- H100s are in high demand
  - cost & hardware availability are concerns
  - we're investigating alternatives, e.g. Intel AMX CPU-based acceleration



Remote Attestation

Users want to verify what the workload processing their data is actually doing

A **transparent release** process yields reproducible, externally verifiable builds of the TEE container workload

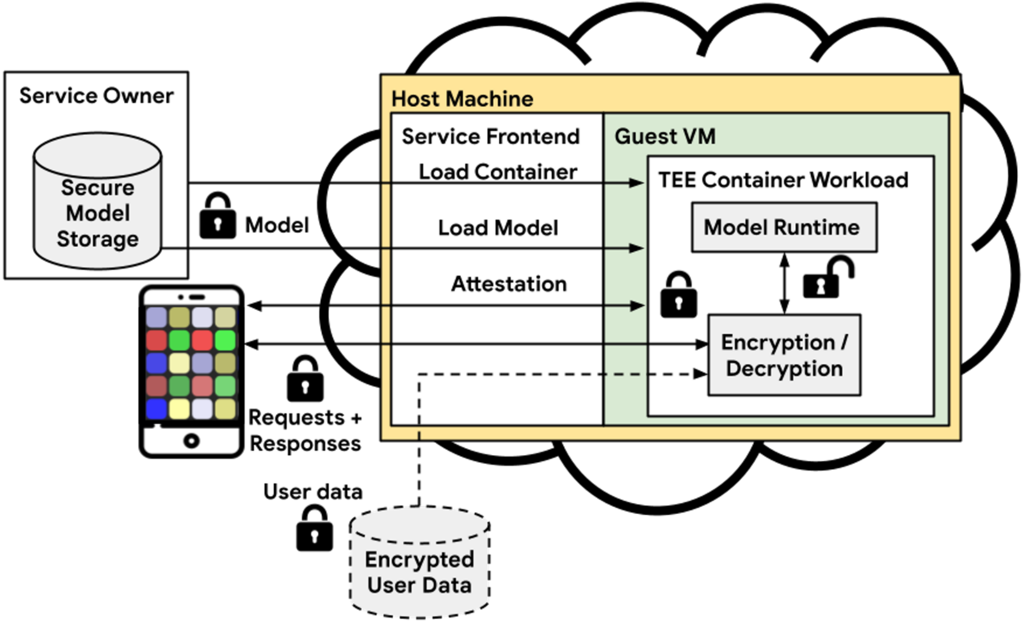
- **Attester**
  - Remote machine undergoing verification
  - When **challenged**, securely communicates **evidence** with **Verifier**
- **Verifier**
  - Stores database of known good measurements (reference-values)
  - Compares **Attester's** evidence with reference & generates **Attestation Report**
- **Relying Party**
  - Client, that trusts the **Verifier,** and relies on **Attestation Report** to determine if the **Attester's** state matches expectations

Confidential Neural Computing

An ML framework that enables generative AI training and inference in secure enclaves

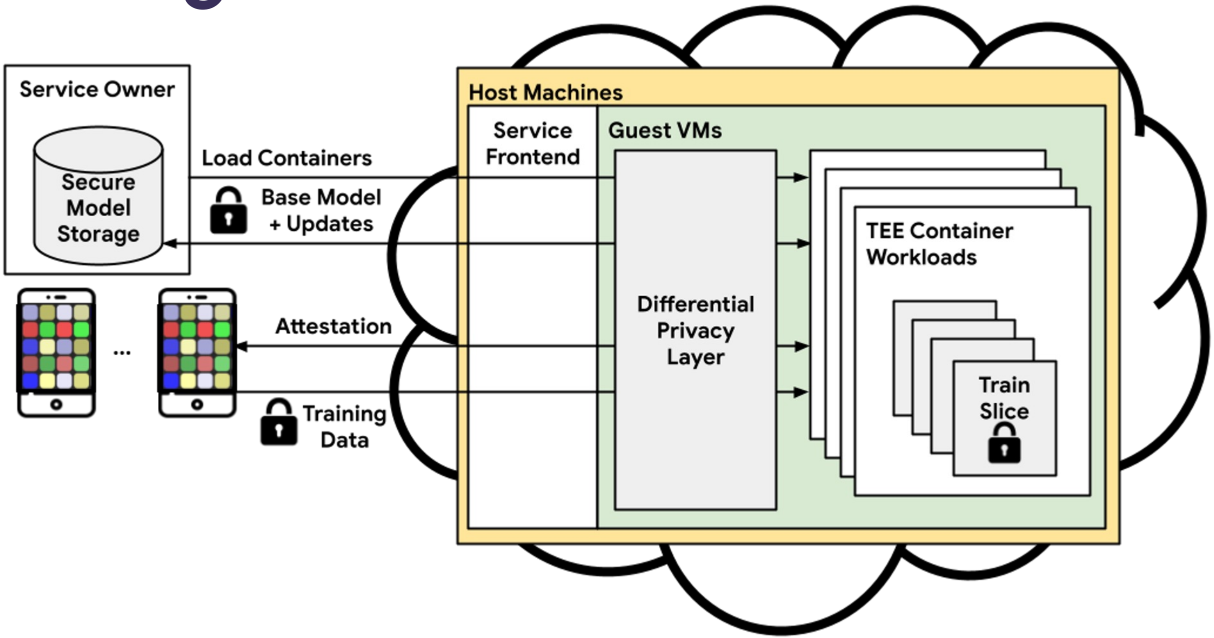
- targets **Trusted Execution Environments**
- leverages **confidential CPU + accelerators**
- is built for **Remote Attestation**
- supports **Privacy via Confidentiality + Transparency**

Inference



Service Owner
Secure Model Storage
Model
Host Machine
Service Frontend
Load Container
Load Model
Attestation
Guest VM
TEE Container Workload
Model Runtime
Encryption / Decryption
Requests + Responses
User data
Encrypted User Data

Training



Service Owner
Secure Model Storage
Load Containers
Base Model + Updates
Attestation
...
Training Data
Host Machines
Service Frontend
Guest VMs
Differential Privacy Layer
TEE Container Workloads
Train Slice

Differential Privacy

- For training, we want the model to learn from realistic samples of user data, ***without*** learning individual private information
- Differential Privacy prevents this by introducing controlled noise into datasets
- Appropriate privacy guarantees can be made through adjusting noise based on ε & δ values
    - ε = the Privacy Loss Budget
    - δ = the failure probability
- Limitations
    - privacy - accuracy tradeoff
    - computational cost

Ongoing explorations

High Performance AI in TEE

- Compute platform & hardware
  - Benchmark & optimize performance for confidential H100 GPU, Intel AMX
  - Target multi-GPU & multi-node environments
- Confidential frameworks
  - Google Cloud: Confidential VMs & Confidential Space
  - Different configurations with Project Oak, e.g. on-prem solutions

ML Infrastructure for Privacy

- Support Private Inference & Private Training
- Attestation and end-to-end encryption between model service & client
- Private model artifact protection
    - Public infrastructure dynamically loads the private model via encrypted channel
    - Infra can impose constraints to the dynamic model
- Training pipeline with private data protection
    - Integrates with Differential Privacy to efficiently run workloads in TEE with accelerators

Contact us!

project-cnc-team@google.com