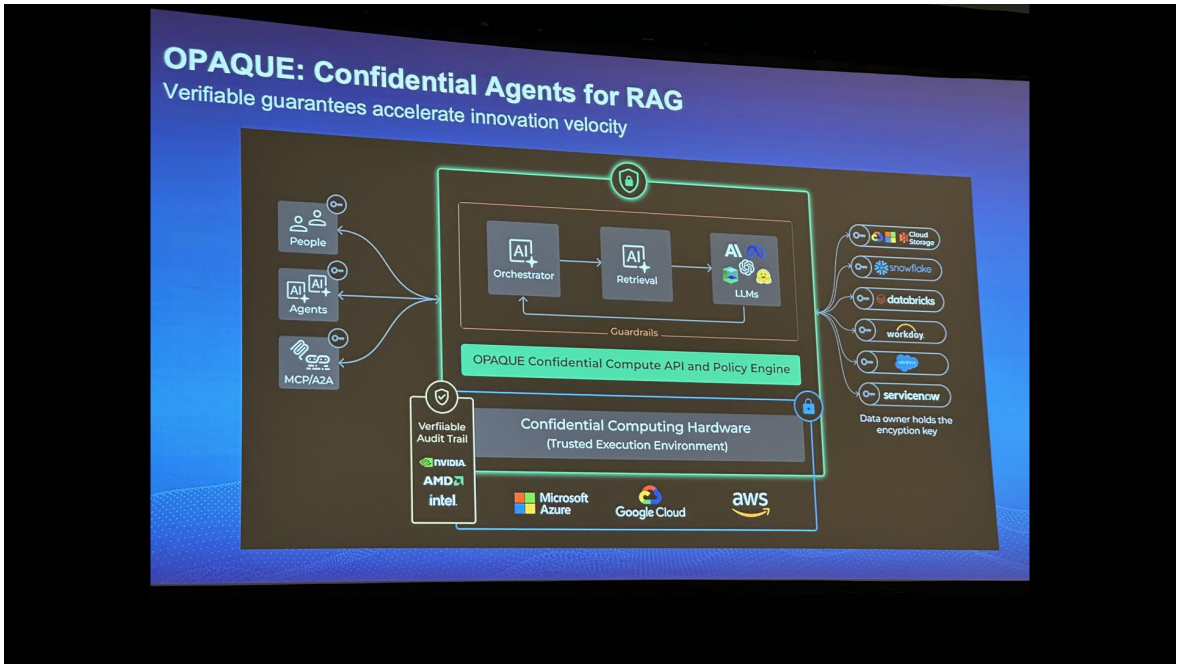
Action Items & Further Reading

- [ ] Add link for Charlie Bell's "expanding microsoft's secure future initiave"
- [ ] Read more about credit card chips being TEEs
- [ ] Read "Why should I trust your code?" whitepaper
- [ ] Read C-FedRAG paper: arXiv:2412.13163

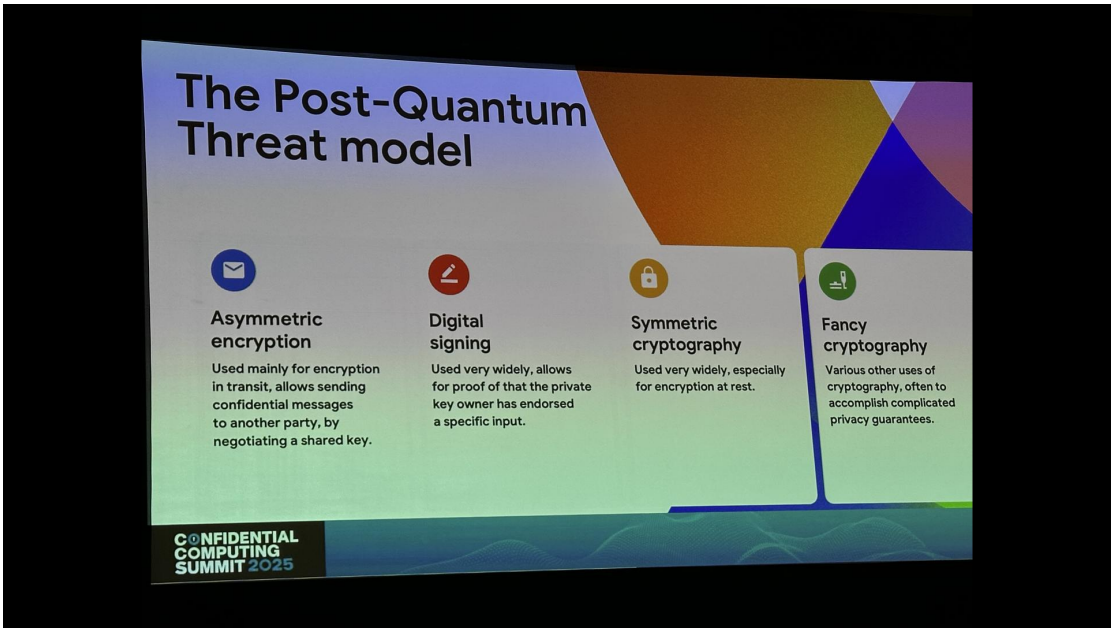
Title: Trust at machine speed: building the future with verifiable infrastruature

- **Speaker**: Aaron Fulkerson from Opaque
- **Core Idea**: Agents work at machine speed, not human speed. This allows for massive efficiency gains, but also means they can cause damage at machine speed. The key is to build in verifiable guarantees to accelerate innovation velocity safely.
- **Opaque: Confidential Agents for RAG**: Opaque provides a platform for confidential AI agents, particularly for Retrieval-Augmented Generation (RAG). The system architecture ensures that data owners retain control of their encryption keys, and all processing happens within a confidential computing TEE, with a verifiable audit trail.
    - **Architecture Overview**: The system allows users (people, agents, machines) to securely interact with an AI pipeline (Orchestrator, Retrieval, LLMs) that processes data from various enterprise sources (Snowflake, Databricks, Salesforce, etc.). This entire process is enclosed within a confidential computing hardware environment (TEE) supported by major cloud providers and hardware vendors, ensuring that data is protected and all actions are auditable.



Title: Confidential computing in a quantum world

- **Speakers**: Nelly Porter and Sam Lugani from Google Cloud
- The talk introduced **Willow**, a 2024 Google quantum computing chip. A key point was the immense power of quantum computers and the threat they pose to current cryptographic standards.
  > Willow is very interesting character, because it can perform incredibly fast... it's very powerful thing... the interesting thing about the component is built in willow node, is that great power is catastrophic. And it's very important because there's every computer that operate is expansion.
- **Urgency of Migration**: Crypto migration is a complex and time-consuming process. The speakers emphasized that the post-quantum threat is not a distant problem and that organizations need to start preparing now.
  > The chances that this process is not coming sooner than later is absolutely high. Don't trust people to say, don't worry about this. I absolutely believe that it will happen in a very, very close time.
- **Post-Quantum Threat Model**: The threat impacts several areas of cryptography.

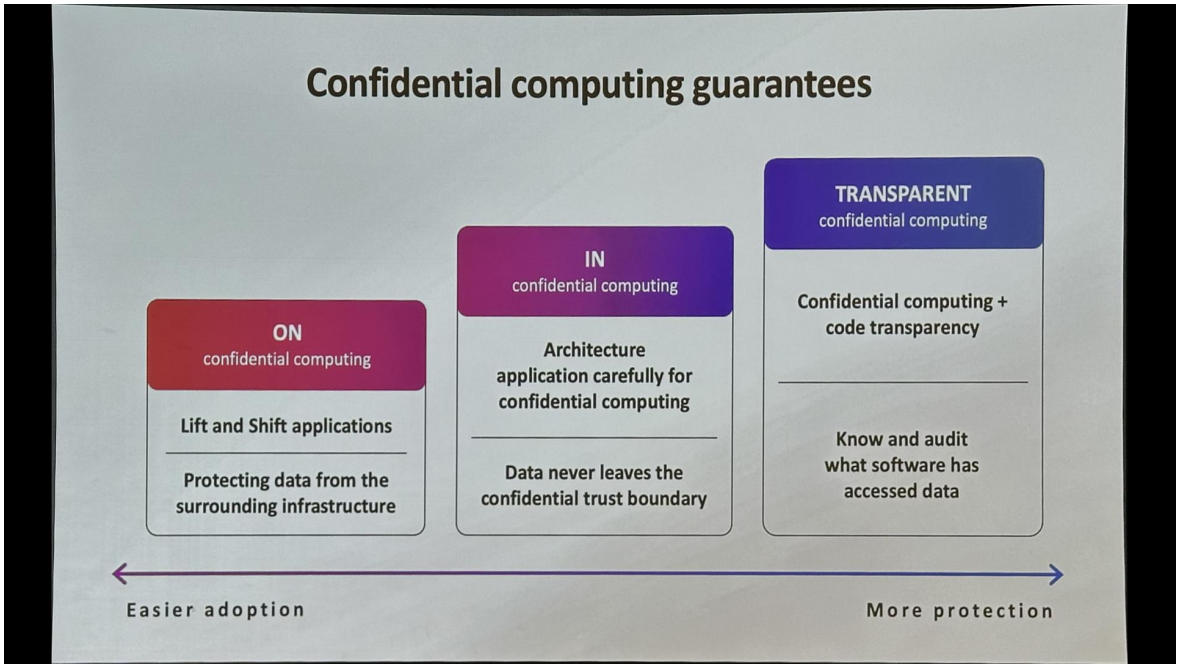


    - **Asymmetric encryption**: Used for negotiating shared keys for confidential communication.
    - **Digital signing**: Widely used to prove that a private key owner has endorsed a specific input.
    - **Symmetric cryptography**: Widely used, especially for data encryption at rest.
    - **Fancy cryptography**: Other complex cryptographic uses, often for sophisticated privacy guarantees.
- **Google's Quantum-Safe Initiatives**:
    - **Confidentiality in Transit**: Chrome and all public Google websites have migrated to quantum-safe encryption.
    - **Digital Signatures**: Google Cloud KMS is previewing support for the new FIPS-205 (ML-DSA)

      and FIPS-204 (ML-KEM) standards for quantum-safe digital signatures.

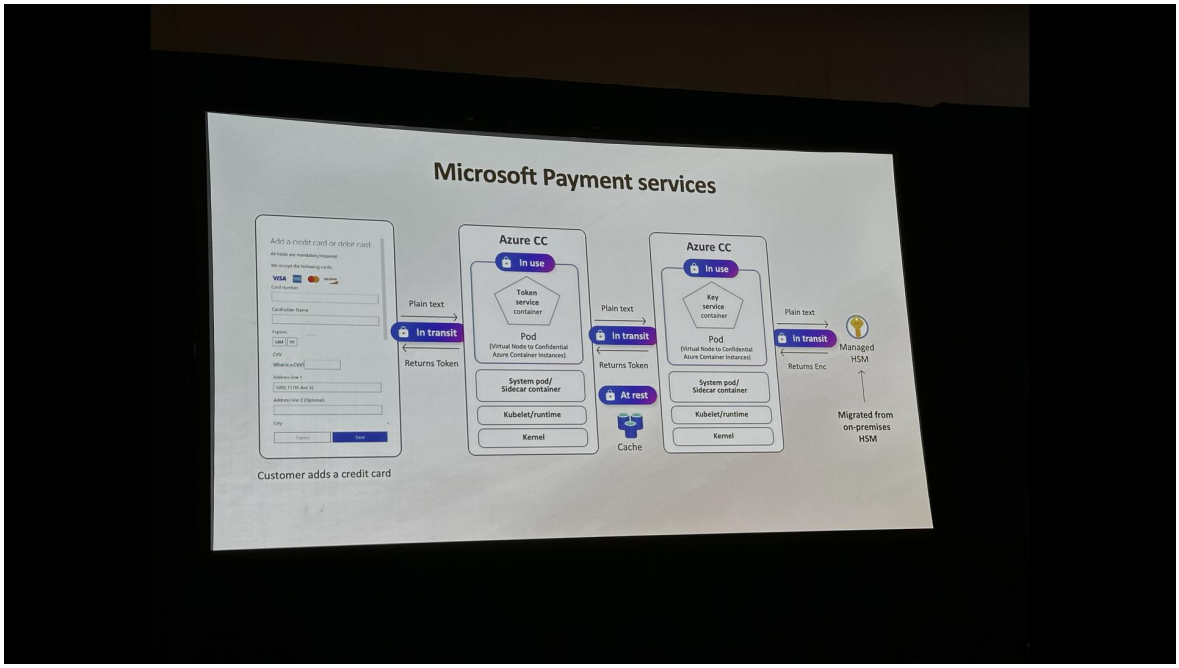
Title: Confidential computing: defining a spectrum of guarantees

- **Speaker**: Mark Russinovich from Microsoft Azure
- Mark introduced a spectrum to define different levels of confidential computing guarantees, which helps clarify the trade-offs between ease of adoption and the level of protection.



1. On Confidential Computing

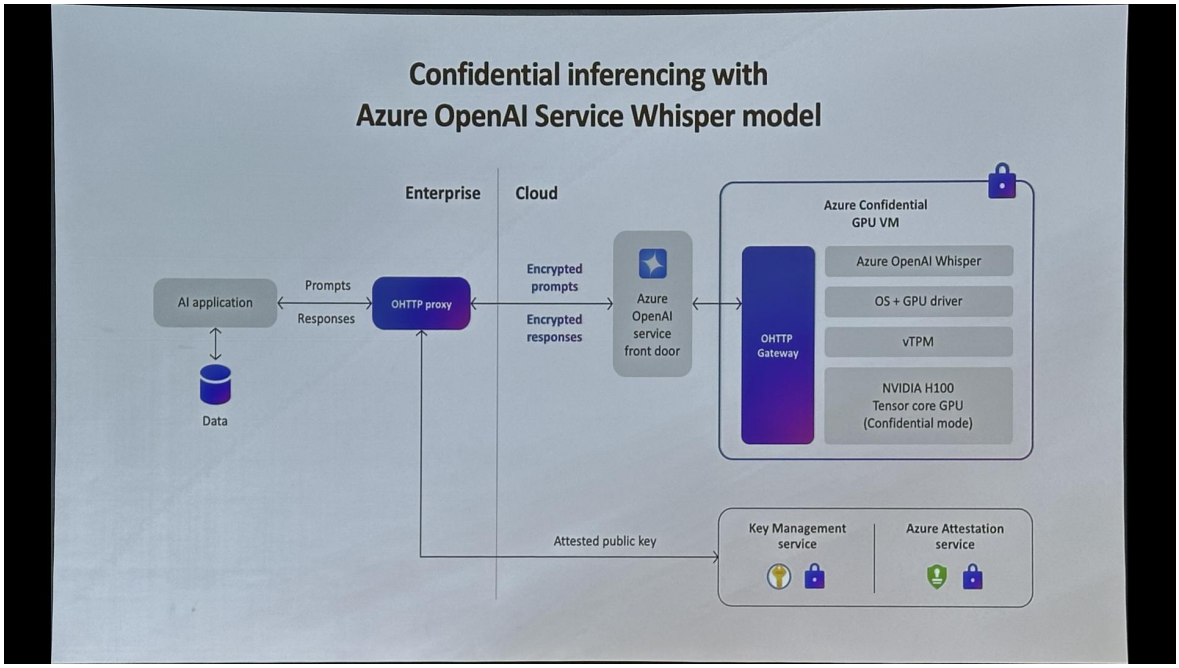
- This level focuses on **lift-and-shift applications**, protecting data from the surrounding infrastructure (e.g., the cloud provider) . It offers the easiest adoption path.
- **Security Above All Else**: This is part of a broader "Secure Future Initiative" at Microsoft, championed by Charlie Bell. The goal is to make security the top priority.
- **Use Case: Protecting Token Signing Keys**: A key service that was previously on-premises using an HSM was migrated to Azure Confidential Container Instances (ACI). The tokenization service and key service containers run in separate confidential pods, interacting with a Managed HSM to protect sensitive financial data like credit cards. This protects the data while in use from the cloud infrastructure.



2. In Confidential Computing

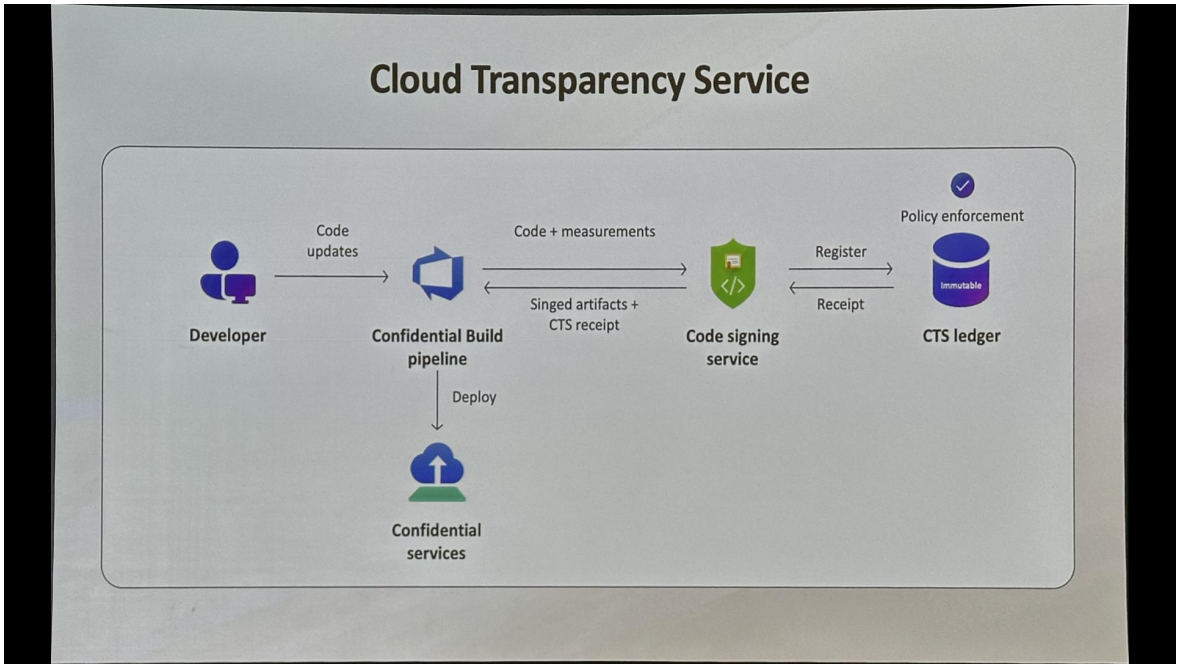
- This level requires applications to be architected more carefully for confidential computing, ensuring that data never leaves the confidential trust boundary.
- **Use Case: Confidential Inferencing with Azure OpenAI Service Whisper Model**: A financial customer needed to transcribe sensitive speech without exposing the data.
    - **Architecture**: Encrypted prompts are sent from the enterprise through an OHTTP Gateway to the Azure OpenAI service. They are processed inside an Azure Confidential GPU VM (NVIDIA H100)

      protected by another OHTTP Gateway. The VM's integrity is verified by the Azure Attestation Service, and keys for encrypting/decrypting the prompts and responses are managed by a Key Management Service. The result is an end-to-end encrypted channel where even Microsoft cannot see the prompts or responses.

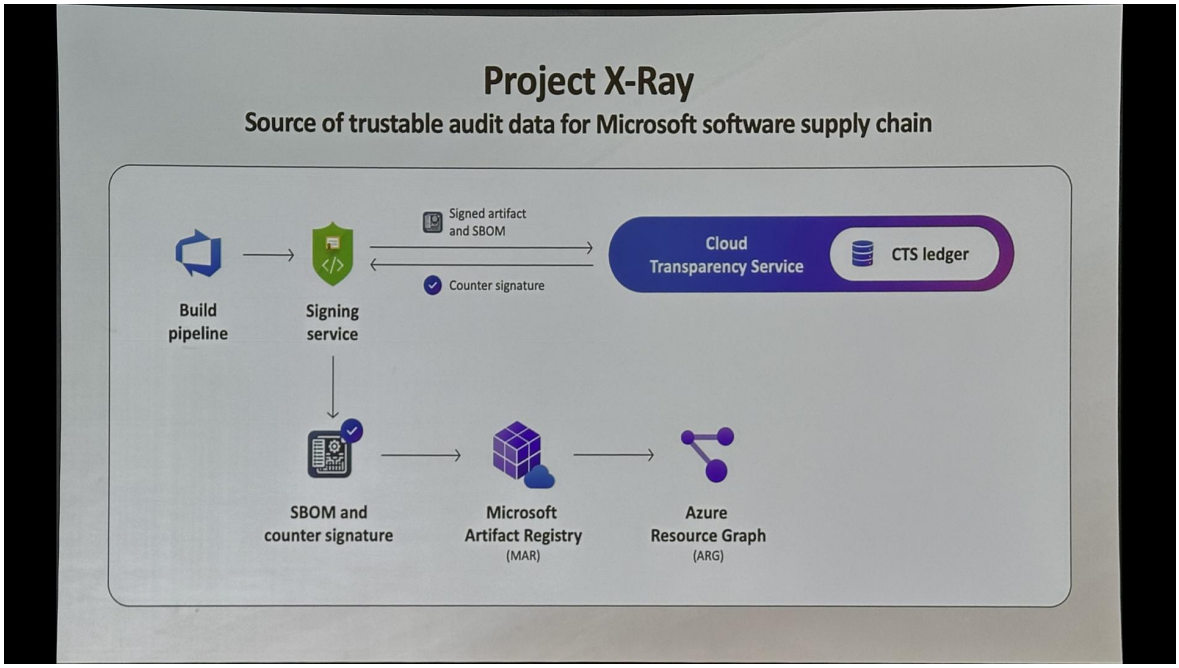


3. Transparent Confidential Computing

- This is the highest level of protection, combining confidential computing with code transparency. The goal is to allow users to **know and audit what software has accessed their data**.
- **The Problem**: How can you trust the library or binary that goes into the TEE? How do you detect misbehavior?
  > ...even auditing the code, if you don't know what you are looking for, you might not be able to find any signs of misbehavior. But if you suspect there's misbehavior and evidence of behavior, having the guaranteed ability to go find where that misbehavior located, and then hold the perpetrator is a very powerful disinfectant for something misbehavior. And so that's what we are going, is being able to provide guaranteed auditing for whatever has access to the data.
- **Cloud Transparency Service (CTS)*
  *: A service that provides a verifiable, immutable ledger of all software releases.
  - **Workflow**: A developer's code update goes through a confidential build pipeline. The code and its measurements are sent to a code signing service, which registers them in the CTS ledger after enforcing policies. The pipeline then receives a signed artifact and a CTS receipt, which can be used to gate deployments into confidential services.



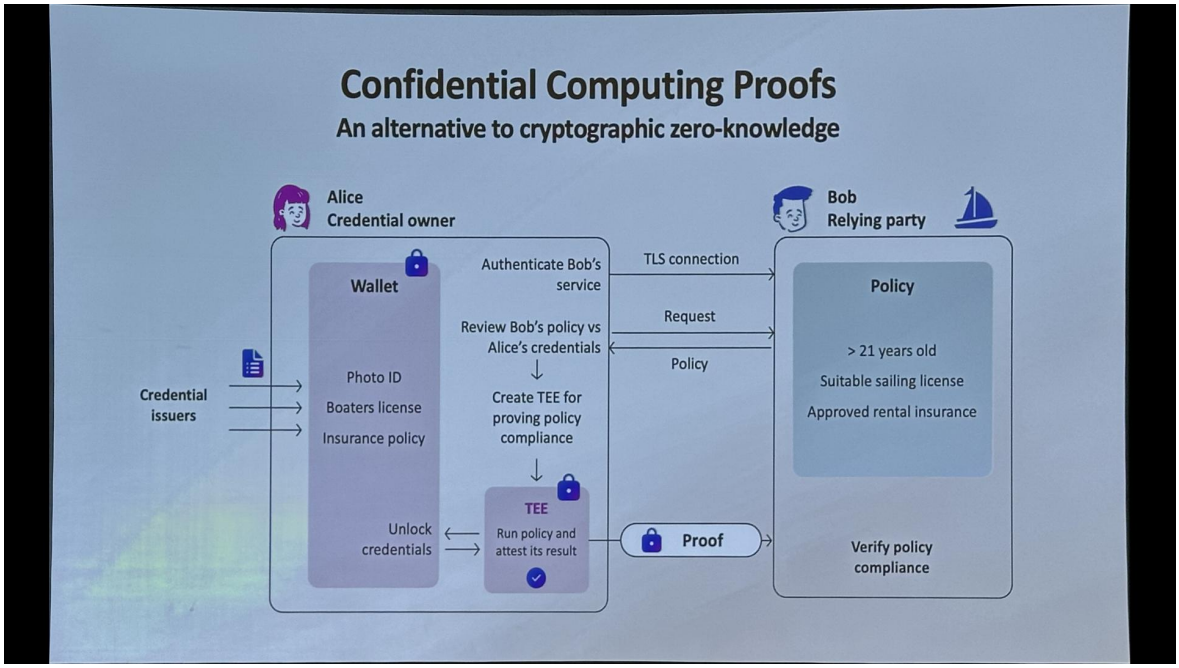
* **Project X-Ray**: This project uses the CTS to be the source of trustable audit data for Microsoft's software supply chain, signing SBOMs (Software Bill of Materials)
and storing them for auditing.



* **Confidential Computing Proofs**: An alternative to cryptographic Zero-Knowledge Proofs (ZKPs)
for proving policy compliance without leaking unnecessary personal info.

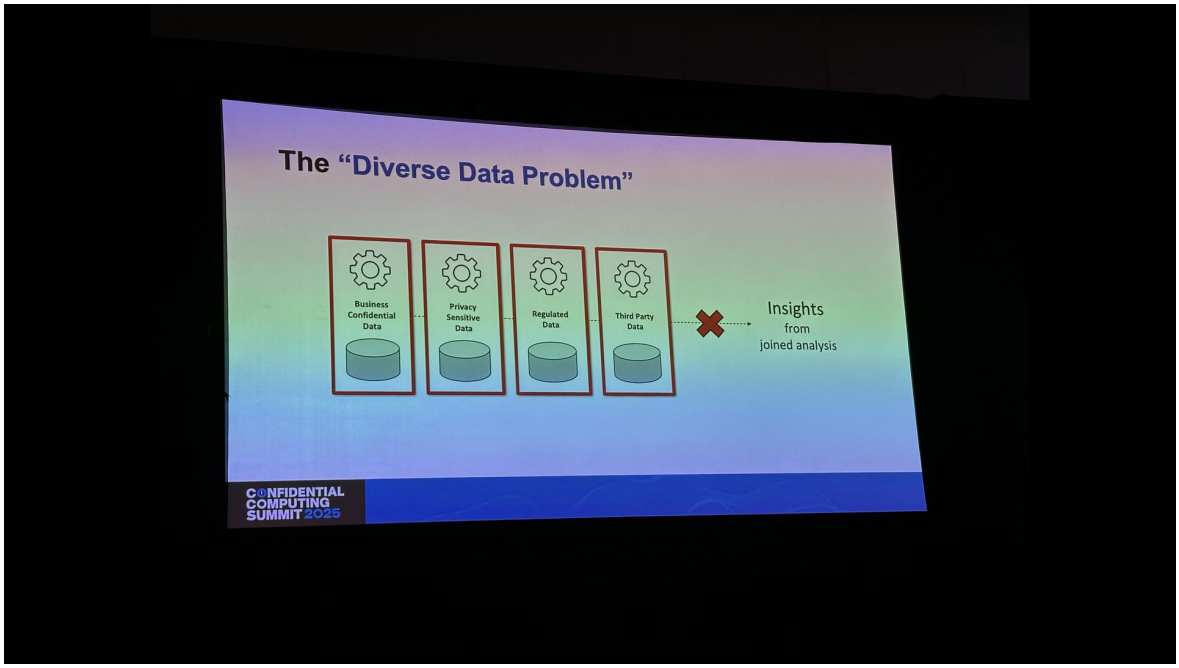
* **The Challenge with ZKPs**: They are computationally expensive, complex, and rigid (you need to know the specific information to prove ahead of time).

> *...you're having to give somebody a lot of personal information that's completely irrelevant to what you're trying to do... This is just the huge violation of our privacy.*
> ** **The Solution**: Use a TEE to run a policy check and attest to the result. For example, to prove you are over 21, your wallet unlocks your signed driver's license inside a TEE. The TEE runs a transparent, auditable piece of code that checks the age and outputs a signed attestation saying "True" without revealing your birthdate, address, or other PII.*



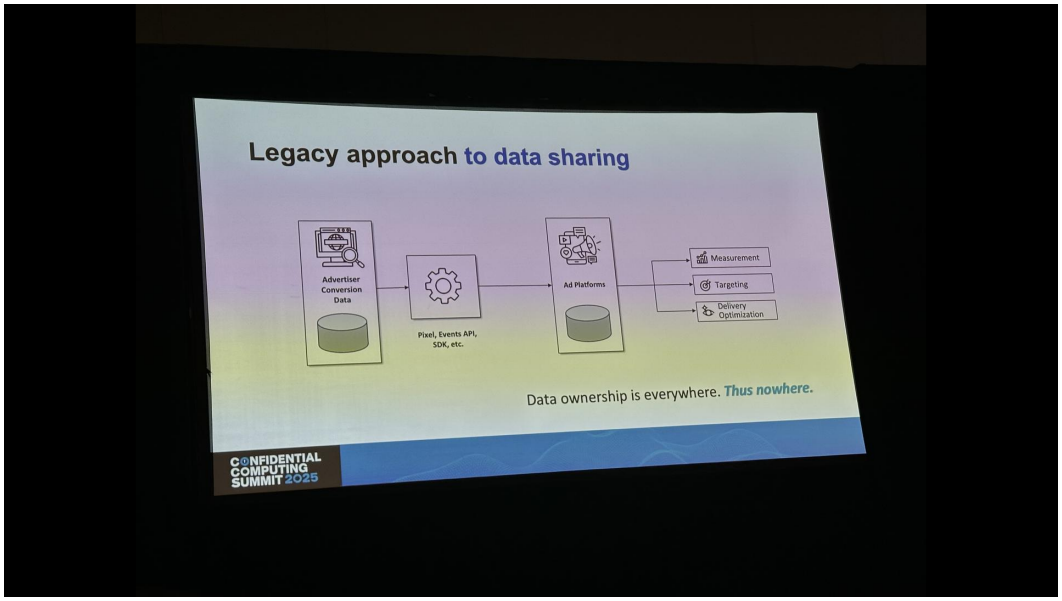
Title: Harnessing diverse data: how one advertising leader uses confidential computing to drive ROI

- **Speakers**: Mike Reed from Intel, Graham Mudd from Anonym at Mozilla
- **The Diverse Data Problem**: Businesses have many types of siloed data (confidential, privacy-sensitive, regulated, third-party) that cannot be easily combined for analysis due to security and privacy restrictions.

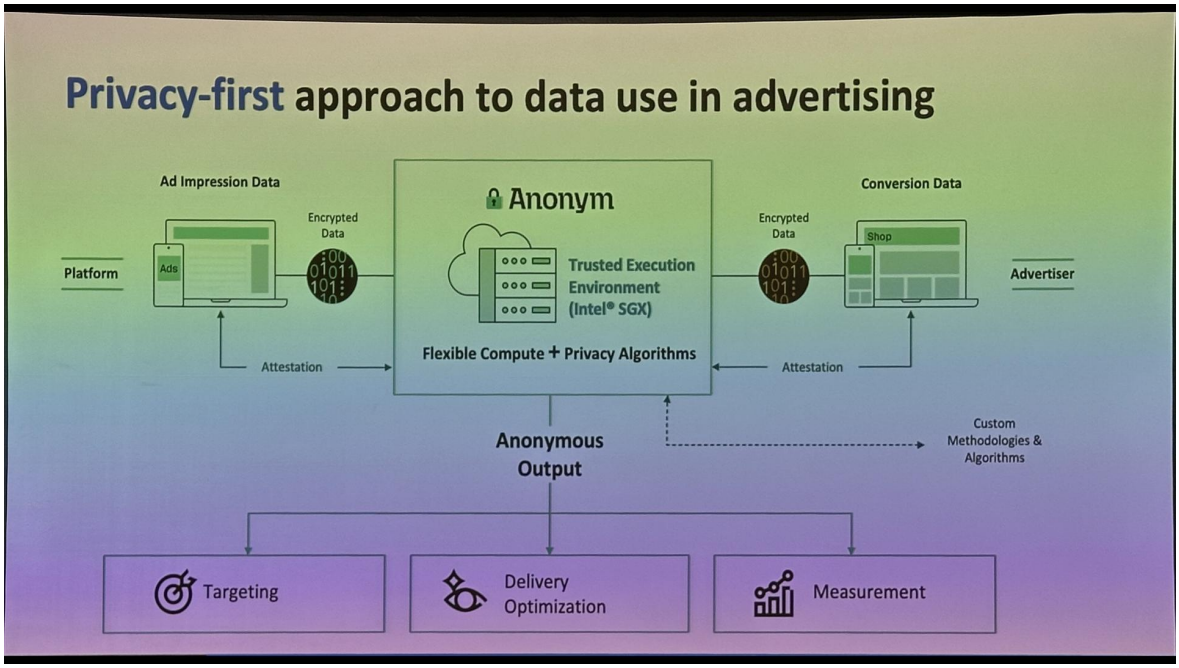


* **Legacy Approach**: The traditional advertising model involves extensive data sharing using tracking technologies like pixels and cookies. This data is used for measurement, targeting, and delivery optimization. The result is that "Data ownership is everywhere. Thus nowhere."
* The speaker gave a common example:

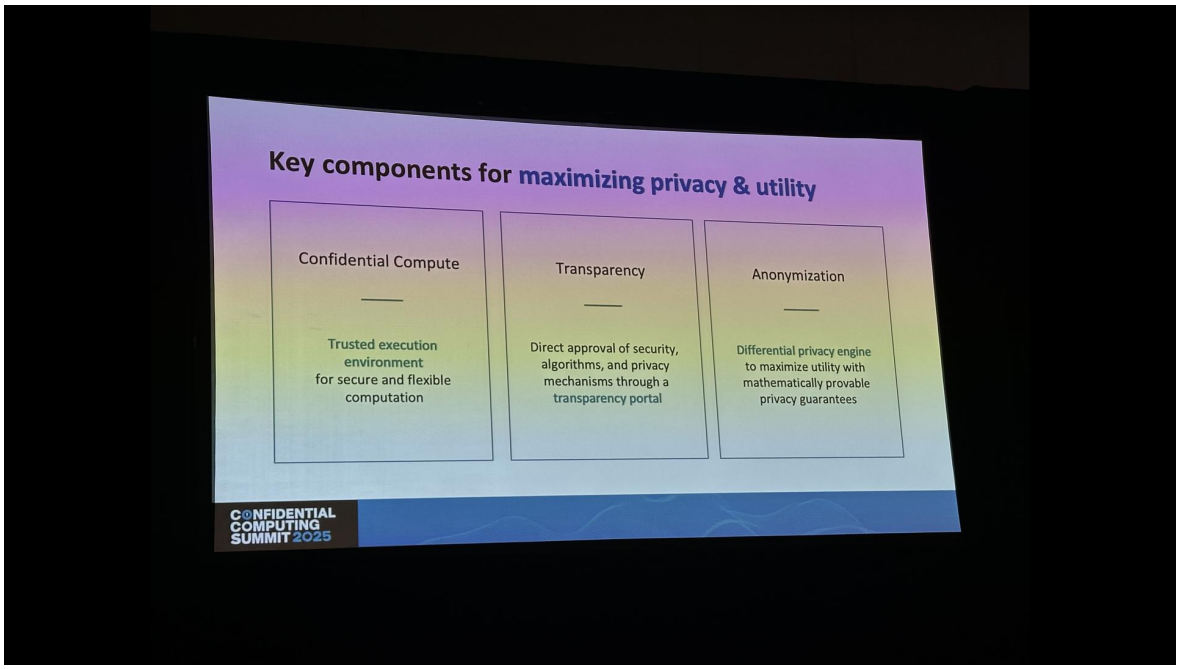
> *You go shopping for a pair of shoes, on a particular website... and then you start seeing that ad following you all over the internet, right? Well, that's because that apparel website is sharing more information with all the ad platforms they advertise on...*



** **Privacy-First Approach**: The new approach must provide the same business utility without compromising privacy.*
If you adopt a privacy-preserving technology, and that results in your ads working less effectively, your advertising will end up at a disadvantage... So we have to find ways to allow them to do things like measurement and targeting and delivery optimization but in ways that are equally as effective as what they're doing today.
** **Anonym's Architecture**:*
** Ad impression data and advertiser conversion data are encrypted at the source.*
** The encrypted data is sent to Anonym's platform, which runs inside an Intel SGX Trusted Execution Environment (TEE)*
.
** The TEE's integrity is verified via attestation.*
** Inside the TEE, the data is decrypted, and privacy-preserving algorithms are run on the joined data.*
** The system outputs anonymous, aggregated results that can be used for targeting, optimization, and measurement without ever exposing user-level data.*



* **Key Components for Maximizing Privacy and Utility**:



Confidential Compute	Transparency	Anonymization
Trusted execution environment for secure and flexible computation	Direct approval of security, algorithms, and privacy mechanisms through a transparency portal	Differential privacy engine to maximize utility with mathematically provable privacy guarantees

* **Learn More**:
* anonymco.com (Mozilla)

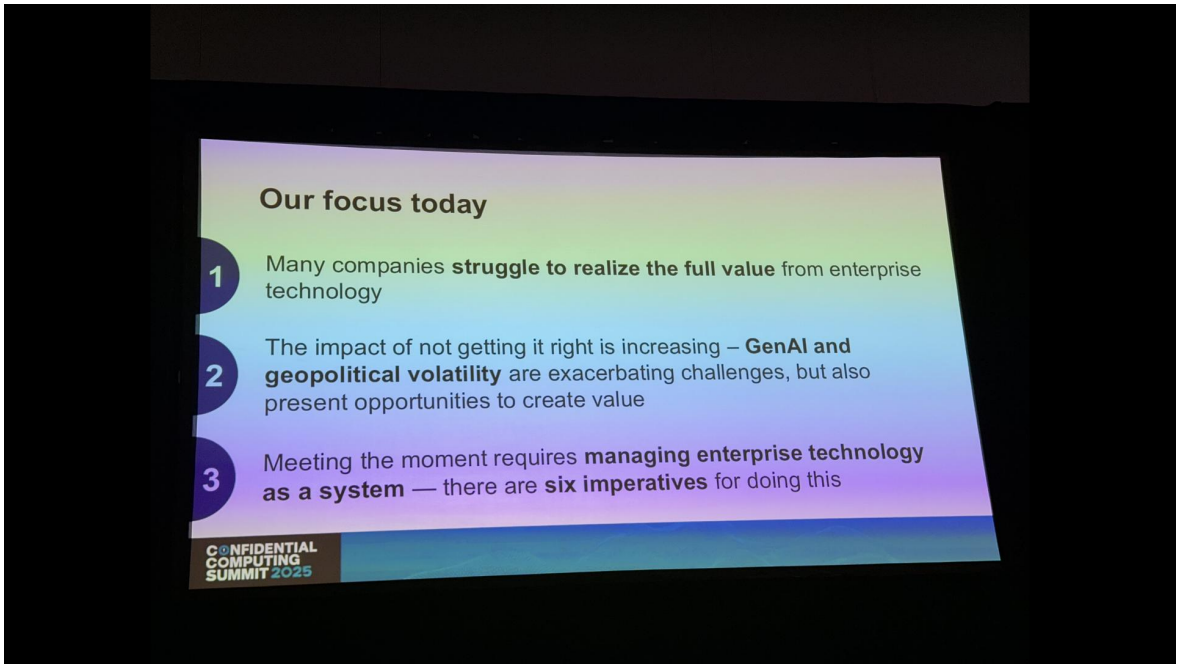
* `intel.com/confidentialcomputing`

Break, social time

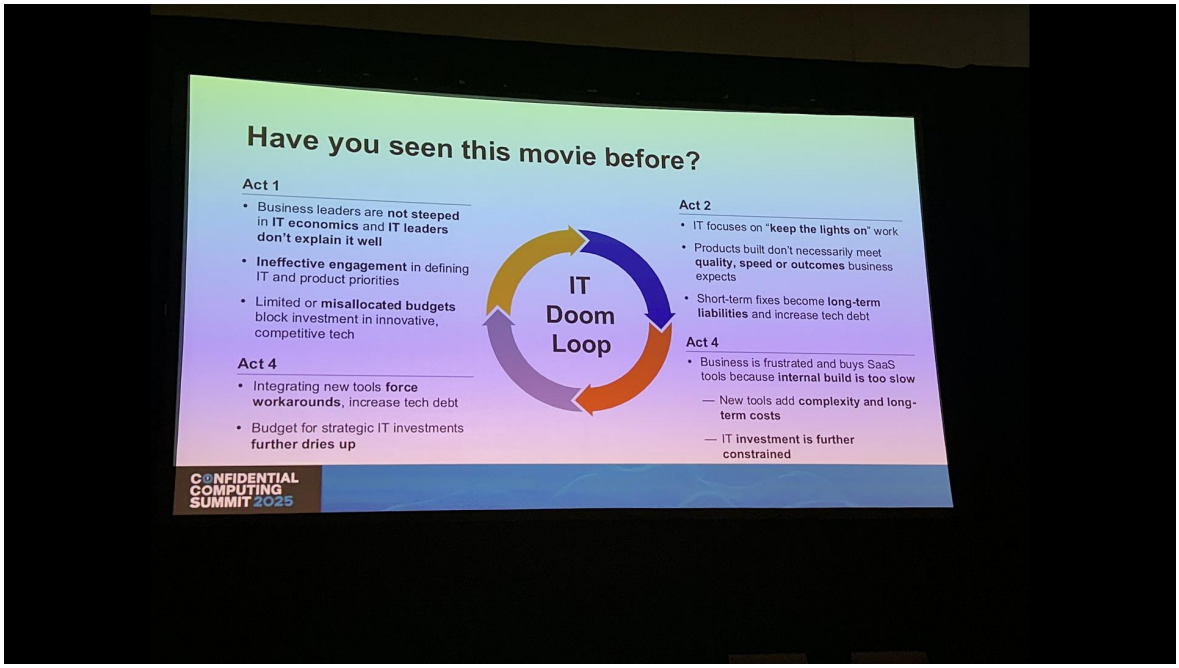
- **Chat w/ Dan (Engineer Director) from Intel**:
  - **TEE Memory Encryption**: All data in memory (RAM) is encrypted by default. The memory controller holds the key. When the CPU needs to process data, the controller passes the key to fetch and decrypt the data for the computing unit. Only encrypted data is ever written to or stored in RAM.
  - **Overhead**: This encryption/decryption process happens at RAM I/O time. Intel claims the performance overhead is around 3%, while Opaque (who builds solutions on top) estimates it's closer to 10%.
  - **GPU Support**: The Nvidia H100 supports TEE for memory on a single machine but does not currently support TEE across a multi-machine cluster. GPU memory is tricky because it doesn't provide general access like system RAM. Dan will discuss this more in his 3:30 PM session.

Title: The next generation technology agenda

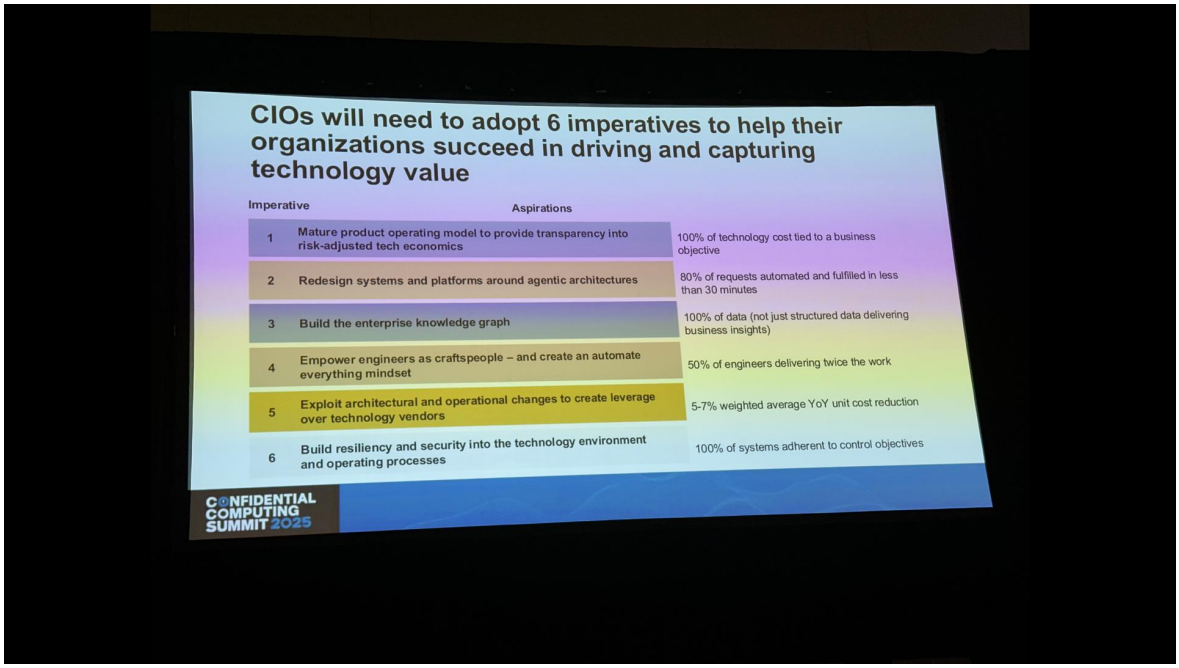
- **Speaker**: James Kaplan, CTO McKinsey
- **Context**: Enterprises face challenges from GenAI and geopolitical volatility, requiring them to manage technology as a unified system, not a collection of silos.



* **The "IT Doom Loop"**: A vicious cycle where businesses, frustrated with the slow pace of internal IT, buy external SaaS tools. This adds complexity and cost, further constraining the IT budget and increasing technical debt, which makes IT even slower for the next innovation cycle.



* **CIOs need to adopt 6 imperatives to help their organizations succeed**:



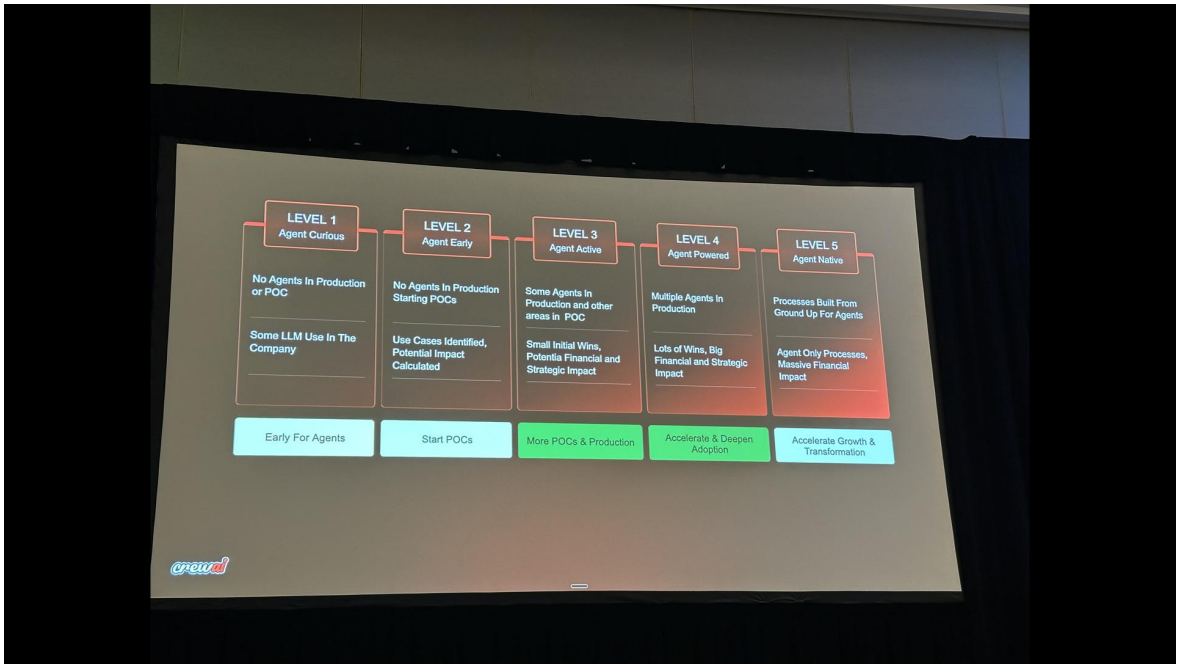
Imperative	Aspirations
1	Mature product operating model to provide transparency into risk-adjusted tech economics
2	Redesign systems and platforms around agentic architectures
3	Build the enterprise knowledge graph
4	Empower engineers as craftspeople – and create an automate everything mindset
5	Exploit architectural and operational changes to create leverage over technology vendors
6	Build resiliency and security into the technology environment and operating processes

* **Question**: Does your IT org run as a system or a series of silos? The speaker emphasized that all six imperatives are interconnected and necessary.

> *It's sort of like asking what's more important a car than an engine? Do I need an engine in building a car? Do I need an engine or do I need brakes or do I need a suspension? Like huh? You're not going very far without each of those things.*

Title: AI Agents at Scale: insights and trends from millions of agent interactions

- **Speaker**: Joao Moura from CrewAI
- (User notes this was pretty much the same as yesterday's workshop)
  .
- **Key Insight**: Enterprises moving beyond prototyping agents care most about three things:
  > ...the three things that companies really care about when this is stable, interoperability, observability, governance. That's it.
- **Agent Adoption Maturity Model**: The speaker presented a 5-level model for how companies are adopting AI agents, noting that most are currently at Level 1 or 2, with Levels 4 and 5 being largely aspirational today.



Level	LEVEL 1: Agent Curious	LEVEL 2: Agent Early	LEVEL 3: Agent Active	LEVEL 4: Agent Powered	LEVEL 5: Agent Native
Status	No Agents In Production or POC	No Agents In Production Starting POCs	Some Agents in Production and other areas in POC	Multiple Agents In Production	Processes Built From Ground Up For Agents

Title: (Panel discussion)

This panel, titled "Beyond Explainability: Why Provable Trust Is the New Foundation for Enterprise AI," discussed the real-world problems and adoption hurdles for confidential computing in the enterprise AI space.

- **What are the most real AI threats right now?**
  - **Mark Russinovich**: The biggest risks are prompt injection and uncontrolled information flows that violate policy. He cited a recent GitHub Copilot leak of private repo data into the public. Confidential computing can help by putting agents and their models inside trust boundaries to protect conversation histories and data.
  - **Nelly Porter**: The key problems are around the identity and authorization of agents. How do we distinguish agents from humans and ensure they are properly authenticated and attested?
  - **Sean (Joao Moura)**: The biggest vulnerability is the "agent swarm." We are creating a massive number of agents, but we don't have mature, secure protocols for how they communicate with each other. Each agent is another attack surface, and they are often created by people without deep security training.
- **Are customers willing to pay more for confidential computing?**
  - **Mark Russinovich**: "Our goal is to have that pricing be the same as not having it." He acknowledged some current cost drivers like workload migration challenges, but the goal is to eliminate the cost penalty so it becomes a default security layer.
  - **Nelly Porter**: Google has a small "confidential tax" but it's dropping. The value is in unlocking highly sensitive customers who previously couldn't move to the cloud. This opens up a "brave new world" of workloads.
- **What are the biggest points of friction for adoption?**
  - **James Kaplan**: "Do we even know how to take advantage of it?" Many enterprise development teams are unsophisticated and don't know how to architect applications for confidential computing in a way that is cost-efficient, secure, and performant.

    - **Nelly Porter**: A primary focus has been eliminating friction. The "lift and shift" model where customers can just check a box to make a VM confidential is a key strategy.
    - **Panel consensus**: Complexity is still a major hurdle. While vendors are working to abstract it, moving beyond simple "lift-and-shift" to more advanced "transparent" models requires platform and security engineering expertise.

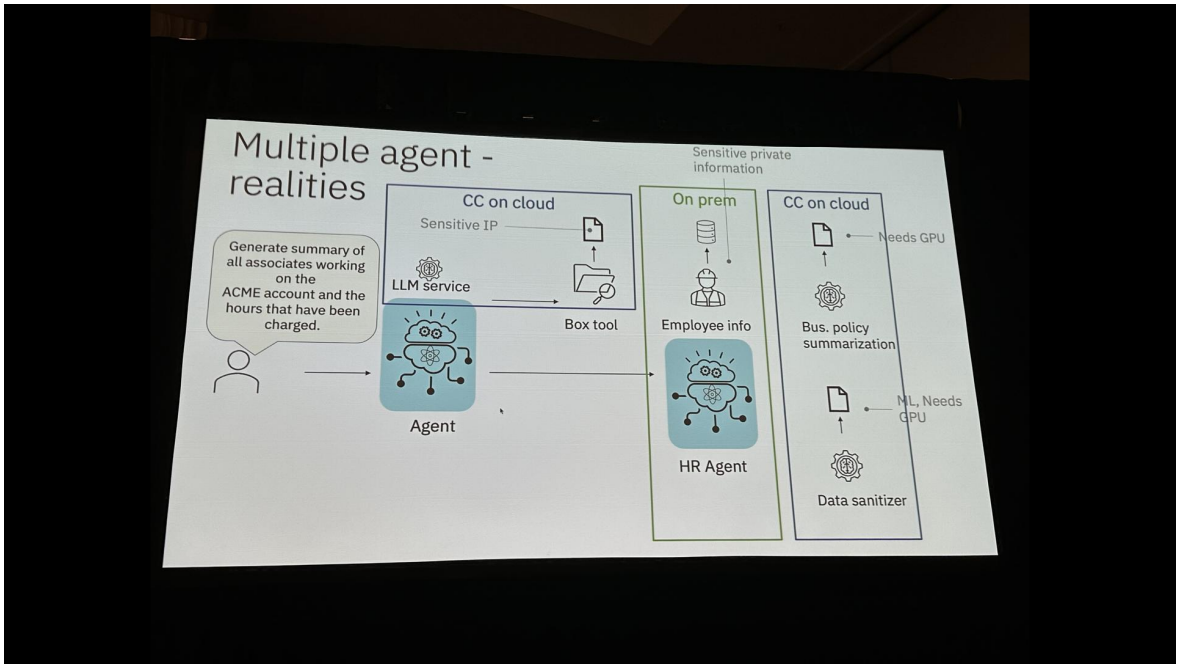
Lunch chat w/ Mark Bower from Anjuna security

- Interesting learning: The chip on a modern credit card is a form of Trusted Execution Environment (TEE).
- When a user enters their PIN for authentication, the chip performs a cryptographic operation. The backend server only ever sees the encrypted result, never the raw PIN, ensuring the secret is protected even if the backend is compromised. It can also use data like transaction history for validation within its secure boundary.

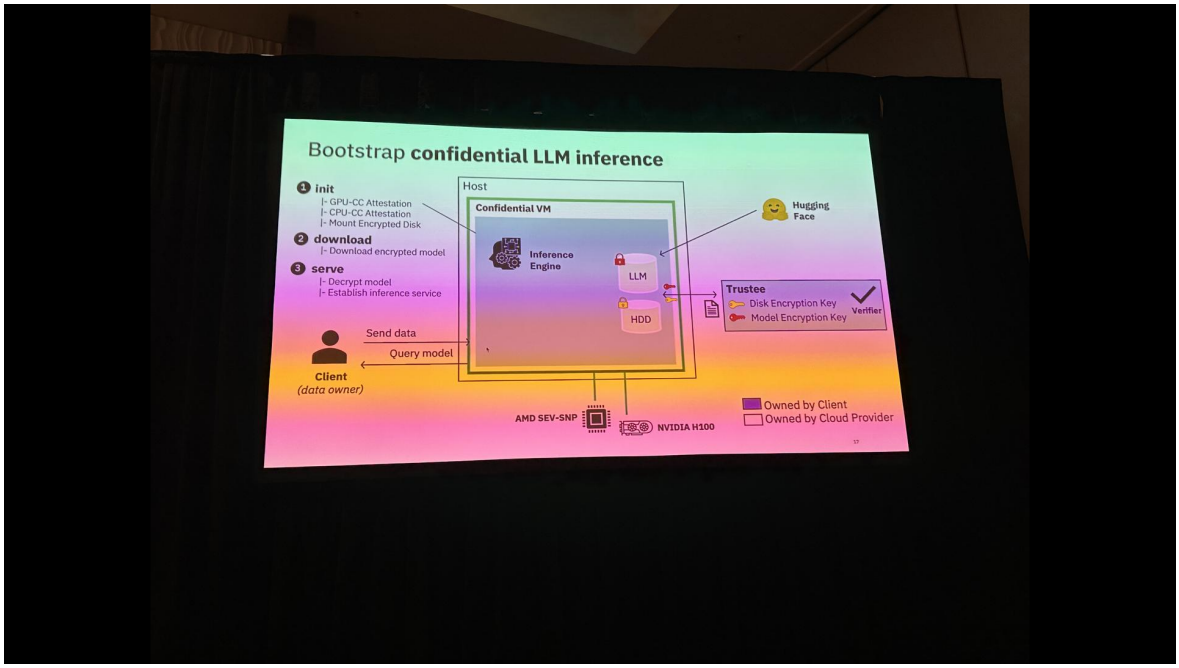
Breakout sessions

Title: Distributed confidential AI Agents

- **Speaker**: Michael Le, Julian Stephen from IBM
- **The Core Problem**: Agentic AI requires access to sensitive data, often from multiple parties who have mutual distrust.
  > The nature of the genetic system is that it is inherently owned by very different parties. You have data owner... model owner... infrastructure owner... and they are not very trusting with each other, right? You have this mutual distrust problem.
- An example workflow shows a request that requires an agent in a cloud CC environment to coordinate with an on-premise HR agent (handling sensitive employee info) and other cloud services for policy summarization and data sanitization.



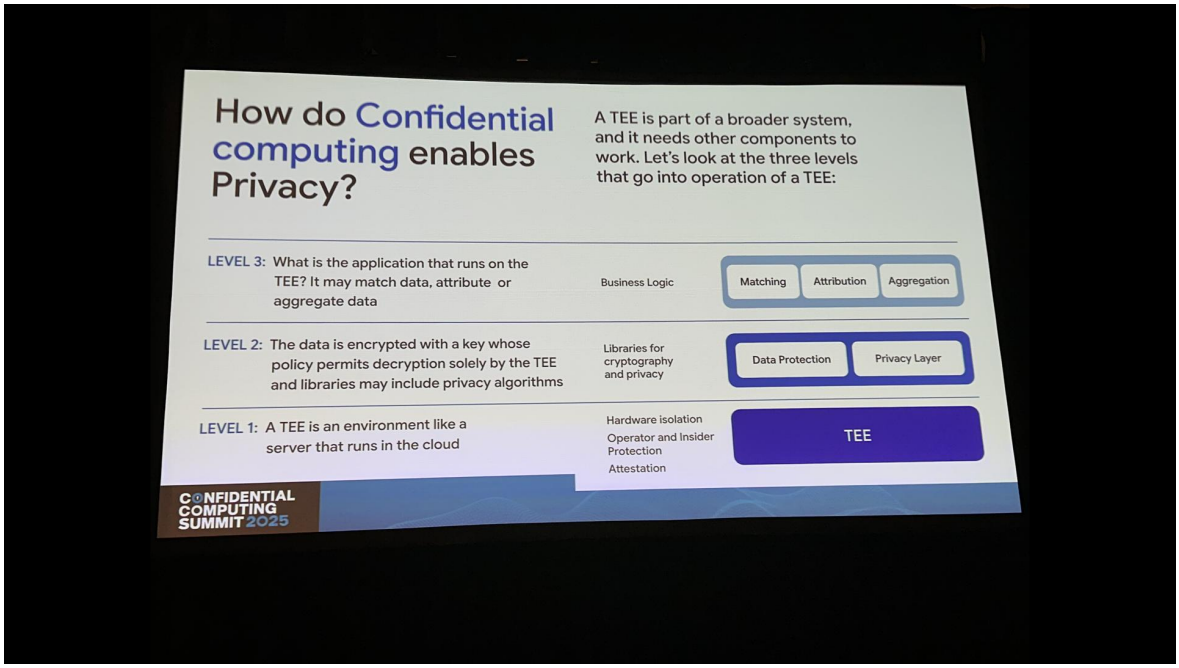
* **Challenges & Considerations for Deploying Confidential Agents**:
* **Attestation**: Confidential data (like disk encryption keys for volume mounts)
should only be moved to the VM *after* successful attestation.
* **Model Loading**: Model load times into a confidential VM are currently higher, which can impact service startup.
* **Performance Tradeoffs**:
* GPU-to-GPU interconnects (e.g., NVLink) are currently not encrypted within a multi-GPU VM, which can affect the trust model.
* Deploying on Kubernetes requires CC-aware orchestration.
* **Multi-Node Attestation**: A cluster of CC nodes needs a way to perform cluster-side attestation to form a single trust domain.
* **Bootstrap Confidential LLM Inference Process**: A secure process for initializing a confidential LLM service.



1. **Init**: The VM performs CPU and GPU attestation. After verification, it mounts an encrypted disk using a key released by a trusted party.
2. **Download**: It downloads the encrypted model (e.g., from Hugging Face)
.
3. **Serve**: It decrypts the model using another key released by the trustee and establishes the inference service.

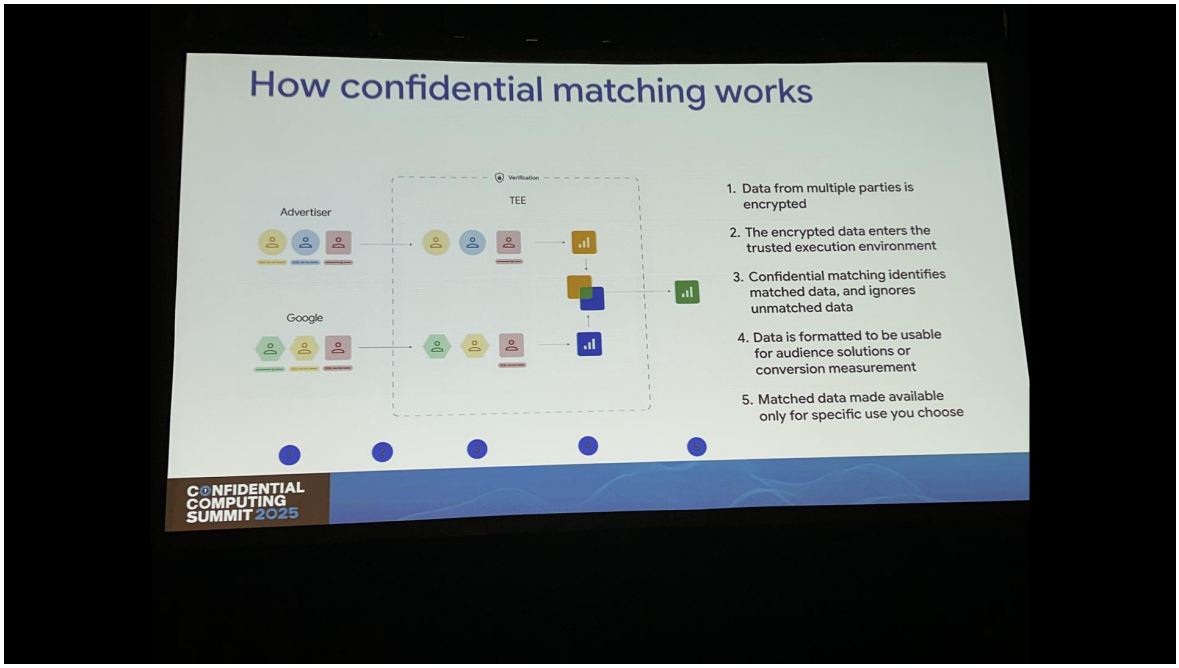
Title: Real-world application of TEEs for digital advertising at Scale

- **Speaker**: Chanda Patel from Google
- **Context**: Google Ads needs to personalize ads, which requires user data, but must do so in a way that is private and scalable for everyone from "Joe Plumber" to large retailers.
  > We want to make privacy so simple that they don't need to understand about confidential VMs or cloud. Things just work for them out of the box.
- **The 3-Level System**: A TEE is part of a broader system.

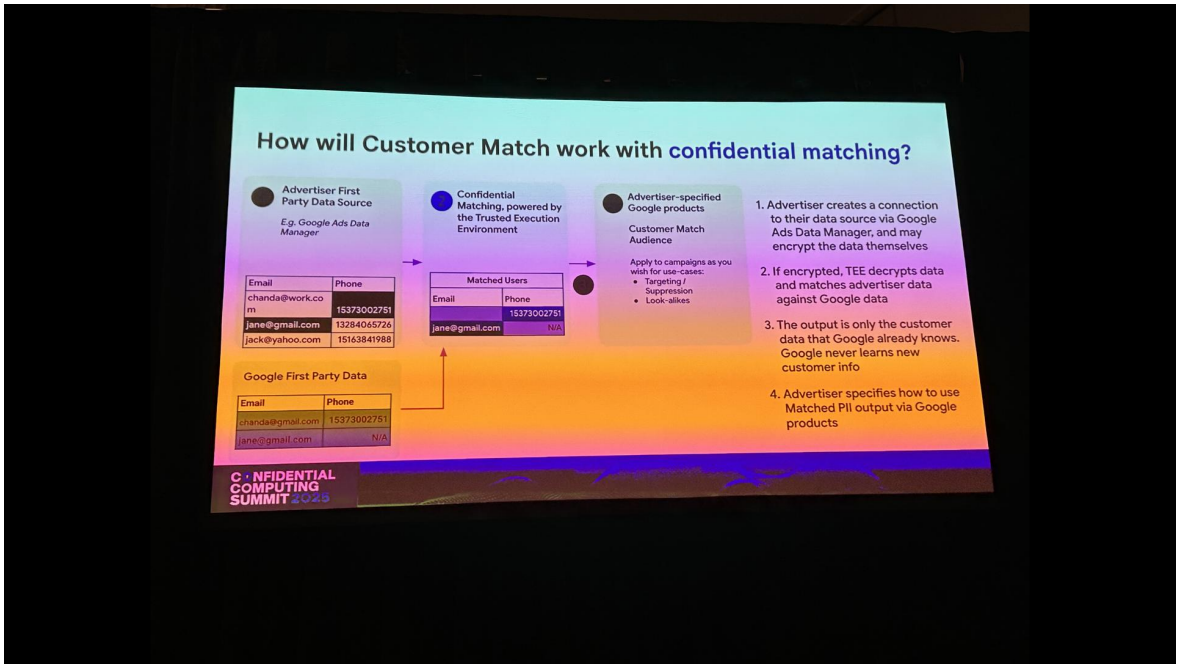


* **Level 1: The TEE**: The base infrastructure layer that provides hardware isolation, operator protection, and attestation.
* **Level 2: Libraries**: Data is encrypted with a key whose policy permits decryption *only* by the TEE. This layer contains open-source, auditable libraries for cryptography and privacy. The hash of the code is critical for trust.

> *If any line of code changes, that means the hash of that binary changed. And that can break the whole process, meaning that the key decryption will fail...*
> ** **Level 3: Business Logic**: The specific application running in the TEE, such as matching, attribution, or aggregation.*
> ** **How Confidential Matching Works**:*
> *1. Data from multiple parties (e.g., an advertiser and Google) is encrypted.*
> *2. The encrypted data enters the TEE.*
> *3. The matching logic runs inside the TEE, identifying matches and ignoring non-matches.*
> *4. The matched data is formatted for use.*
> *5. Only the specifically matched data is made available for the approved use case.*



* **Customer Match Example**: An advertiser provides their list of customers. Google provides its list. The matching happens in the TEE. The output is *only the set of customers that Google already knew about*. Google never learns about the advertiser's customers that aren't on its own platform.



* **Google Tag Gateway**: Extends this protection to data collected from browsers. When a user interacts with a website, the Google tag

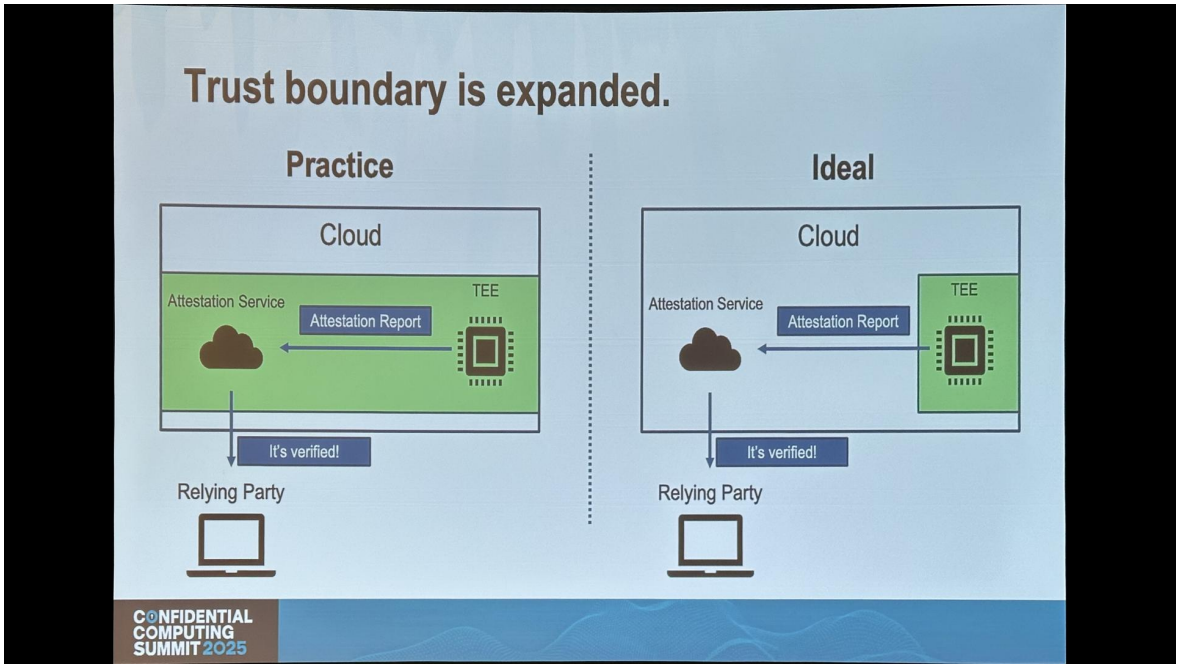
encrypts the event data in the browser before sending it to Google's servers, where it is processed only inside a TEE.

Title: PirateShip: append-only ledger with (mostly)

trusted execution environments
* **Speaker**: Shubham Mishra
* **Core Idea**: While TEEs raise the barrier for attacks by providing integrity protection, we shouldn't assume they will remain secure forever. Systems need to be designed to handle rare TEE compromises and correlated failures.
* **A Ledger should**:
1. Understand the reality of TEE compromises.
2. Offer flexible protections.
* **Proposed Solutions**:
* **Against Correlated Failures**: Use a heterogenous cluster of TEEs from different hardware vendors (e.g., Intel, AMD) to achieve platform fault tolerance.
* **Against Rare TEE Compromises**: Use auditing. The PirateShip ledger is designed for this.

Title: Trustless Attestation Verification in Distributed Confidential Computing

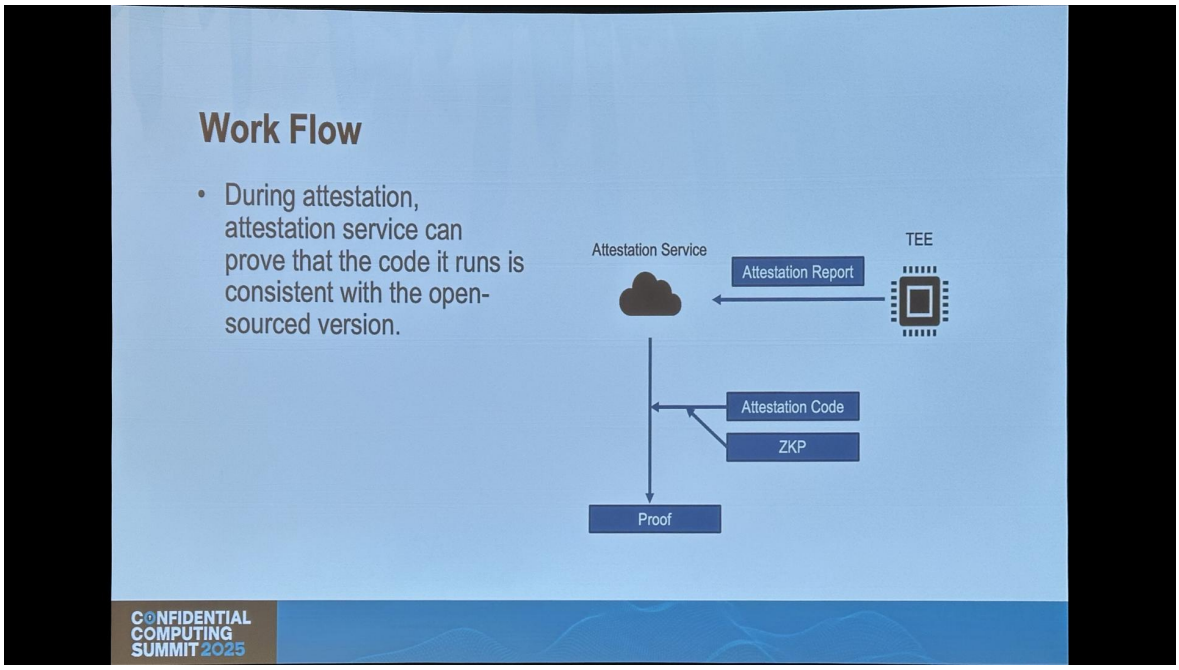
- **Speaker**: Donghang Lu from TikTok
- **The Problem**: When using a cloud service for remote attestation, the trust boundary expands to include the cloud provider's attestation service. The ideal scenario is to trust only the TEE hardware itself.



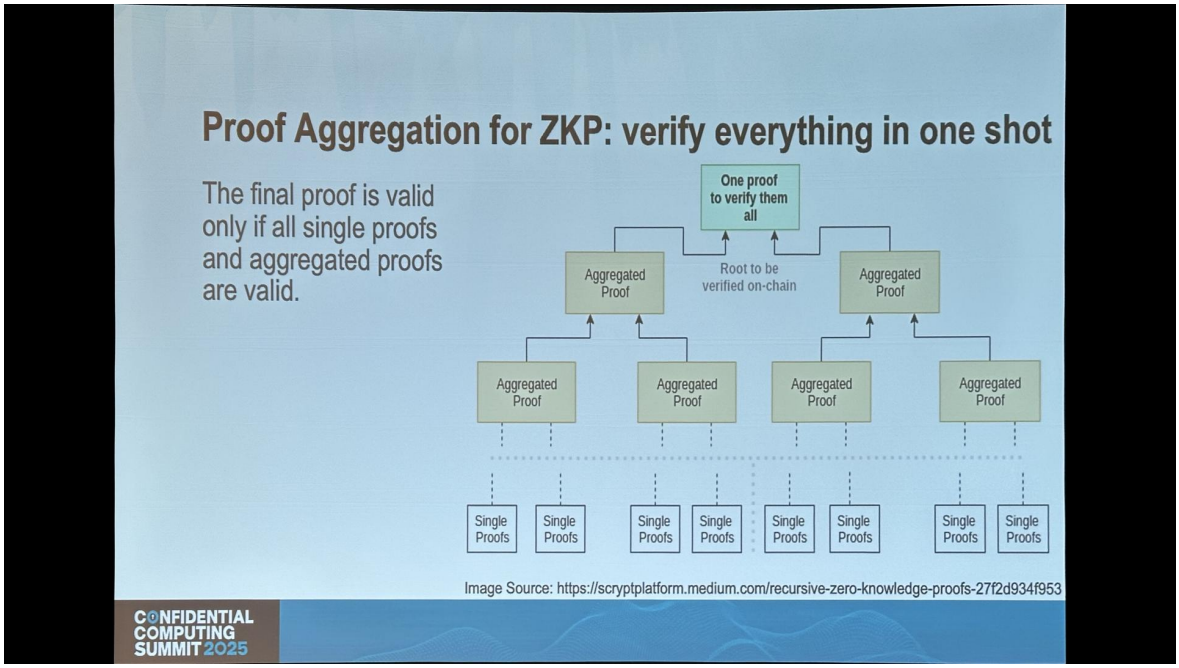
* **The Solution**: Ask the attestation service to prove itself. This can be done using Zero-Knowledge Proofs (ZKPs)

.

> *Instead of trusting attestation service, you can trust math.*
> ** **How it Works**: The attestation service code is open-sourced. When it verifies a TEE's attestation report, it also generates a ZKP proving that it executed the correct, open-source verification logic. The relying party can quickly verify this ZKP.*



* **Distributed Confidential Computing**: When multiple TEE nodes work together, they require mutual attestation. Verifying each one individually is inefficient.
* **Proof Aggregation for ZKP**: This technique allows multiple ZKP proofs to be aggregated into a single, final proof that can be verified in one shot. A load balancer could generate a proof for each TEE and then an aggregated proof for the whole cluster.

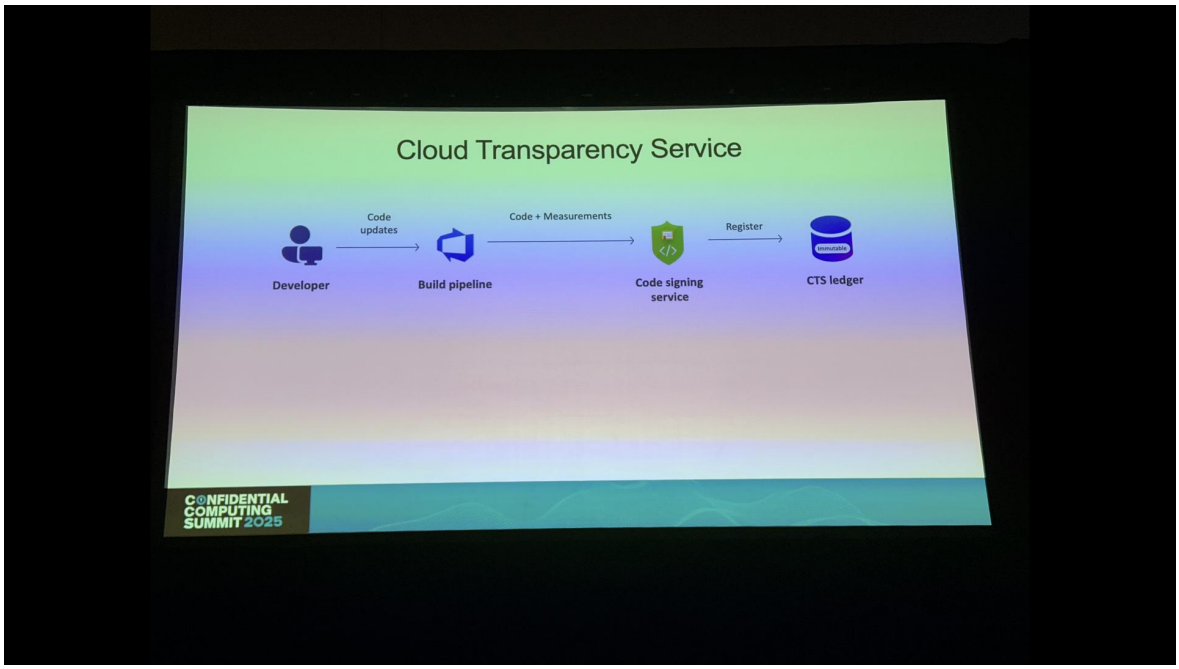


* **Performance of ZKP**: Performance depends heavily on the chosen protocols and frameworks. The speaker mentioned prototyping with two open-source frameworks: **Circom** (circuit-based)
and **RiscZero** (ZK Virtual Machine).

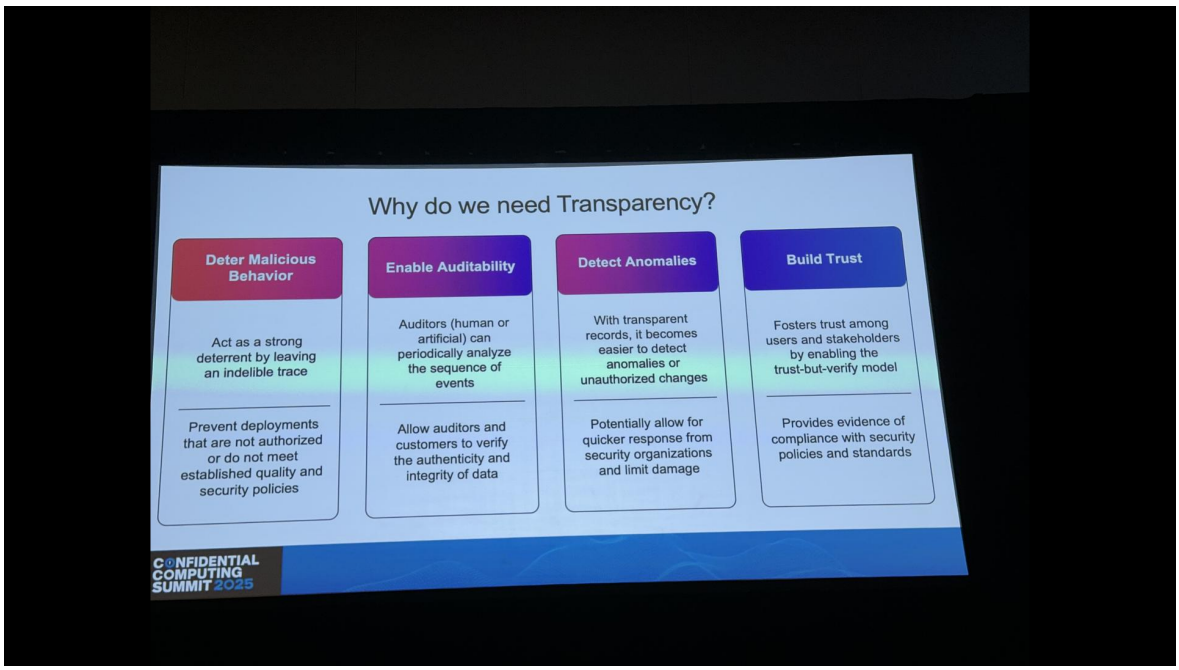
Title: Achieving auditability and trust through cloud transparency service on Azure

- **Speaker**: Kartik Prabhu from Azure
- **Cloud Transparency Service (CTS)**: A confidential and attested Azure service that provides a tamper-proof, append-only ledger for software releases.
- **Workflow**: When a developer updates code, the build pipeline registers the code's measurements and other metadata (e.g., Git commit hash) with the CTS ledger. CTS enforces policies (e.g., identity verification) before adding the entry and issuing a

cryptographic receipt. This receipt can then be used to gate deployments, ensuring only authorized and compliant code runs in production.

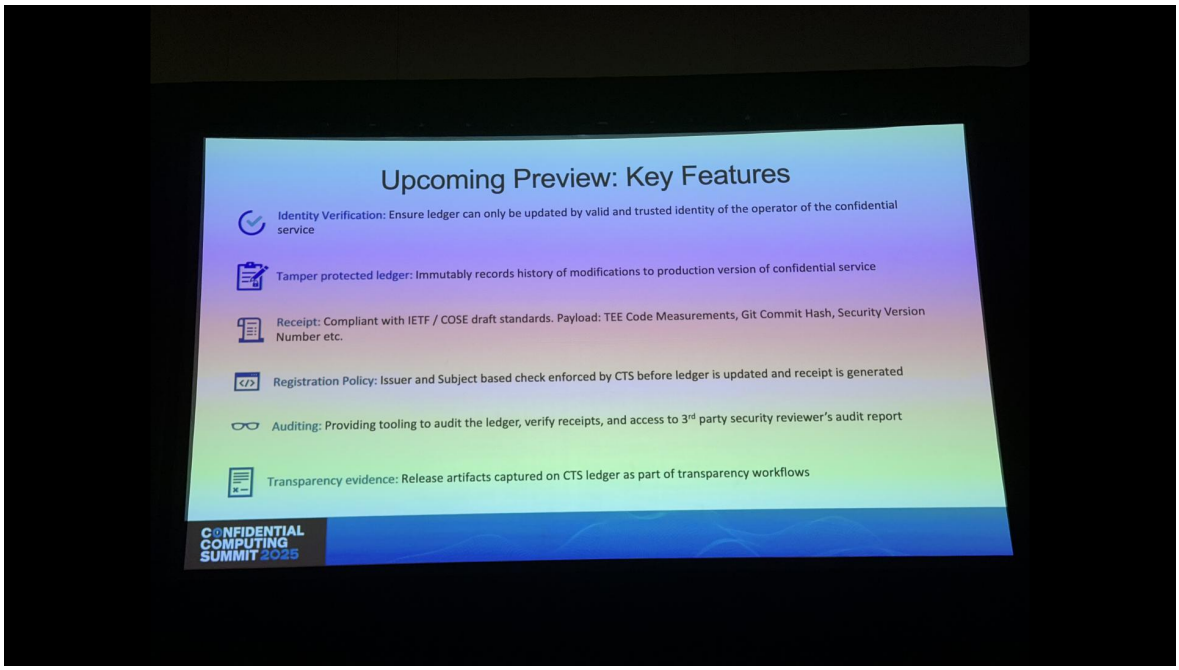


* **Why is this needed?**:



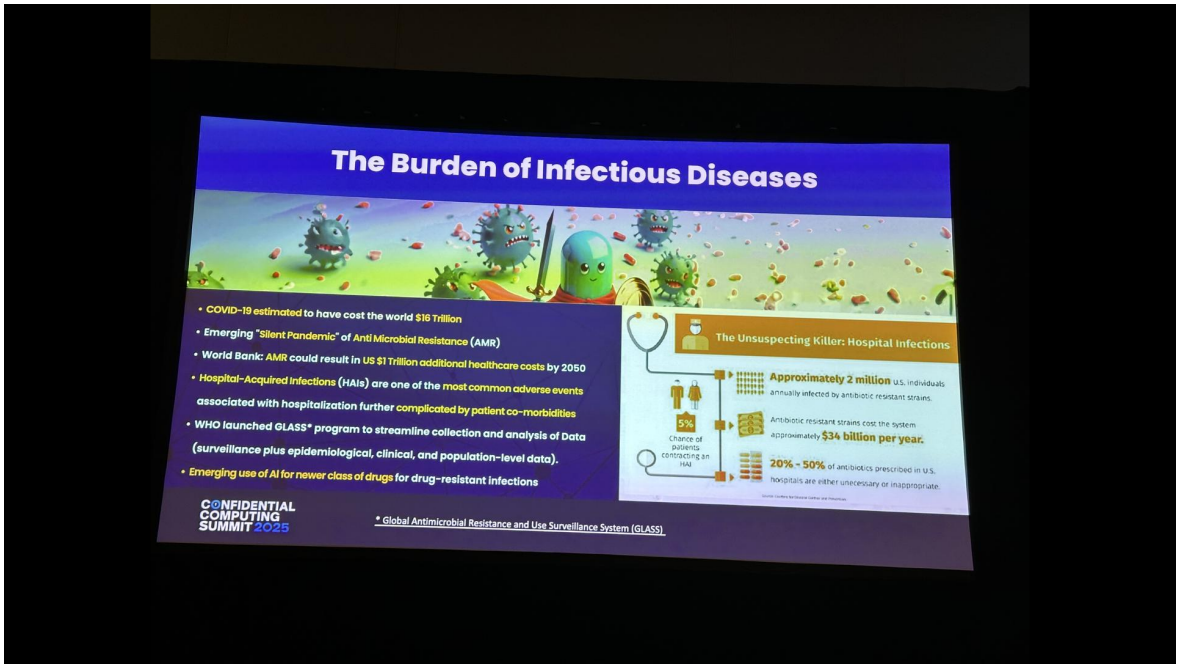
Deter Malicious Behavior	Enable Auditability	Detect Anomalies	Build Trust

| Leaves an indelible trace of any changes, deterring attackers. | Allows auditors to periodically analyze the history of events. | Makes it easier to detect unauthorized changes or anomalies. | Fosters a "trust-but-verify" model for users and stakeholders. |
* **Trust in CTS**: The service itself is built on confidential computing, is open-source, uses reproducible builds, is audited by third parties, and all its own upgrades are transparently recorded on the ledger.
* **Upcoming Preview**: A preview is coming for CTS integration with Microsoft Azure Attestation (MAA)
and Managed HSM. Key features will include identity verification, a tamper-protected ledger, COSE-compliant receipts, policy enforcement, and auditing tools.

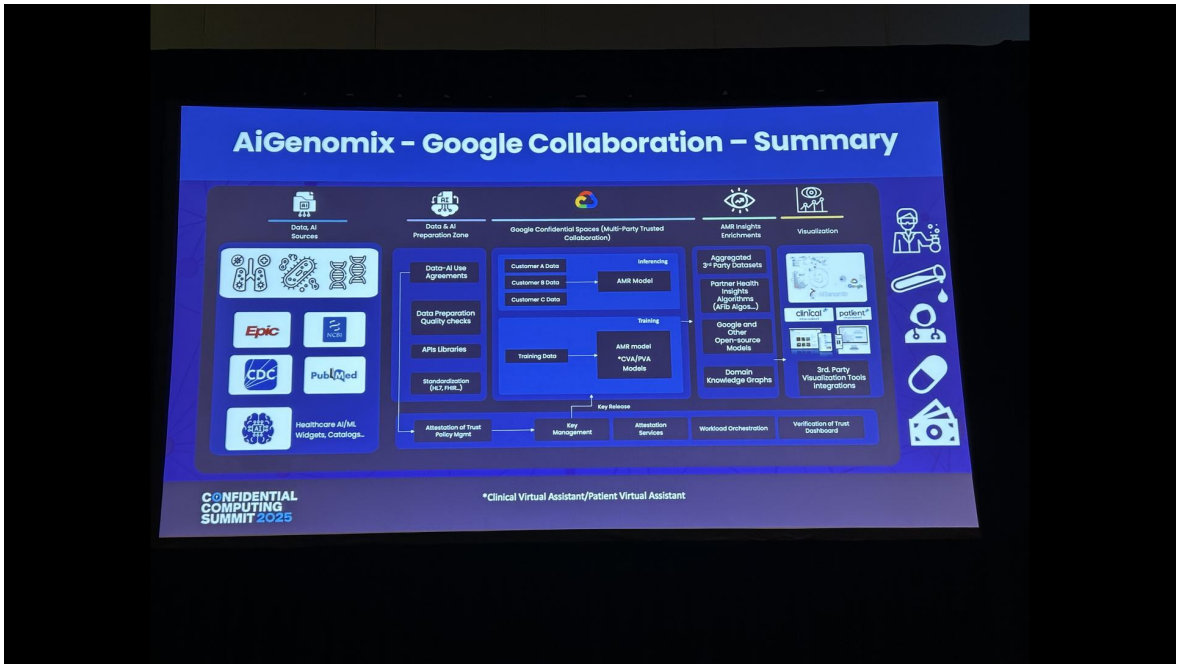


Title: Trusted insights for healthcare

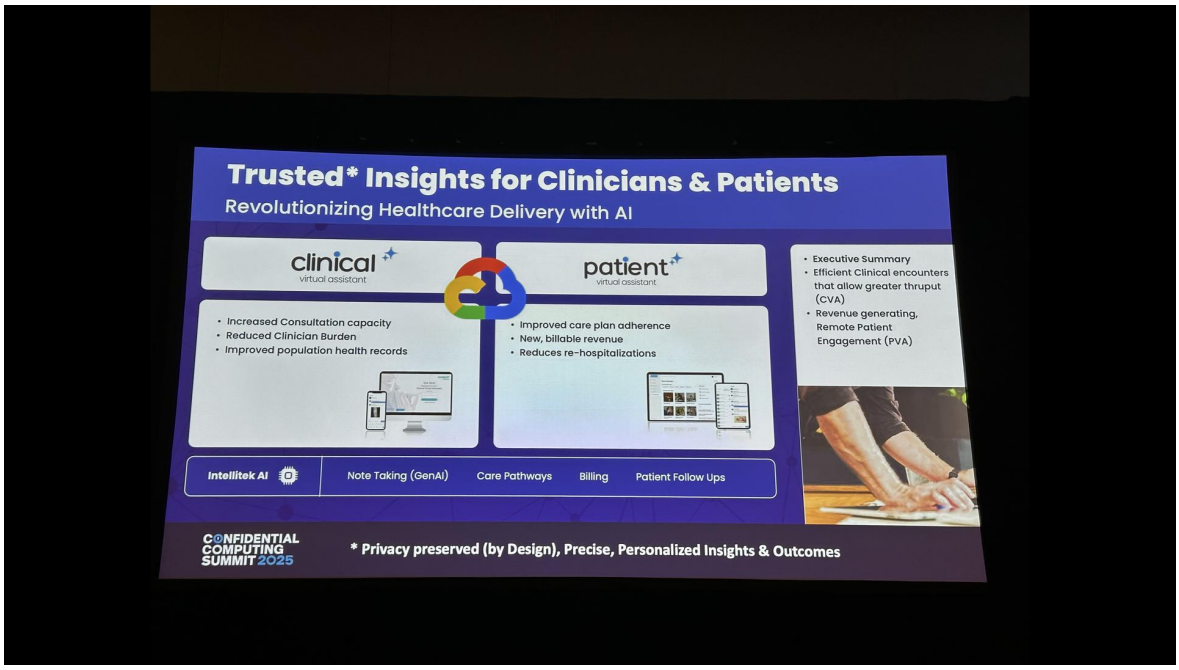
- **Speaker**: Jonathan Monk, Harsh Sharma from AiGenomix
- **The Problem**: The burden of infectious diseases, particularly the "silent pandemic" of Anti-Microbial Resistance (AMR)
and Hospital-Acquired Infections (HAIs). Analyzing sensitive patient and genomic data is key to fighting this, but privacy is paramount.



* **Solution**: AiGenomix is collaborating with Google Cloud to build a trusted AI platform for healthcare. The architecture uses Google Confidential Spaces for multi-party trusted collaboration.
* Multiple data providers (hospitals, research centers) can contribute data for AI model training and inferencing within a secure environment.
* The platform is underpinned by a foundational layer for attestation, key management, and verification, ensuring that data use agreements are programmatically enforced.



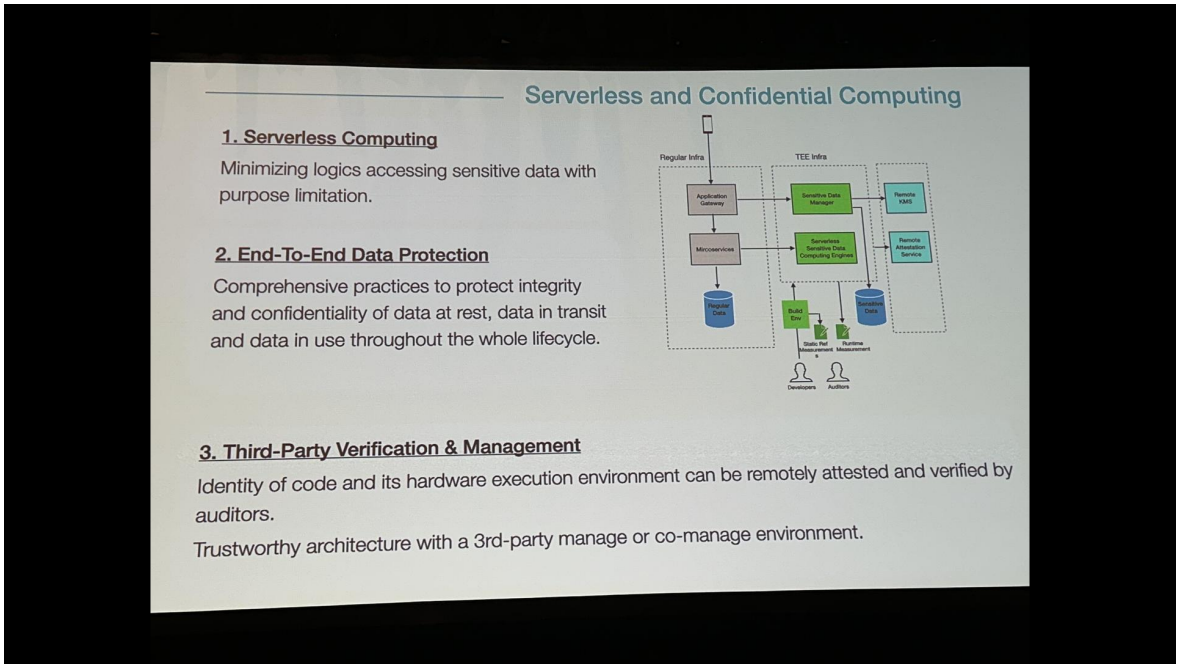
* **Outcome**: The platform generates trusted, privacy-preserved insights for clinicians (via a Clinical Virtual Assistant)
and patients (via a Patient Virtual Assistant).



Title: Safeguarding sensitive data access at scale with confidential computing

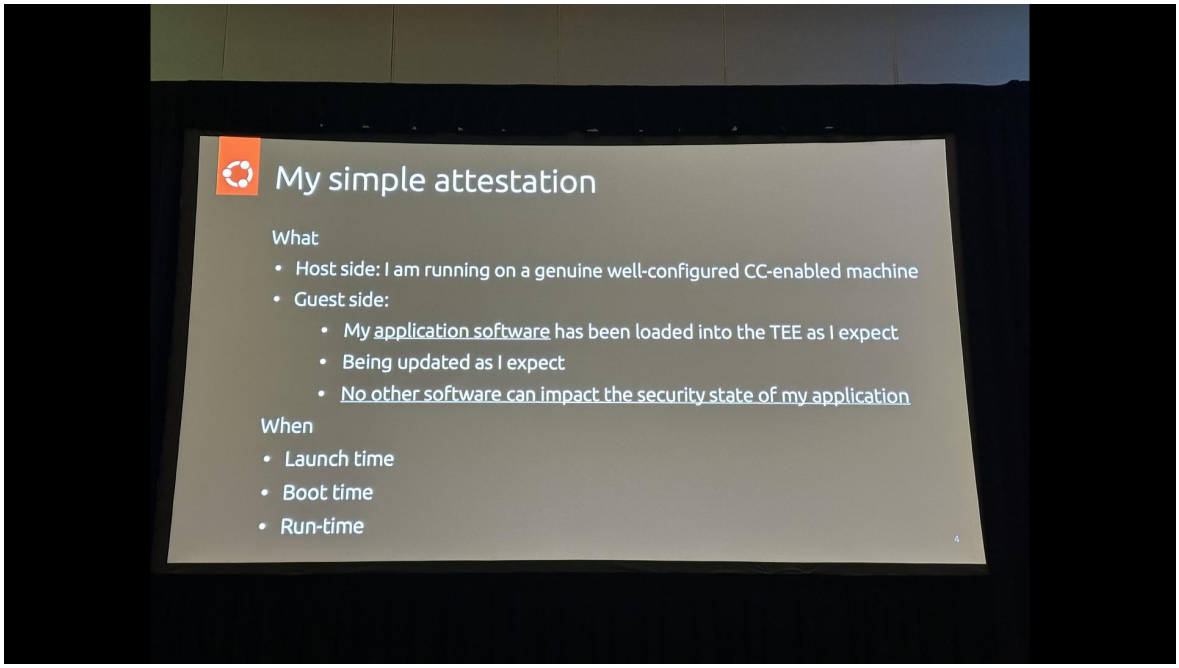
- **Speaker**: Yizuo Tian and Mingshen Sun from TikTok
- **Challenge**: How to protect data with fine-grained purpose limitation in a flexible microservice architecture, while also providing verifiable transparency to third parties.
- **Solution**: A framework combining serverless and confidential computing.
    1. **Serverless Computing**: Isolate the minimal logic that accesses sensitive data into serverless functions that run in a TEE. This narrows the scope of what needs to be reviewed and audited.
    2. **End-to-End Data Protection**: Data remains in a desensitized or encrypted form throughout its lifecycle, except for the moments it is being processed inside the TEE.
    3. **Third-Party Verification**: Use remote attestation and a third-party Key Management Service (KMS)

. The service provider technically has no access to the keys needed to decrypt sensitive data outside the attested TEE.

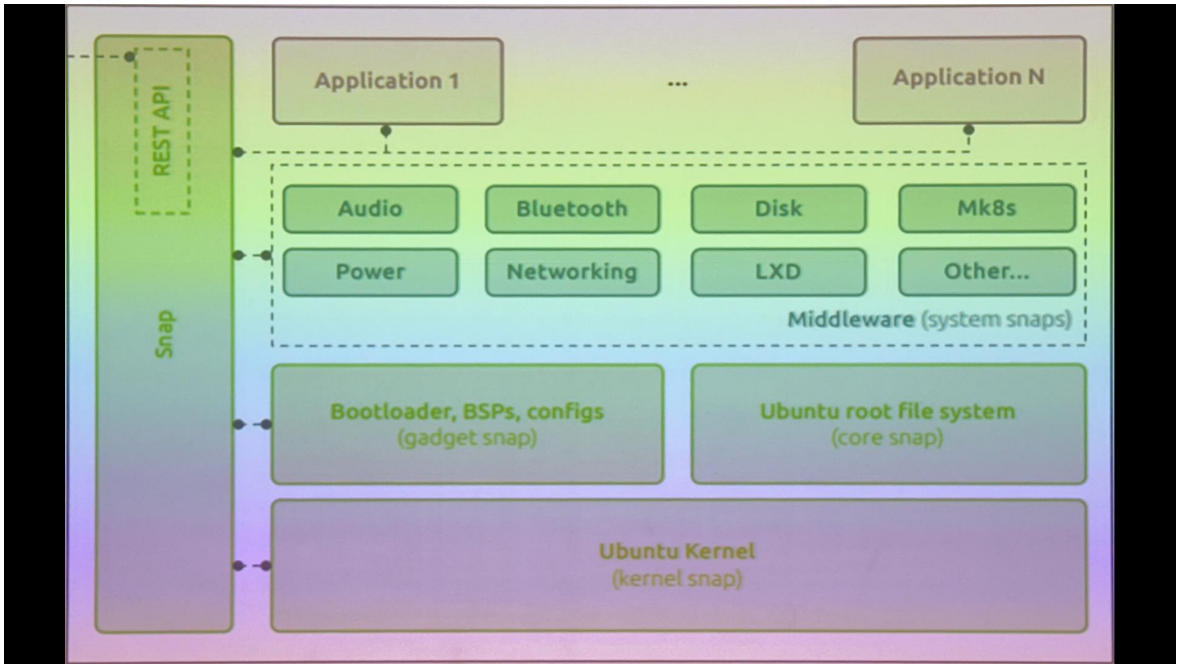


Title: Ubuntu Core: unlock ISV innovation in multi-cloud environments

- **Speaker**: Hugo Huang, Ijlal Loutfi
- **The ISV Dilemma**: ISVs are early adopters of confidential computing but face a tension between security and usability. Process-based TEEs (like SGX)
  offer tight security but are complex. VM-based TEEs are easy to use ("lift and shift") but "pollute" the attestation with the entire OS, not just the ISV's app.



* **The Solution**: Use a modular OS to achieve separation of concerns *within* the VM.
* **Ubuntu Core**: A version of Ubuntu built on "snaps"—sandboxed, modular, read-only application packages.
* **Modularity**: The OS is composed of distinct snaps (kernel, core, gadget, applications)
. This provides clear boundaries.
* **Provenance & Integrity**: Each snap is digitally signed and has a hash. Snapd (the agent) verifies this hash on installation. A new feature uses `dm-verity` to provide runtime integrity, extending the chain of trust from the hardware root of trust all the way up to the application.

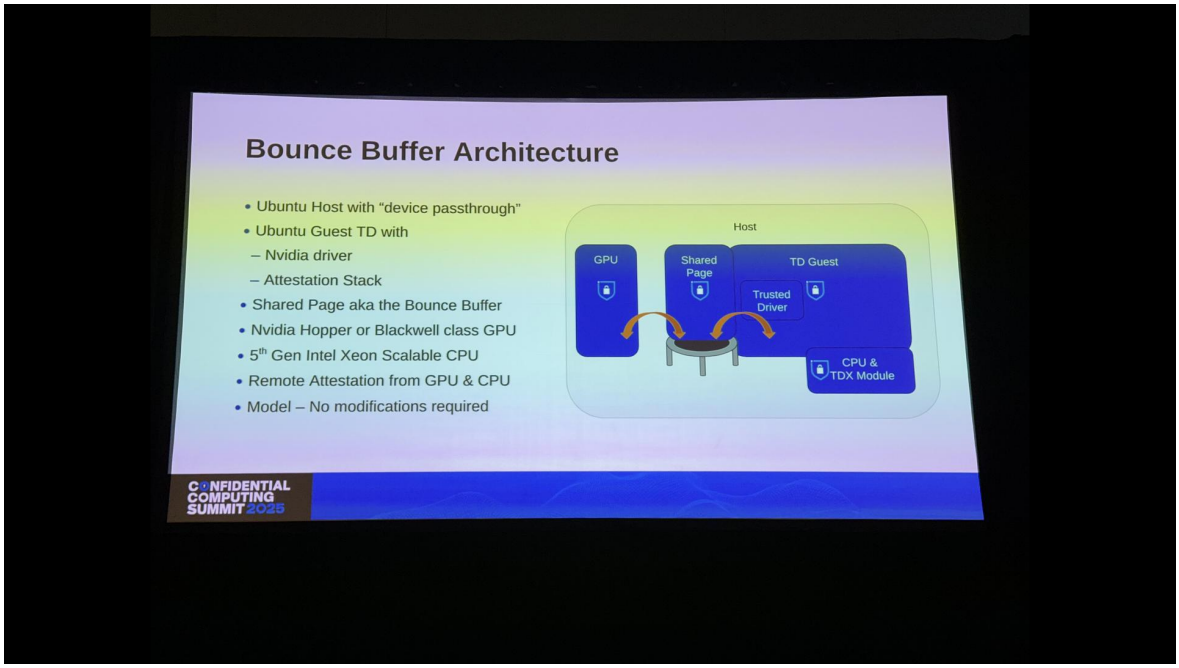


* **From Edge to Cloud**: Ubuntu Core's architecture, originally designed for IoT devices, is organically suited for CVMs because both are "appliances" that need to be secure, isolated, and transactionally updatable. Ubuntu Core will be available for Google Cloud CVMs in Q3 2025.

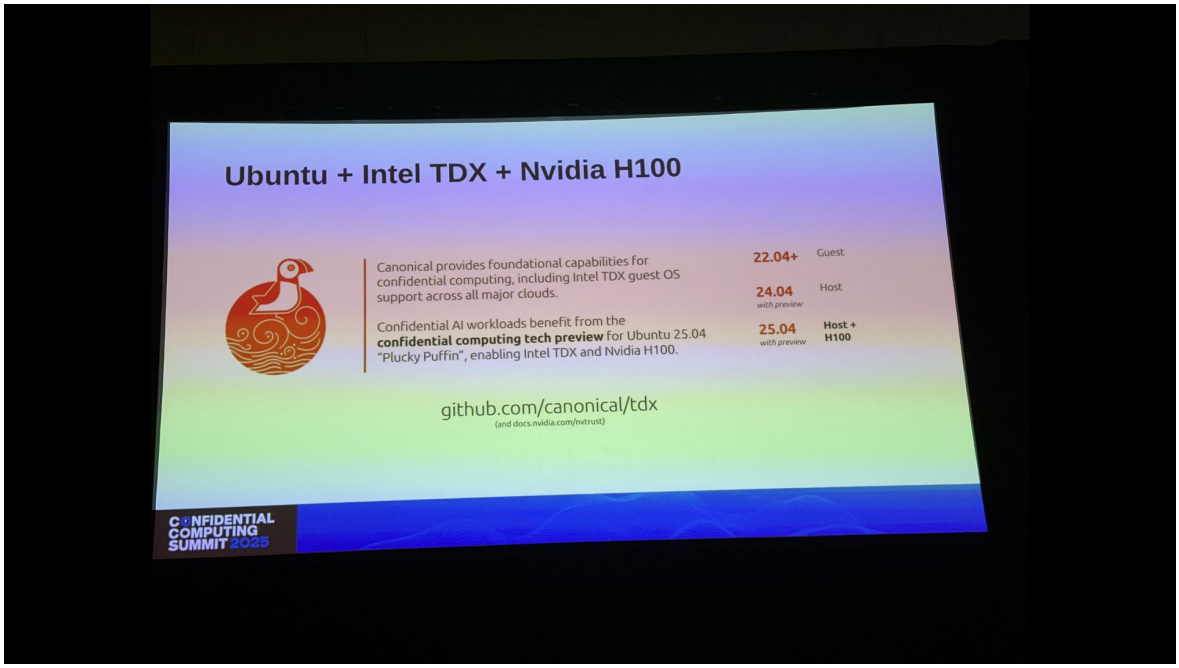
Title: Building a Confidential AI Hardware and Software Solution Stack

- **Speakers**: Dan Middleton from Intel, Chris Schnabel from Canonical, Rob Nertney from Nvidia
- **The Challenge**: How to make the CPU and GPU communicate securely when the communication path travels through the untrusted host OS.
- **Bounce Buffer Architecture**:
  - A secure channel is established between a trusted driver in the confidential VM (TD Guest)

    and the GPU.
  - A small, shared memory region (the "bounce buffer" or "shared page") is created that both the guest and the host can access.
  - Data from the VM is encrypted by the CPU, "bounced" through this shared page to the GPU, which then decrypts it for processing. The process is reversed for the response.



* **Solution Stack**: **Ubuntu + Intel TDX + Nvidia H100**. Canonical provides the OS support (Ubuntu 25.04 now has a tech preview for Host + H100 support)
, Intel provides the CPU with TDX, and Nvidia provides the H100 GPU with a confidential mode.
* **Code**: `github.com/canonical/tdx`

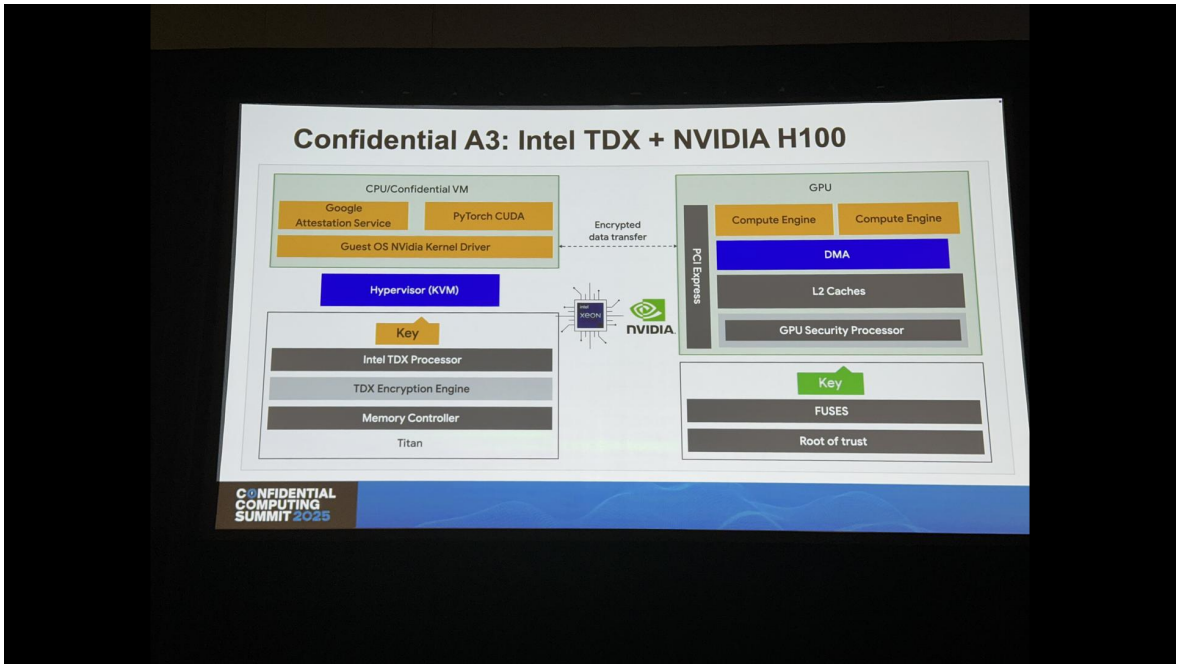


* **The Future: Getting Rid of the Bounce Buffer**: The bounce buffer

has a performance cost. The next generation will use "TDX Connect" (based on the TDISP standard in PCIe)
, which allows the CPU and GPU to communicate securely at full line rate without software encryption, effectively making confidential computing "just work" like the transition from HTTP to HTTPS.

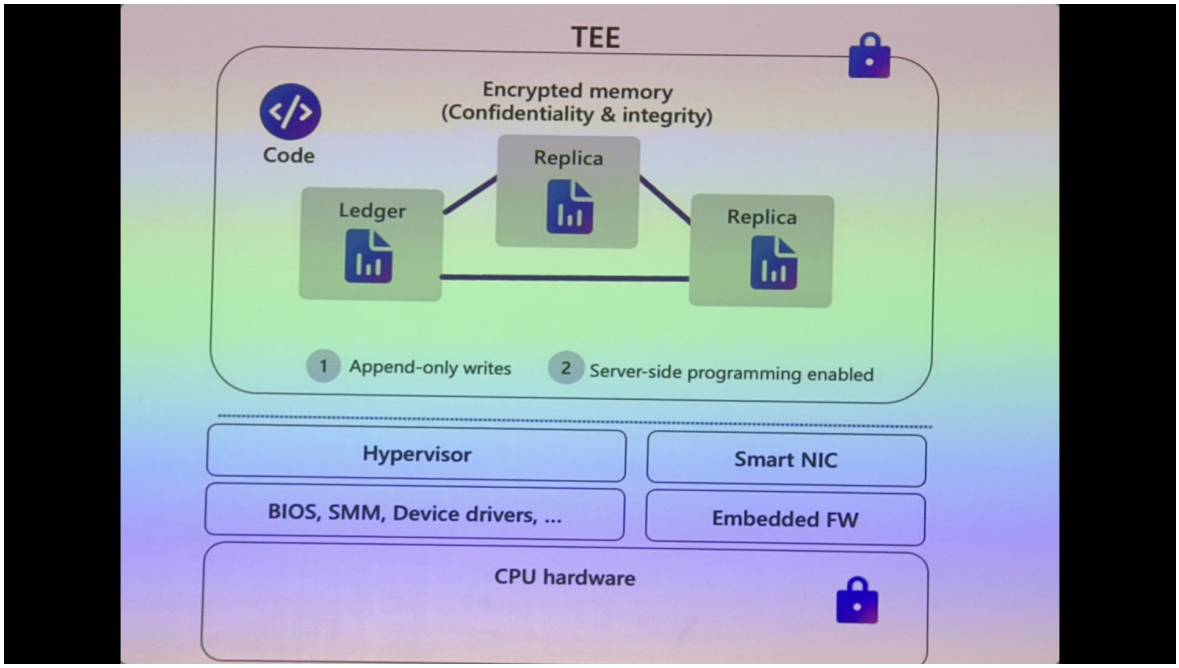
Title: Confidential AI: securing the future of intelligence in a privacy-first world

- **Speakers**: Ranjit Narjala, Amit Patil from Google Cloud
- **The Goal**: Protect both the data and the AI models throughout their lifecycle (production, tuning, consumption).
- **Confidentiality Spectrum**: From most secure to most usable: Enclaves -> Confidential VMs -> Attestation & Transparency. The right choice depends on the application's needs.
- **Google's Confidential Cloud Portfolio**:
    1. **CC Infrastructure**: Confidential VMs (AMD SEV & Intel TDX), Confidential GPUs (H100), Attestation Service, Confidential Hyperdisk.
    2. **Confidential GCP Services**: Confidential GKE nodes, Dataproc, Dataflow (Beam).
    3. **Confidential Applications**: Confidential Spaces for multi-party collaboration and Confidential AI.
- **Confidential A3**: The platform combining Intel TDX CPUs with Nvidia H100 GPUs, available in public preview. The architecture ensures encrypted data transfer between the CPU's confidential VM and the GPU over the PCIe bus.

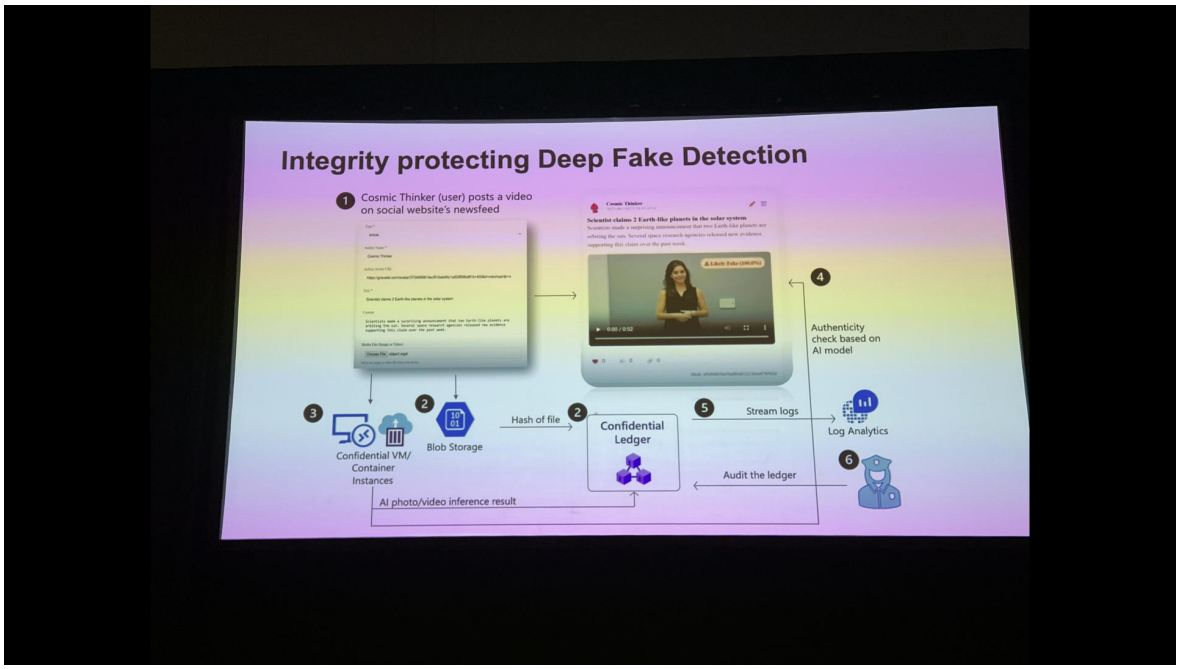


Title: Securing data and records through Azure confidential ledger

- **Speakers**: Shubhra Sinha Kamath and Yagnesh Setti from Azure
- **Azure Confidential Ledger**: A service providing tamper-proof, append-only, verifiable audit trails, built on confidential computing (CCF)
  . It runs in a TEE, ensuring data integrity and confidentiality.



* **Case Studies**:
* **Integrity-Protected Deepfake Detection**: A video is posted, a hash is stored in the ledger. A confidential AI model analyzes the video for deepfakes, and its result is also recorded in the ledger. This creates an immutable, auditable trail of the content and its verification status.

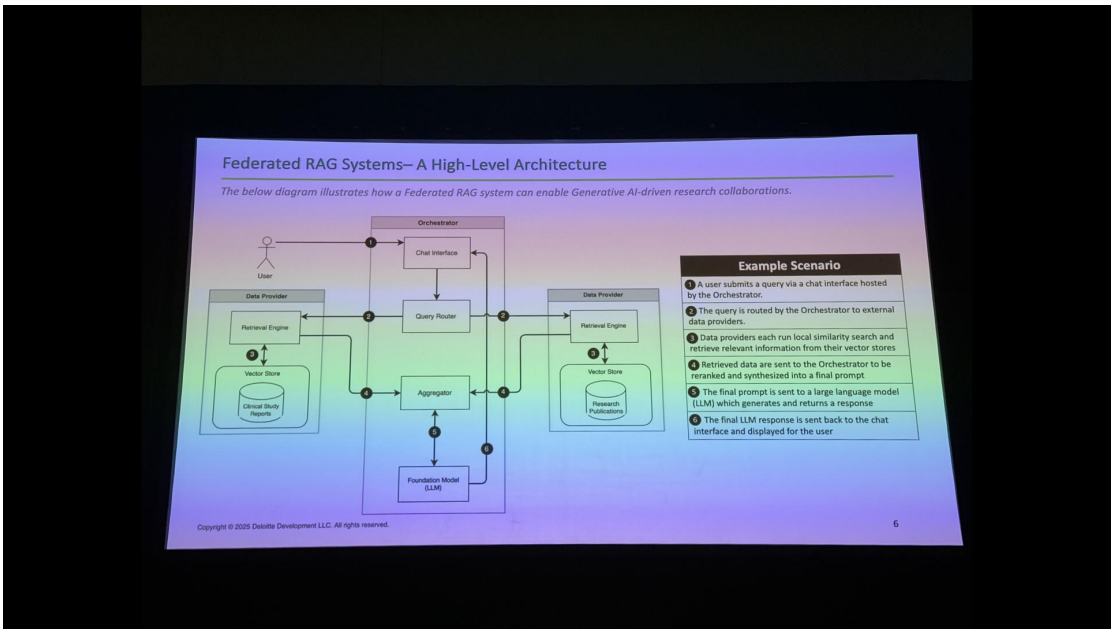


* **AI Dataset Records**: BeeKeeperAI uses the ledger to record algorithmic and dataset relationship logs from their AI computations, providing a verifiable record for regulatory bodies.
* **Carbon Measurement**: Carbon Asset Solutions uses the ledger to provide integrity for its carbon measurement platform, recording IoT data from farms.
* **SOX Compliance**: Microsoft's internal compliance team uses the ledger to manage evidence for SOX compliance, creating a reliable source of truth for actions and data existence.

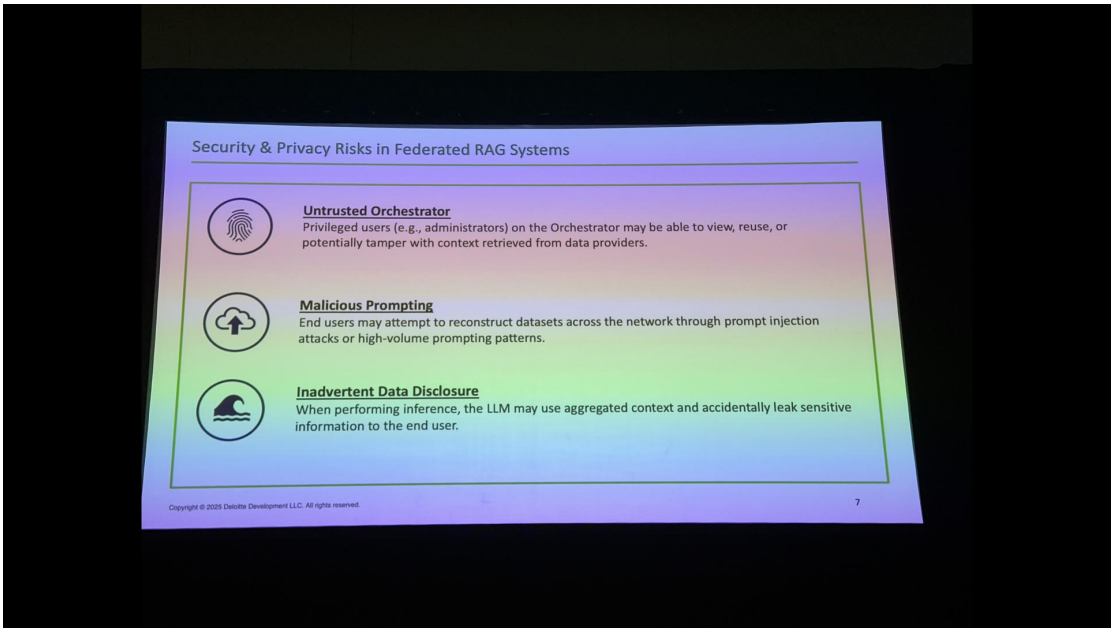
Title: Confidential Federated RAG systems: Enabling secure and robust clinical applications

- **Speaker**: Minh-Tuan Nguyen from Deloitte

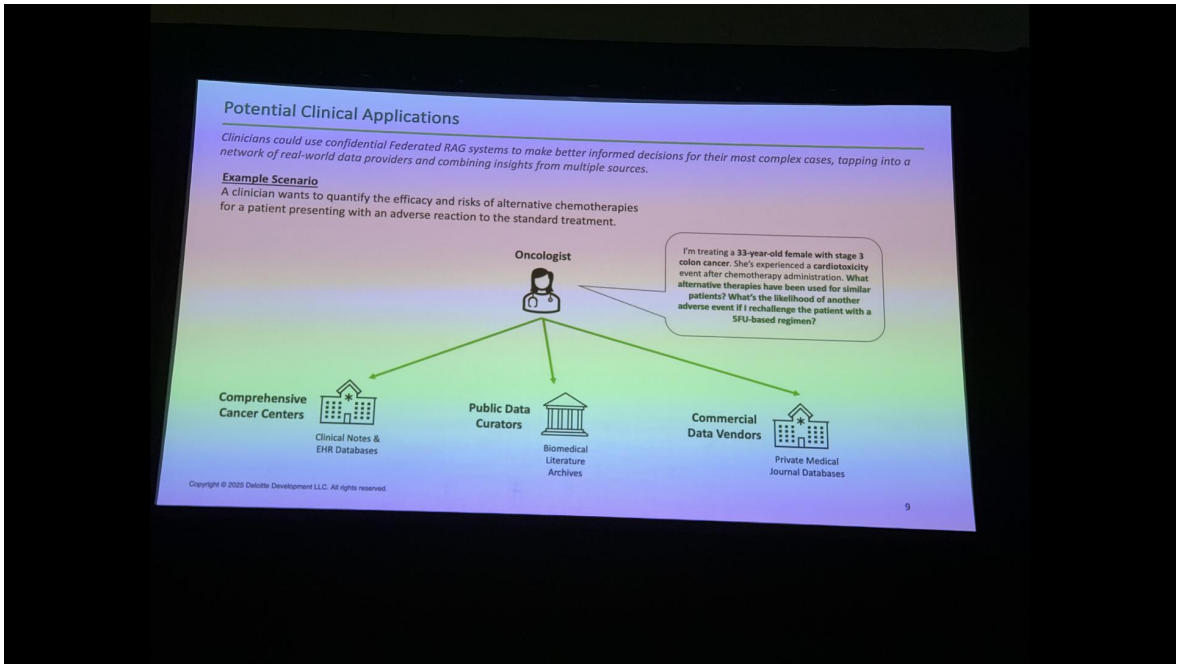
- **Federated RAG System**: An architecture to allow RAG systems to securely retrieve and synthesize information from a network of distributed data providers without centralizing the raw data.
    - **Components**: Federated embedding models, a federated data catalog for discovery, and a federated retrieval mechanism where an orchestrator routes prompts to the relevant data providers.
- **High-level Architecture**: A user queries a central orchestrator, which routes the query to multiple data providers. Each provider runs a local search on its own vector store and returns relevant context to the orchestrator, which aggregates it and uses an LLM to generate a final answer.



- **Security and Privacy Risks in Federated RAG**:



    - **Untrusted Orchestrator**: The orchestrator might view or tamper with the aggregated context.
    - **Malicious Prompting**: Users could try to reconstruct datasets via prompt injection.
    - **Inadvertent Data Disclosure**: The LLM might accidentally leak sensitive information from the aggregated context in its response.
- **Confidential Federated RAG (C-FedRAG)*
  *: The solution is to place the aggregator and the LLM inside a confidential computing environment. Data providers send their encrypted context to this secure environment, which is the only place it is decrypted, processed, and used for generation.
- **Example Scenario**: An oncologist treating a patient with an adverse reaction to chemotherapy can query a network of sources (cancer centers, public data curators, commercial vendors) to find alternative therapies used for similar patients, without any of the data providers having to share their raw clinical data.



* **Reference**: Paper on the topic: *C-FedRAG: a confidential federated retrieval-augmented generation system*, arXiv:2412.13163.