

Action Items & Further Reading

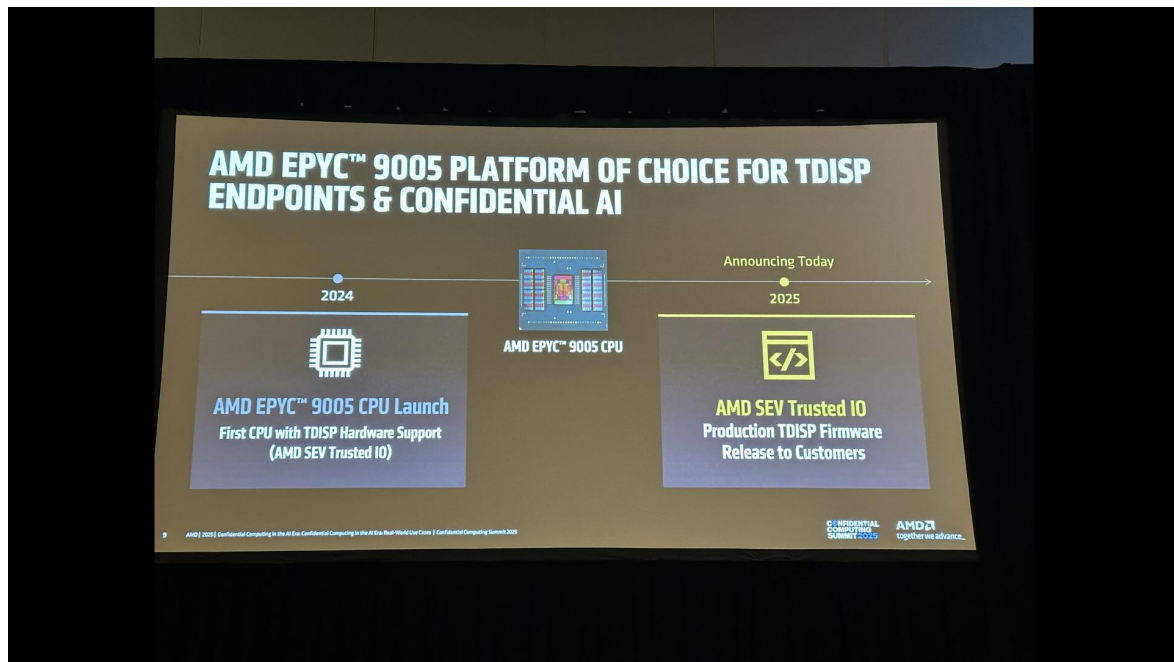
- [] Read the Meta whitepaper 'Private Processing for WhatsApp Overview'

Conference Notes: Confidential Computing Summit 2025

Title: Confidential computing in the AI era: Real world use cases

Speaker: Ravi Kuppuswamy (AMD)

- **Confidential computing enables privacy for AI:** The core promise is that not even cloud providers can see the data being analyzed. This is crucial for unlocking the full potential of AI.
 - > "Because if we are not confident about how that data is getting handled, then the use cases that one would actually be able to use and process is going to get limited."
- **Protecting Data, Models, and Weights:** Confidential Computing is critical at every phase of the AI lifecycle, including application development, model training, and inference.
 - > "Protecting the data, the models and weights, and particularly AI data, which is highly sensitive, user application, training and inference applications, you need the fundamentals of a robust framework for securing this data."
- **AMD EPYC 9005 CPU (Turin):** Launched in 2024, it was the world's first CPU with hardware support for TDISP (TEE Device Interface Security Protocol).
- **Today's Announcement:** AMD is releasing the production firmware for TDISP to customers, enabling AMD SEV Trusted IO.
 - > "In 2025 today, we're excited as AMD to announce that the firmware for this and the true application of this processor for SCV trusted IO will be released and available to customers in a few weeks."



- **Use Case: BeeKeeperAI (Healthcare)**

: EscrowAI enables confidential AI collaboration using AMD SEV-SNP. Another real-world example is AstraZeneca using Google Cloud for secure analysis of patient data to determine correct medications.

- **Use Case: WhatsApp (Meta):** Used for private conversations, including by governments, and increasingly for AI features like summarizing messages, drafting replies, etc.

- Meta's Private Processing platform ensures that not even Meta can see the content of user messages being processed for AI features.

- The platform uses AMD EPYC CPUs with SEV-SNP for the confidential VM and communicates with NVIDIA H100 Tensor Core GPUs via the SPDM (Security Protocol and Data Model) protocol. This secure, attested channel protects the entire workflow.

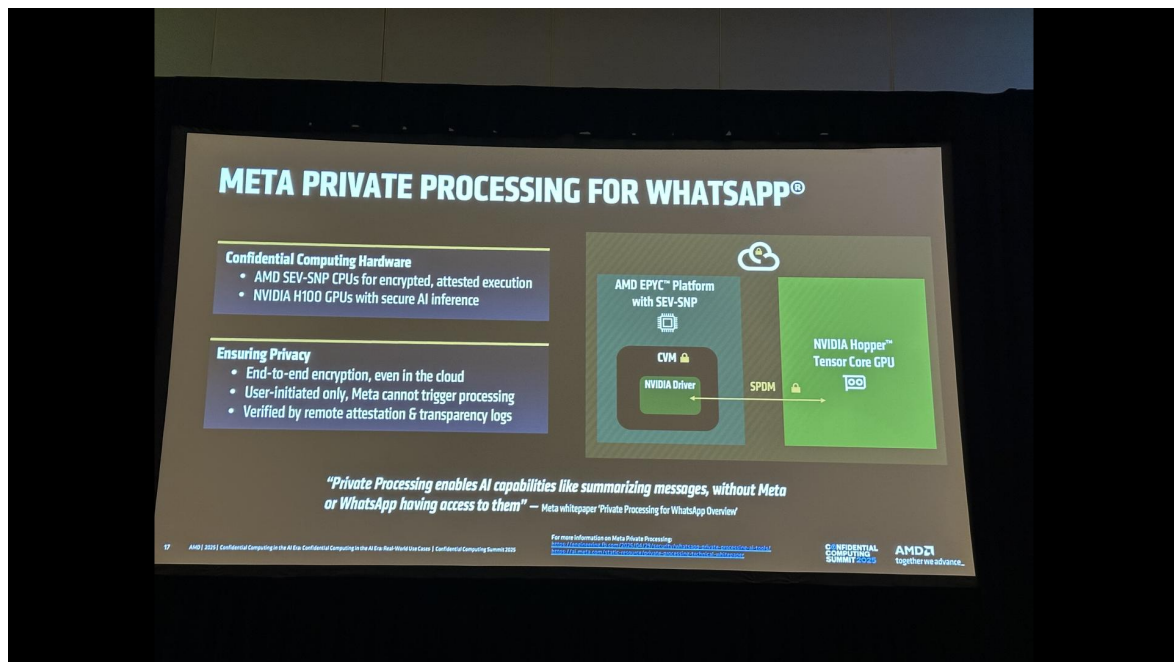
- This collaboration demonstrates an open ecosystem, not a "walled garden."

"Now, this is not a walled garden. People may want to choose and attach any kind of device that they want to go ahead and attach. You do not want the limits of security to be within the walls of any one company."

- **Further Reading:**

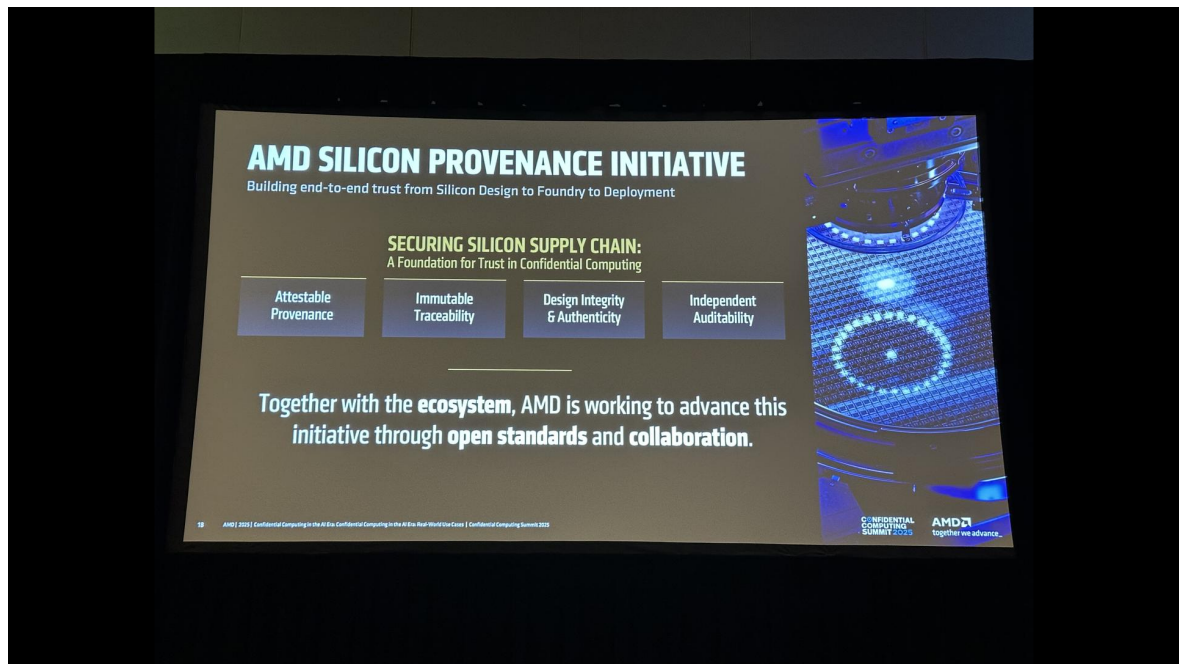
- <https://engineering.fb.com/2025/04/29/security/whatsapp-private-processing-ai-tools/>

- <https://ai.meta.com/static-resource/private-processing-technical-whitepaper>



- **AMD Silicon Provenance Initiative:** An effort to secure the entire silicon supply chain, from design to deployment, based on four key pillars:

1. **Attestable Provenance:** Verifiable records across the device lifecycle.
2. **Immutable Traceability:** Tamper-proof tracking.
3. **Design Integrity & Authenticity:** Ensuring the design is genuine.
4. **Independent Auditability:** Allowing third parties to verify claims.



- **Open Standards and Collaboration:** AMD emphasizes enabling the ecosystem through open standards and collaboration with partners and industry bodies.

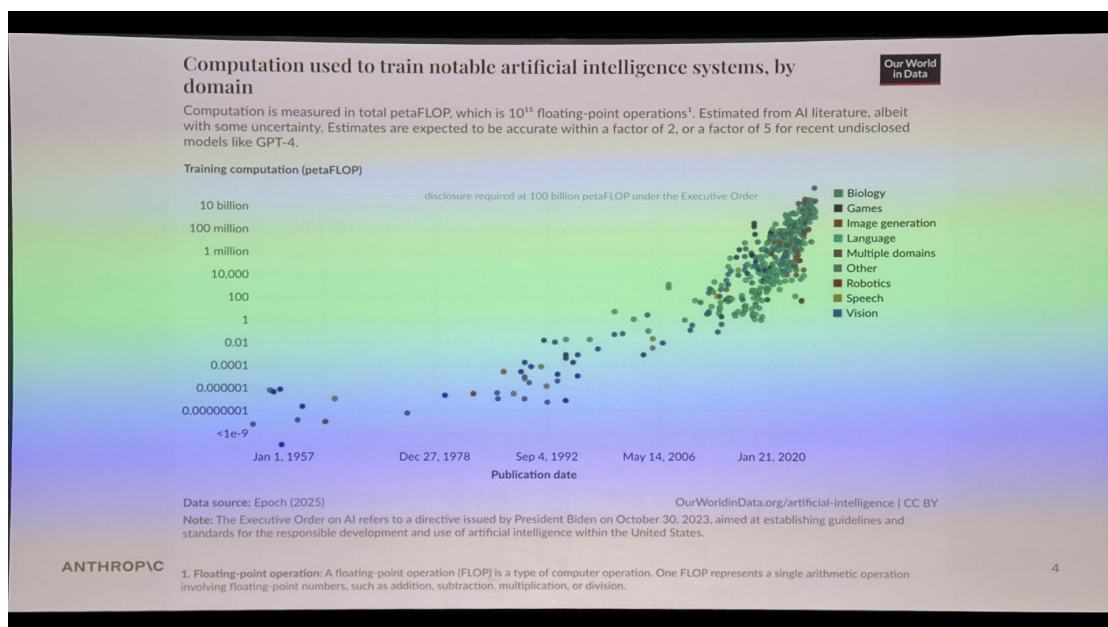
"We are calling upon all of you, the ecosystem here, to build confidential AI around open standards and a trusted supply chain."



Title: Update from the LLM scaling law frontier, leading intelligence increases and cybersecurity implications

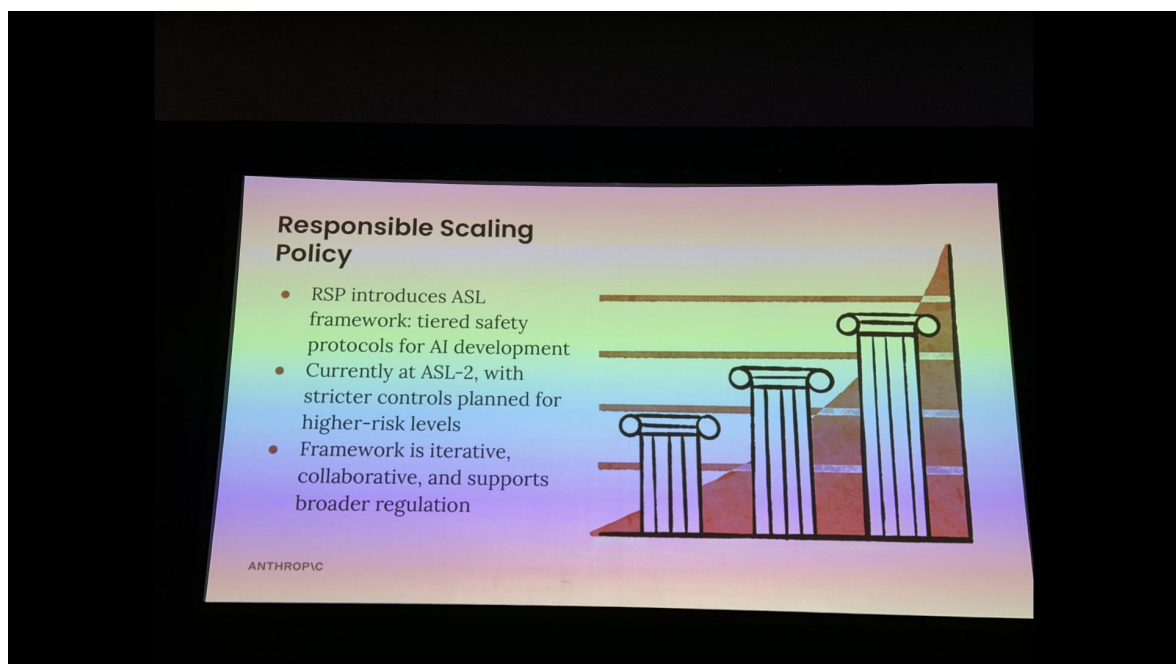
Speaker: Jason Clinton, CISO of Anthropic

- **Scaling Laws Continue:** The amount of computation used to train notable AI systems continues to grow exponentially, following a linear trend on a log scale.
> "This is the most important graph that everyone in this room should internalize... we see roughly a 4X year-over-year increase in the amount of compute that goes into AI models. And that line has been holding for almost 70 years now."
- **Future Projections:** This trend is expected to hold for at least the next few years, meaning we can anticipate models trained with 4x more compute in 12 months and 16x more in 24 months.



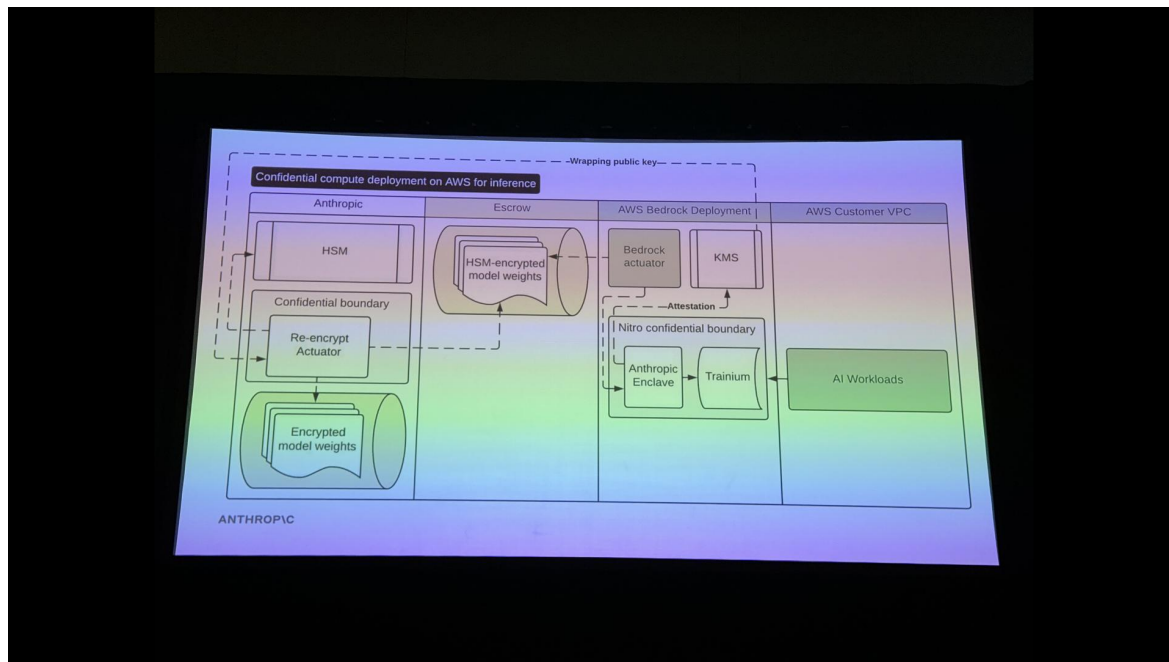
- **Claude's Evolving Role:** The capability of models is projected to evolve from assisting with tasks (2024) to actively collaborating (2025) and eventually pioneering new solutions (2027). The speaker anticipates the first "virtual collaborator" or "virtual employee" systems appearing within 10-12 months.
- **Agents vs. Workflows:**
 - **Workflows:** LLMs and tools are orchestrated through pre-defined paths.

- **Agents:** Tackle open-ended problems where the steps cannot be hardcoded. They can plan, execute, and adapt. The speaker used the analogy of replacing individual "nodes" in a complex business workflow graph with agents, with more capable agents able to take over larger sections of the graph over time.
- **Evolving Memory Ability:** The speaker highlighted the importance of **episodic memory** for agents to learn, adapt, and build context in their specific roles over time.
- **Responsible Scaling Policy (RSP):** Anthropic introduced the Autonomous Scaling Levels (ASL) framework to implement tiered safety protocols as models become more capable.
 - They recently announced achieving **ASL-3**, which involves putting specific guardrails in place to mitigate risks related to CBRN (chemical, biological, radiological, and nuclear) misuse, for example, in drug discovery.



- **Path to ASL-4: Confidential Deployment:** A key step towards higher safety levels involves deploying models in confidential computing environments.
- **Confidential Compute Deployment on AWS:** Anthropic is deploying its models on AWS using Nitro Enclaves and Trainium chips for confidential inference.
- **Goal:** To ensure both customer prompts/completions and the model weights themselves are encrypted and protected while in use.

- **Process:** Model weights are encrypted by Anthropic, held in escrow, and can only be decrypted inside a verified AWS Nitro Enclave after successful attestation. This prevents both the cloud provider and unauthorized actors from accessing the model.



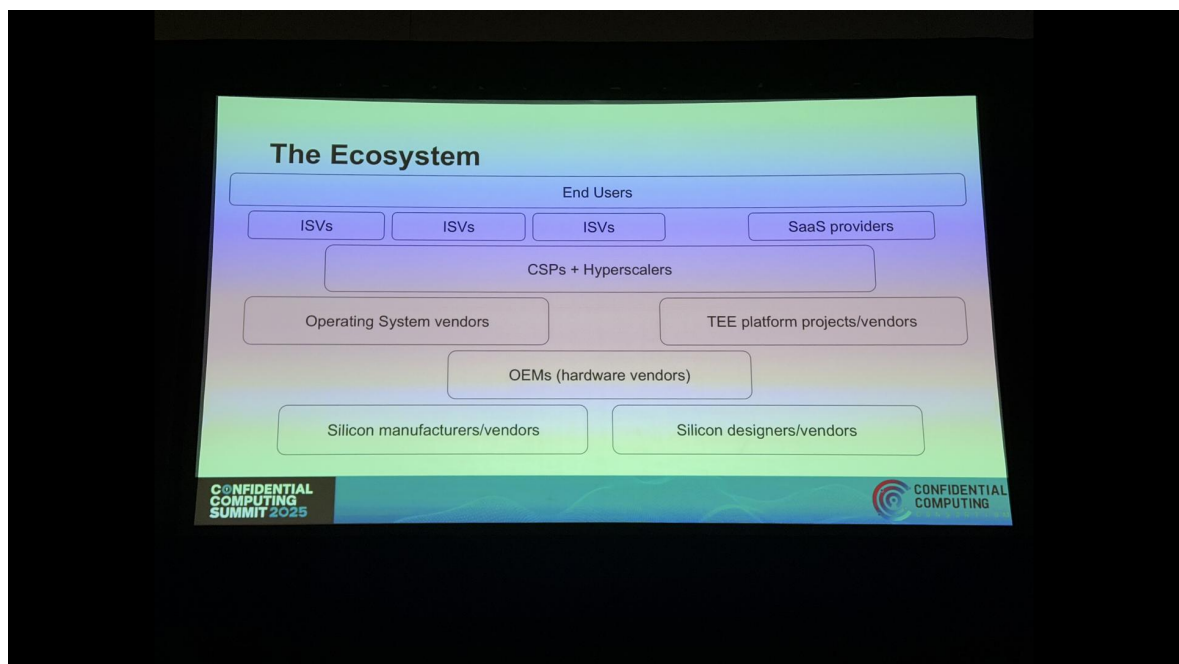
Title: Aligning confidential computing with use cases

Speaker: Mike Bursell from confidential computing consortium - CCC

- **Attestation Process:** Assures that a TEE is running on valid hardware, is correctly set up, and the software inside the TEE is exactly what is expected.
- **Attestation Evidence as a Certificate:** The signed data from an attestation can be treated as a certificate. This is a powerful concept with business value.
 - > "It is a signed bunch of data. And that can be a certificate... You can use it for transport security, TLS, to prove identity. You can prove uniqueness. And you can extend it and use it for signing the output of the application..."
- **Using Certificates for Provenance:** The user asked "how?". The speaker explained that by creating certificates at each step of a multi-step process (like AI training and inference), you can build a verifiable chain of trust.

> "We can trace back all the way to the initial data sets and the models and check the provenance from that certificate, right? So, what we're doing is we've got a process where we're stepping from... piece to piece, step to step, using certificates... to add value..."

- **CCC Ecosystem:** The Confidential Computing Consortium represents the entire ecosystem, from silicon designers and manufacturers at the bottom layer, through OEMs, OS vendors, CSPs, and up to ISVs and end-users. Learn more at confidentialcomputing.io.



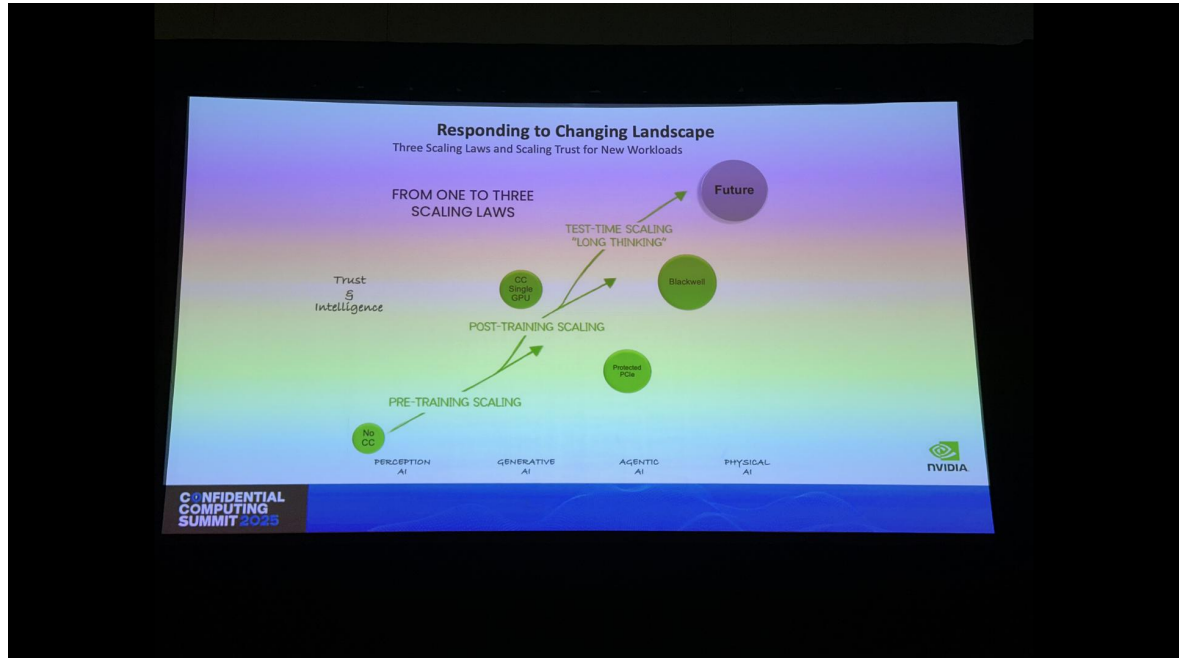
Title: Building the CC ecosystem: how nvidia is shaping the secure AI era

Speaker: Danial Rohrer from Nvidia

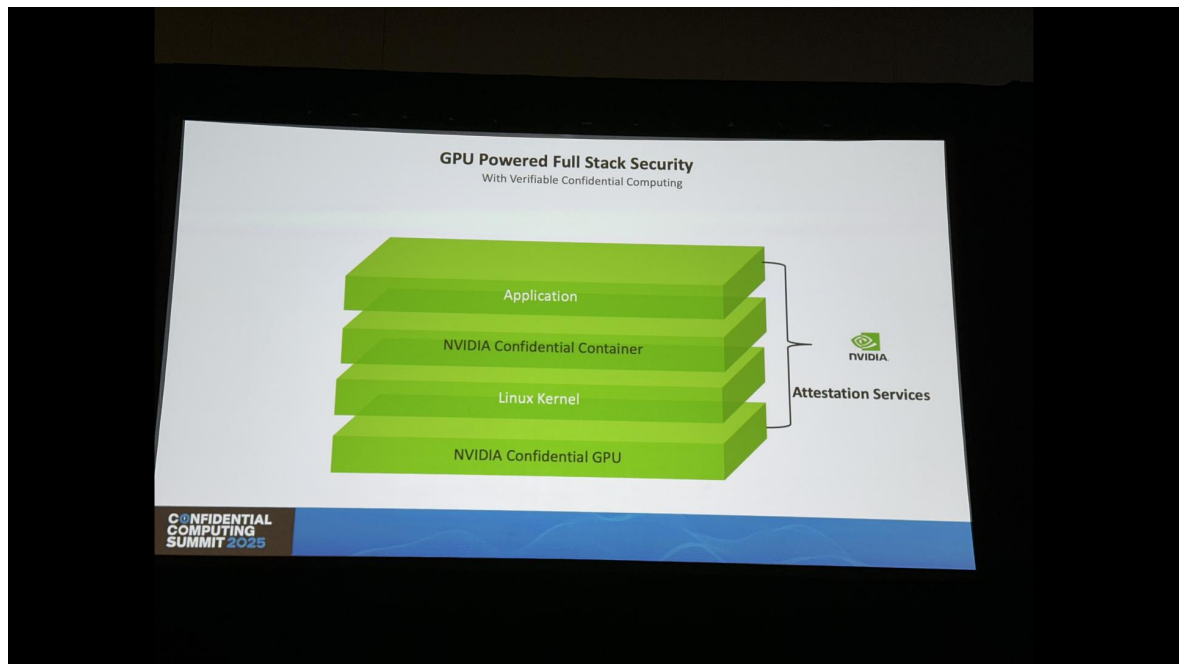
- **From One to Three Scaling Laws:** The AI landscape has evolved, introducing new scaling challenges beyond just pre-training.
 1. **Pre-training Scaling:** The initial law, focused on building foundational models.
 2. **Post-training Scaling:** For tasks like fine-tuning and alignment of generative AI.

3. Test-time Scaling ("Long Thinking")

: For complex, multi-step agentic AI that requires more computation at inference time.



- **Full Stack Security with Verifiable CC:** NVIDIA provides a full stack for secure AI, where each layer is verifiable through attestation.
- **Application**
- **NVIDIA Confidential Container**
- **Linux Kernel**
- **NVIDIA Confidential GPU**
- **Attestation Services:** All layers of the stack are covered by NVIDIA's attestation services, which can be chained with other evidence, such as signed models, to build comprehensive trust.



- **Learn more:** nvidia.com/confidentialcomputing

Title: Making reinvention real with Gen AI: lessons learned from over 2000+ projects

Speaker: Teresa Tung from Accenture

- **Table Stakes vs. Strategic Bets:**

- **Table Stakes:** Foundational, "no regret" investments in GenAI for broad adoption and incremental value (e.g., IT transformation, productivity)
- **Strategic Bets:** Significant, long-term investments with transformative payoffs that are often industry-specific (e.g., asset management, fraud detection).

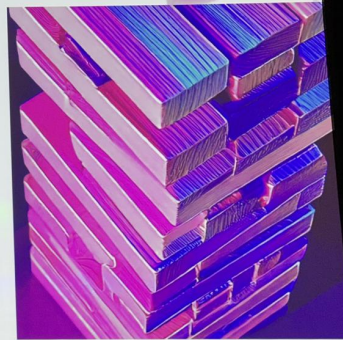
What are "Strategic Bets"?

Table Stakes

Foundational, no regret investments in gen AI that drive broad adoption within an organization, and offer incremental value, typically in productivity and efficiencies
E.g., IT transformation, Gen AI led contact center

Strategic Bets

Strategic bets are significant, long-term investments that have a very large payoff; drive transformative, industry-specific, process-level efficiencies, productivity, innovation and revenue growth
E.g., Asset management, Field worker companion



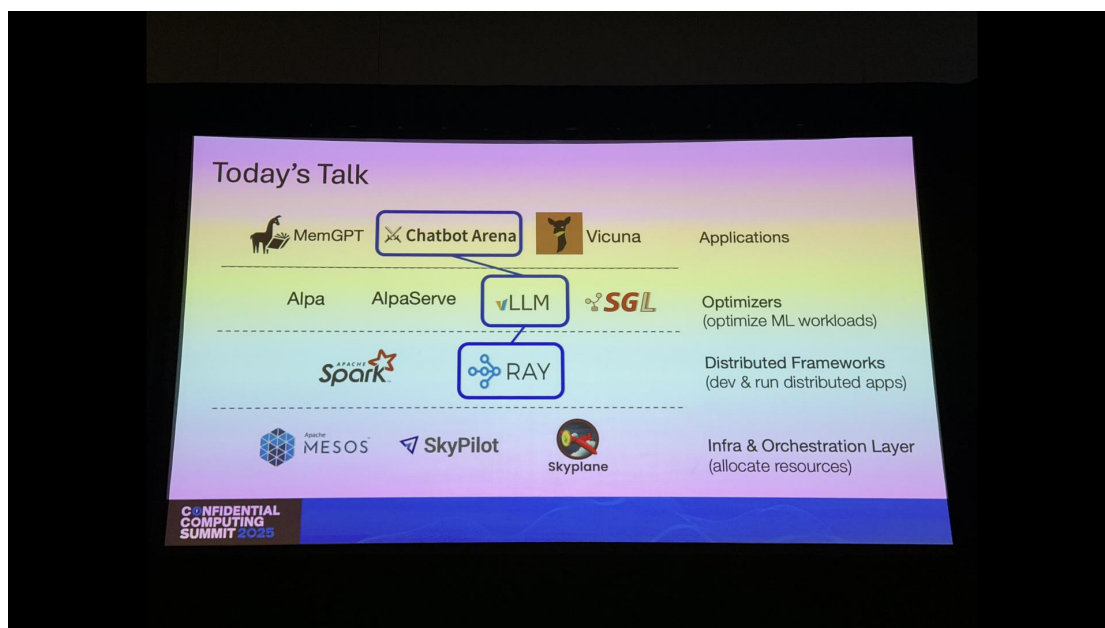
Copyright © 2023 Amazon. All rights reserved.

Title: From scaling AI system to securing them: lessons from building an AI stack

Speaker: Ion Stoica from UC Berkley

- **Ray:** A distributed framework for developing and running AI applications at all stages.
- **Serving Optimizers:**
 - **vLLM:** Uses PagedAttention, a memory management technique inspired by virtual memory and paging in operating systems, to improve LLM serving throughput.
 - **SGLang:** Uses RadixAttention for efficient inference.
- **Data and Evaluation:**
 - **ShareGPT:** A platform for sharing user conversations with models like ChatGPT. This data was used to create the Vicuna model by fine-tuning LLaMa on ~70,000 conversations.
 - **ChatBot Arena:** A crowdsourced platform for evaluating LLMs. It uses a tournament-style, Elo ranking system where humans pick the better of two model responses.
- **Problems with Traditional Evaluation:** Static benchmarks are prone to data contamination and don't effectively capture human preferences, while human evaluators are slow and expensive. The ChatBot Arena model addresses this.

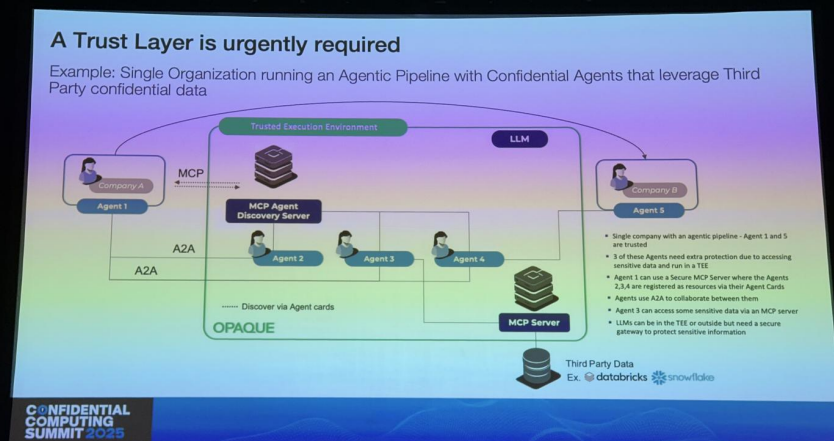
- **Highlighted Stack:** The talk highlighted a specific application stack: **Chatbot Arena** → **vLLM** → **RAY**.



Title: From pilot to production: unlocking enterprise AI through CC

Speaker: (Panel)

- **Industry Adoption:** Major tech firms are embedding confidential computing into their flagship AI services to ensure data privacy.
 - **Apple:** Private Cloud Compute (PCC)
 - **Meta:** Private Processing for WhatsApp
 - **Google:** Confidential Spaces and Vertex AI
 - **Microsoft:** Azure Confidential Computing for AI
 - **Anthropic:** "Clio" for privacy-preserving analysis
- **Trust Layer for Agentic Pipelines:** As workflows become more complex, involving multiple AI agents and third-party data, a trust layer is required. Opaque presented a use case of a confidential agentic pipeline where TEEs and a service mesh provide governance and security.



Lunch w/ Daniel from Meta, talk about cryptography, quantum computing, etc.

- Daniel works on confidential computing for WhatsApp using TEEs.
- Shor's algorithm for quantum computers can break both RSA and ECC, the foundational algorithms of modern public-key cryptography.

Breakout sessions:

Title: Verifiability challenges: for confidential compute workloads

Speakers: Sarah de Haas, Hannah Lee, Giles Hogben from Google

- **Oak:** Google's project for building applications on a confidential computing foundation that can make provable claims about data handling.
- **Core Challenge:** Providers need to demonstrate the trustworthiness (security, privacy, integrity) of their CC infrastructure to users. This is especially true as CC

makes stronger claims about tenant workload processing, particularly for AI privacy.

- **Examples of services making such claims:** Apple PCC, WhatsApp Private Processing, Google Oak, Confidential Federated Analytics, OpenMined's PyTorch on TEEs, and Project Veracruz.

Part 2.

- Confidential Computing increasingly makes claims about tenant/workload processing, esp AI Privacy.
- Examples:
 - Apple's Private Cloud Compute
 - WhatsApp's Private Processing.
 - Google's Project Oak and Confidential Federated Analytics.
 - OpenMined's pytorch on TEEs (Includes Model integrity)
 - Project Veracruz

Google

- **Key Problems to Solve:**

- Tenant binary attestation (making end-users the verifiers)

- Verifying proprietary code and binaries
- Auditability and transparency
- Key provisioning for attestation
- Scalability of trust in distributed systems
- Comparability of binary transparency logs

- **Core Principle:** Don't trust, verify! The goal is to build a robust verifier ecosystem.

Title: A Blueprint for Trust in multi-agent system

Speakers: Hernan Gatta, Rishabh Poddar from Opaque

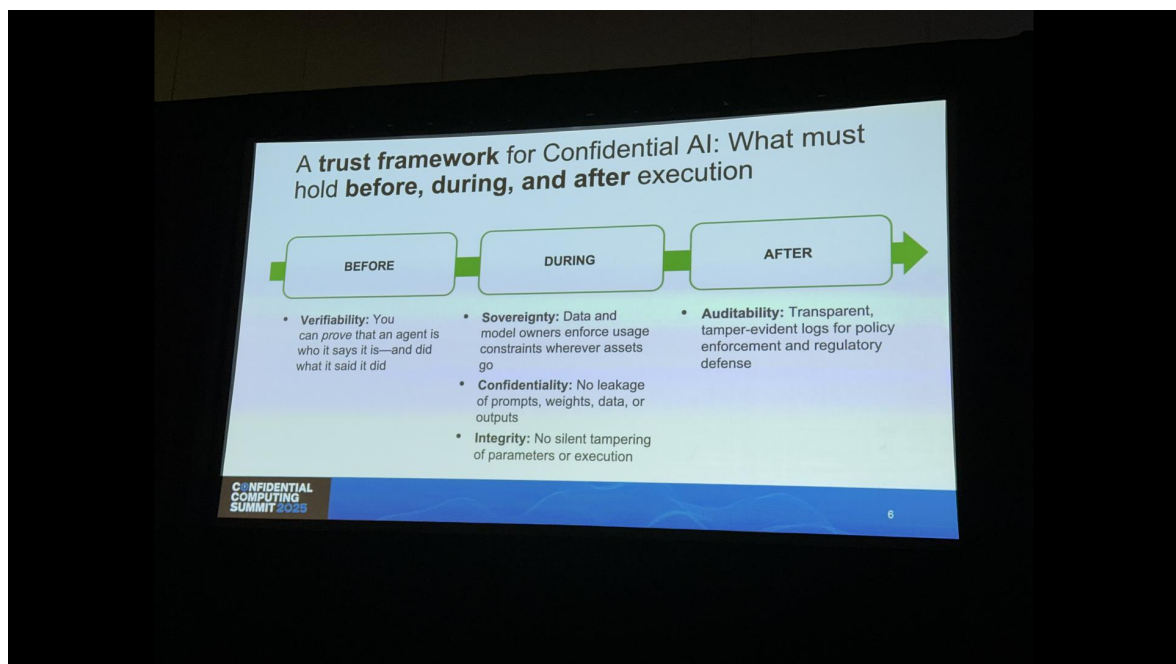
- **Regulatory Shift:** Moving from guidance to enforceable mandates like the EU AI Act.

- **Challenges in Multi-Agent Systems:**

- Fragmented landscape of stakeholders
- Non-determinism and autonomy of agents
- Jurisdictional and compliance boundaries

- **A Trust Framework for Confidential AI:**

- **BEFORE Execution: Verifiability** (Prove an agent is who it says it is and did what it said it did).
- **DURING Execution: Sovereignty** (Data owners control usage), **Confidentiality** (No leaks), **Integrity** (No tampering).
- **AFTER Execution: Auditability** (Tamper-evident logs for enforcement).



- **Enforcing Trust with a Confidential Service Mesh:**

- **Confidentiality & Integrity:** TEEs ensure code and data are protected at runtime. Remote attestation establishes a hardware-rooted chain of trust. The principle is to *trust the workload, not the service provider*.

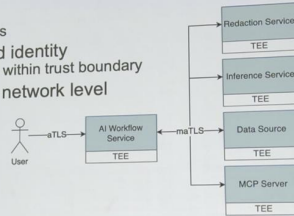
- **Sovereignty:** A **transitive mesh of trust** is created using **Attested TLS (aTLS)**

- * **and** ***mutual aTLS (maTLS)**. This binds the cryptographic channel identity to the workload identity, ensuring the encrypted channel terminates inside a verified TEE. This creates a graph of trust rooted in the user.

Sovereignty

Attested TLS & maTLS Mesh

- Transitive mesh of trust
 - One- and two-way attestation across trust boundaries
- Graph of trust rooted in the user
 - Extends as requests cross trust boundaries
- Binding between channel and workload identity
 - Asserts that encrypted channel terminates within trust boundary
- Assert behavior and guarantees at the network level
 - Transparent and workload-agnostic



CONFIDENTIAL
COMPUTING
SUMMIT 2025

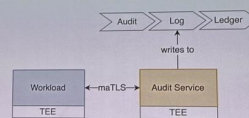
11

- **Auditability:** Achieved through hardware-rooted, tamper-evident logs where traces are accepted only from attested workloads and bound to their identity.

Auditability

Hardware-rooted, tamper-evident logs

- Attested log sources
 - Only accept traces from trustworthy workloads
- Audit logs bound to workload identity
 - Bind inputs and outputs to attestation evidence
- Hash-extend traces
 - Assert append-only semantics
- Optionally decentralize audit logging
 - Ensure no single entity controls audit chain



CONFIDENTIAL
COMPUTING
SUMMIT 2025

12

- **Shared Vision:** Confidential AI as a Public Infrastructure.

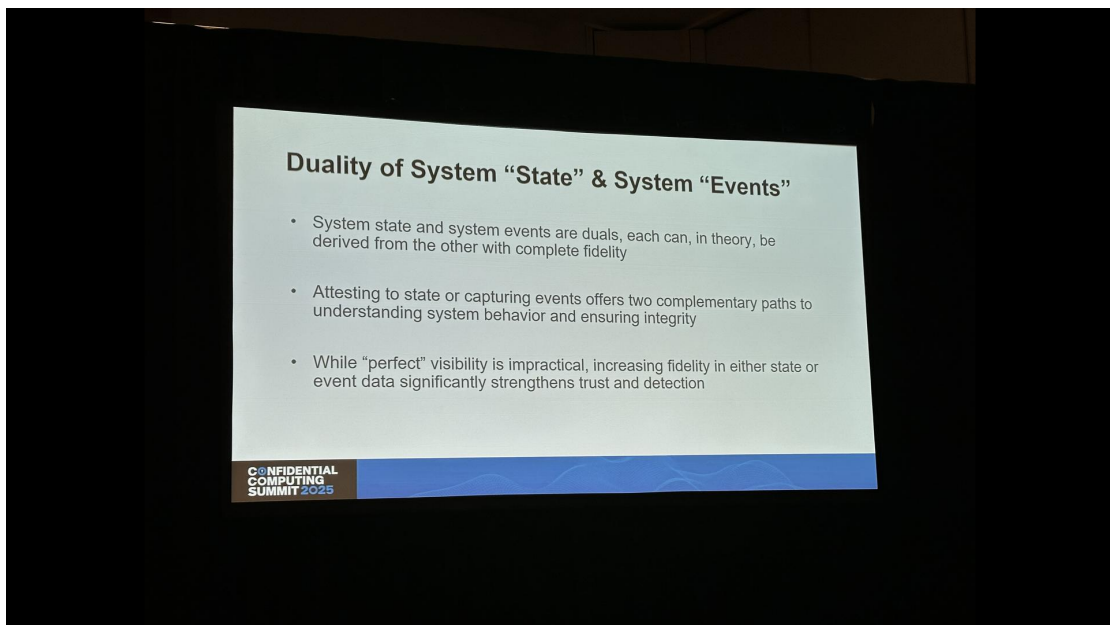
Title: Runtime attested HPC cluster reference architecture

Speakers: Wesley Peck and Jason Rogers from Invary

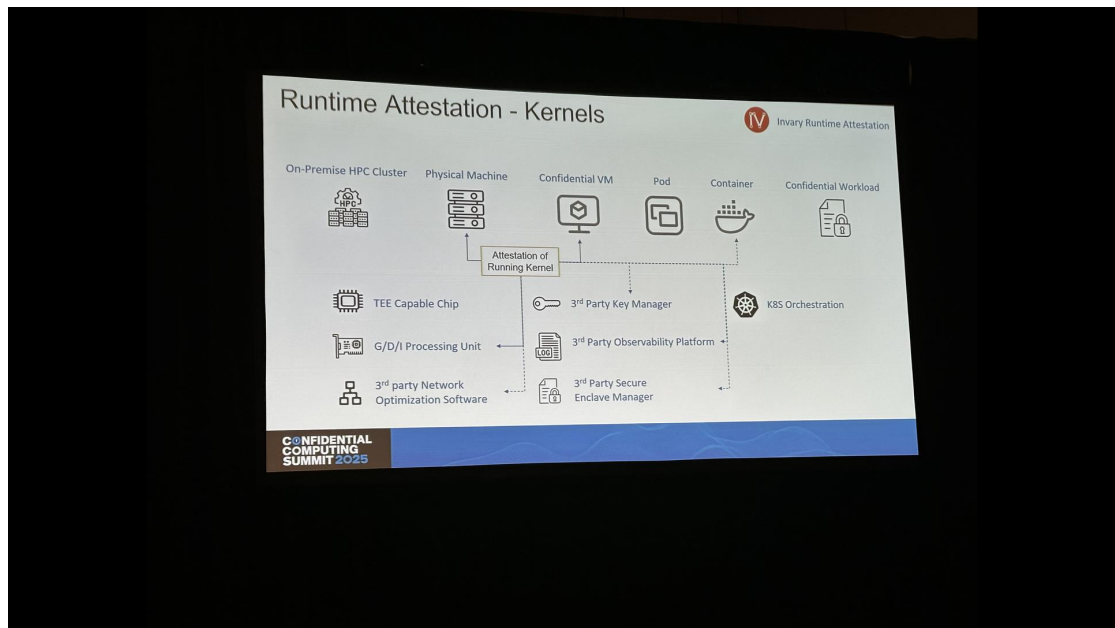
- **Architecture Principles:**

- Avoid reimaging and redeployment for remediation.
- Implement attestation gates at build, deploy, and job execution stages.
- Identify and attest to "hidden" infrastructure.
- Blend state-based attestation with event-based observability.

- **Duality of State and Events:** System state and system events are duals; each can theoretically be derived from the other. Attesting to state and capturing events are two complementary paths to ensuring system integrity.



- **Runtime Attestation of Kernels:** The focus is on verifying the software running in memory, as this is the ultimate source of truth for system behavior. The disk state is less important than the runtime state. A chain of trust is established from the silicon chip up to the running kernel in memory.

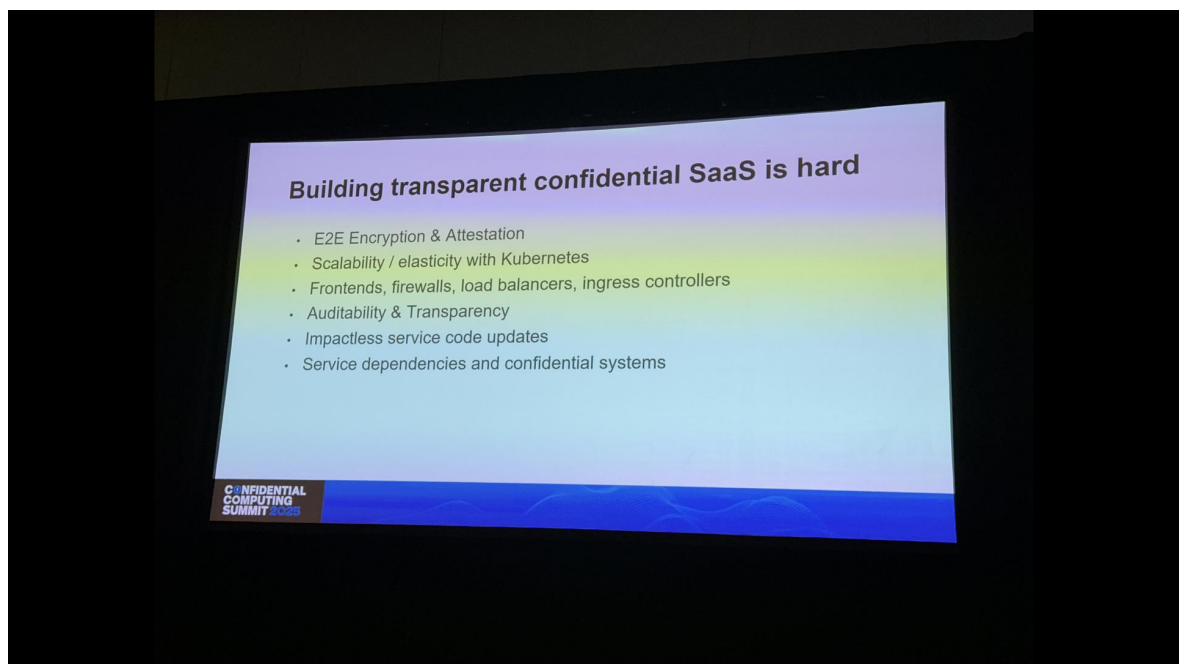


Title: Building confidential inferencing (and other SaaS)

on Azure

Speaker: Antoine Delignat-Lavaud from Azure

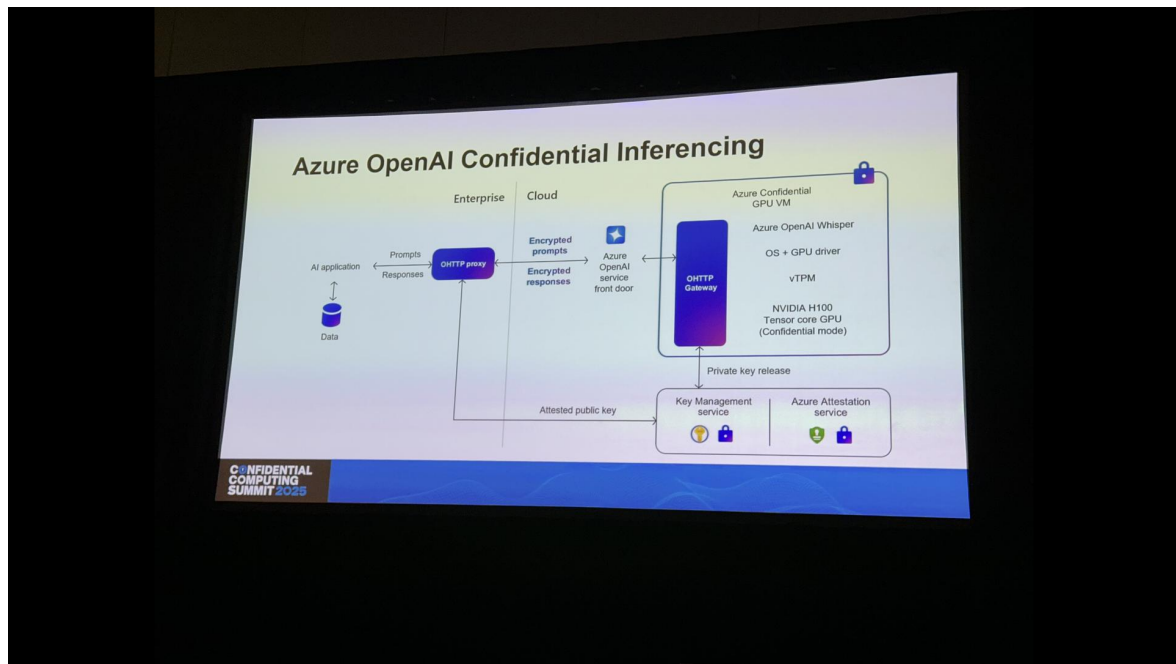
- **Challenge:** Building transparent, scalable, and auditable confidential SaaS is difficult. It involves E2E encryption, attestation, managing load balancers and firewalls, seamless code updates, and handling dependencies.



- **Azure OpenAI Confidential Inferencing:** An architecture that allows clients to use OpenAI models without exposing their prompts to Microsoft or untrusted components.

- The protocol used is **OHTTP (Oblivious HTTP)*

*,



- **Why use OHTTP?**

- It encrypts the request at the application level, so it can pass through untrusted middleboxes.
- Load balancing is stateless, as any TEE with the private key can decrypt the request.
- It is built on semi-static HPKE, allowing for easy request retries.
- It can help anonymize clients.

Why use Oblivious HTTP?

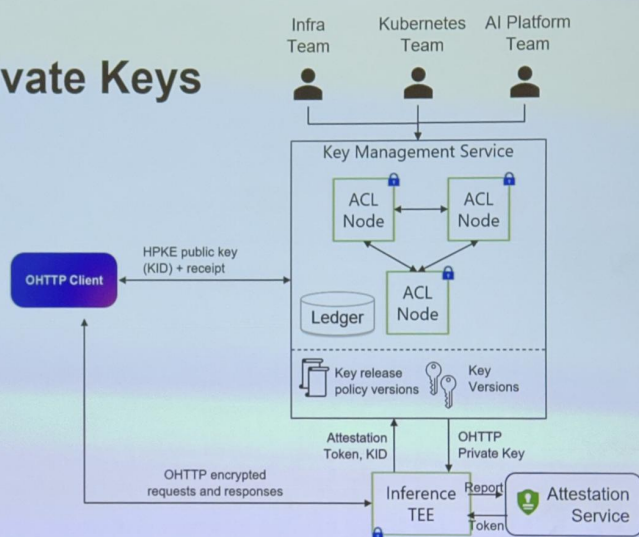
- OHTTP encrypts the request at the application level, so can pass through any L4 or L7 middlebox
 - In real Azure OpenAI systems, at least 4 untrusted middleboxes
- Any TEE that has the OHTTP private key can decrypt any request
 - Load balancing is stateless
- OHTTP is built on semi-static HPKE
 - Failed requests can be retried without client action
- OHTTP can help anonymize clients
- Relatively easy to implement in most client including browser

CONFIDENTIAL
COMPUTING
SUMMIT 2025

- **Managing OHTTP Private Keys:** Keys must be refreshed frequently. Different teams (Infra, Kubernetes, AI Platform) update different parts of the key release policy. This is managed using a key management application built on Azure Confidential Ledger.

Managing OHTTP Private Keys

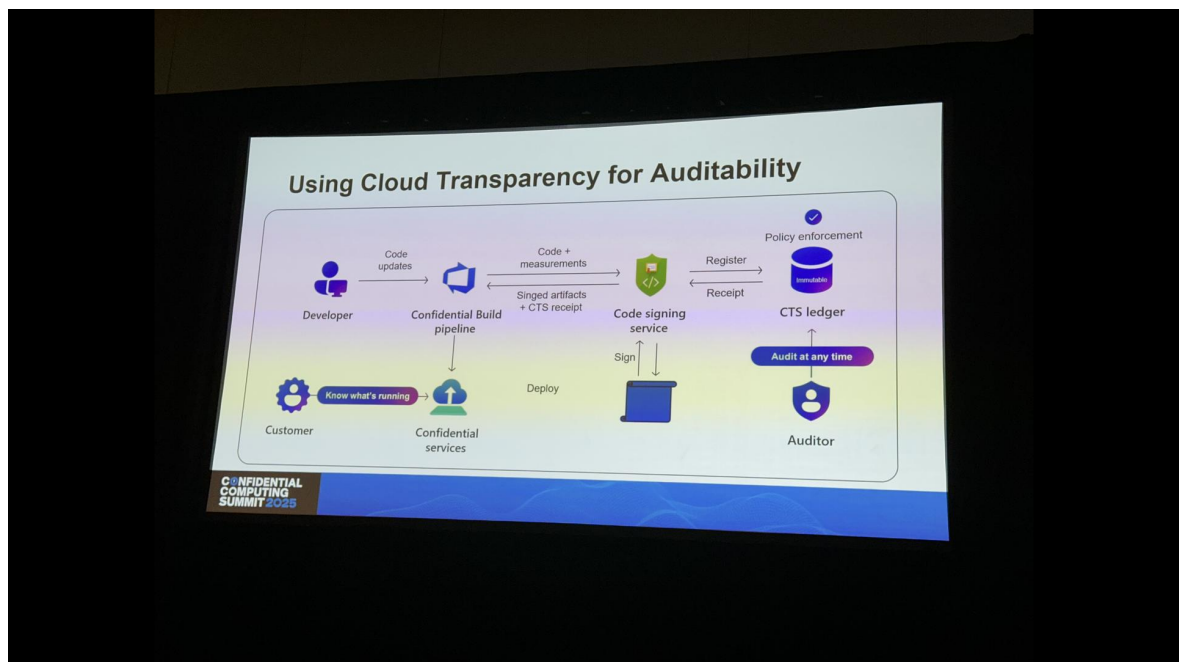
- OHTTP uses 8-bit Key Identifier
- Keys must be frequently refreshed
- Different teams must update different parts of the key release policy:
 - Infra: BIOS/VBIOS versions
 - Kubernetes: base OS image
 - AI Platform: pod deployments
- KID/SKR updates must be transparent
- We built a key management application on Azure Confidential Ledger



[GitHub - microsoft/azure-privacy-sandbox-kms](https://github.com/microsoft/azure-privacy-sandbox-kms)

- **Auditability via Cloud Transparency:** Code updates go through a confidential build pipeline, are signed, and a receipt is registered in an immutable Cloud Transparency Service (CTS)

ledger. This allows developers, customers, and auditors to verify what code is running at any time.

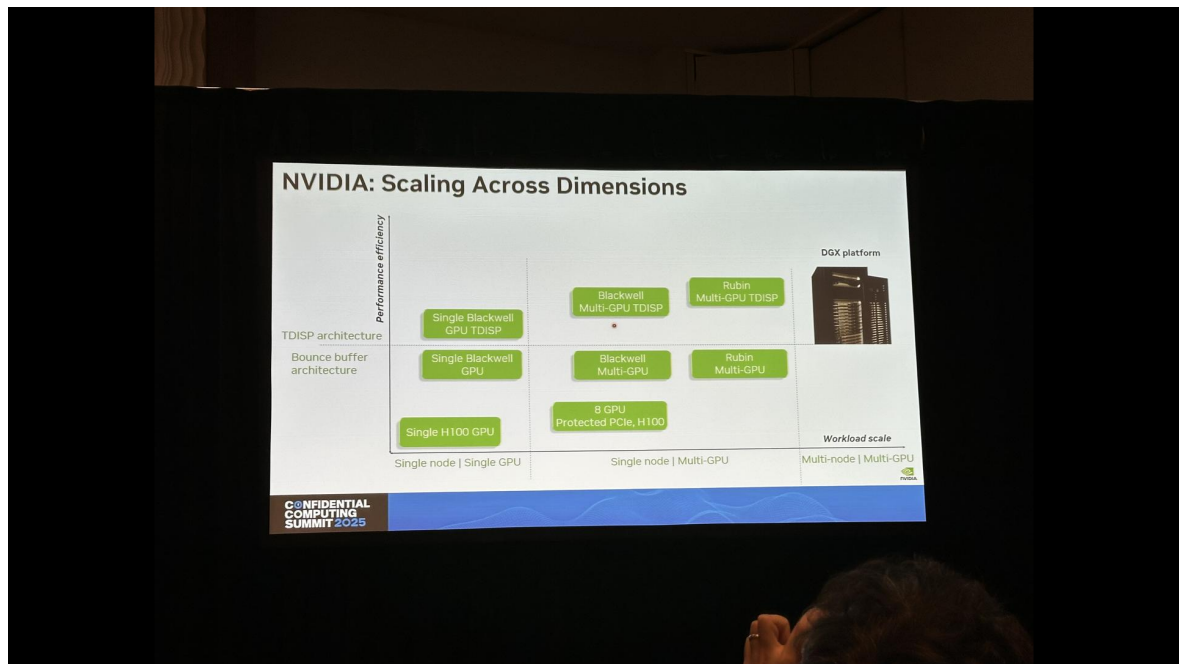


- **Service Dependencies with Attested DNS:** A system for securely discovering and connecting to other confidential services. (Ref: arxiv.org/2303.14611)

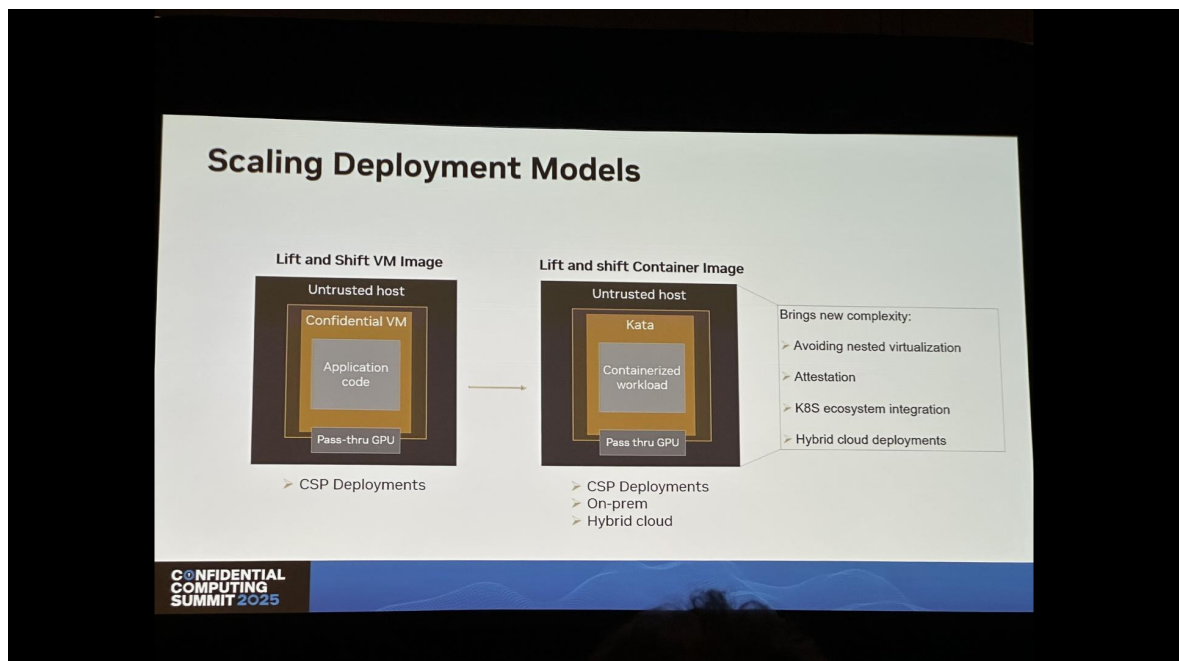
Title: Nvidia North Star: confidential computing everywhere

Speaker: Paul Johnson from Nvidia

- **Bounce Buffer vs. TDISP Architecture:** These are two architectures for enabling confidential computing between CPUs and GPUs.
 - **Bounce Buffer:** A software-based approach where the CPU encrypts data in a shared "bounce buffer" in memory, which the GPU then reads and decrypts. This is less efficient.
 - **TDISP Architecture:** A hardware-based approach using the TDISP protocol over PCIe to create a direct, secure channel between the CPU and GPU. This offers much higher performance efficiency.
 - The DGX platform will support the TDISP architecture for multi-node, multi-GPU deployments.

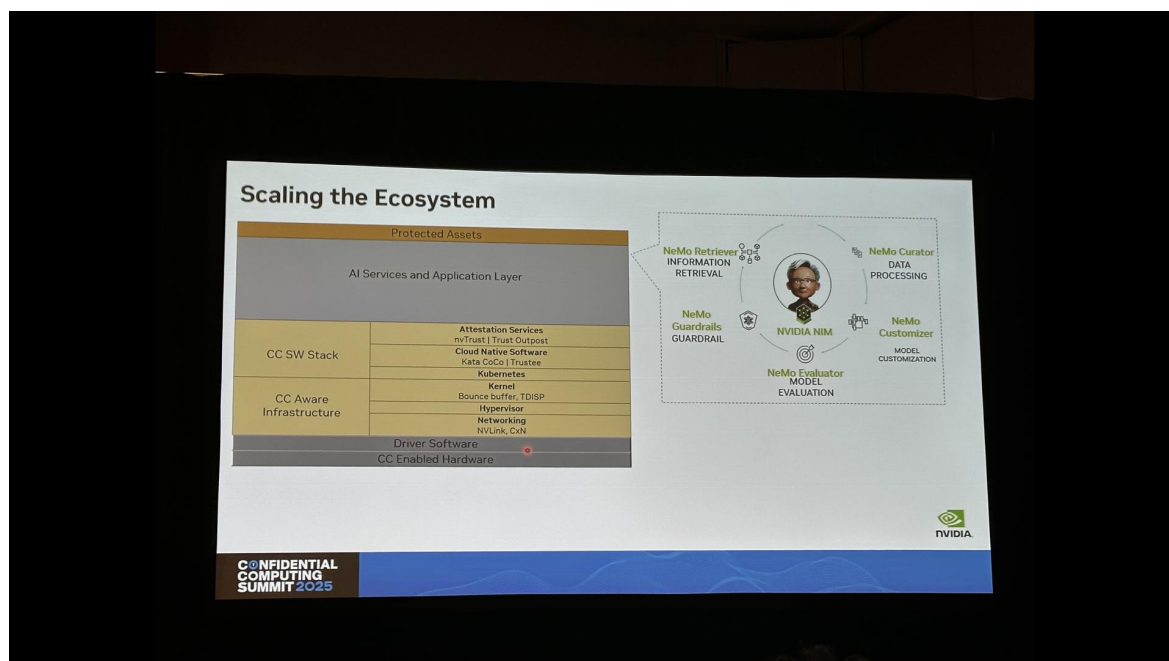


- **Scaling Deployment Models:** Moving from a "Lift and Shift VM Image" to a "Lift and shift Container Image" (using Kata Containers) enables deployment on-prem and in hybrid clouds, but introduces complexity around attestation and Kubernetes integration.



- **Scaling the Ecosystem:** NVIDIA is building out a full stack, from hardware to AI services (like the NeMo framework)

, to enable confidential AI. This includes attestation services, cloud-native software (Kata), and kernel-level support.



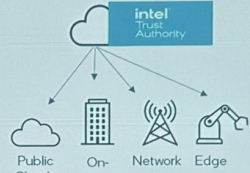
Title: Trust without Borders: independent attestation for hybrid and confidential AI deployments

Speakers: Rene Kolga from Google cloud, Raghu Yeluri from Intel

- **Who provides attestation verification?** A key question is whether verification should be done by the infrastructure provider or an independent third party. Independent verification aligns with Zero Trust principles ("separation of duties")
- **Intel® Tiber™ Trust Authority:** An independent attestation verification service from Intel.
 - Provides a single set of APIs for workload attestation across multiple clouds and on-prem environments.
 - Supports unified attestation for heterogeneous environments, such as those with NVIDIA H100 GPUs and Intel TDX CPUs.
 - Extends the chain of trust up to the workload level.
 - Learn more at: intel.ly/45rmbEY

Intel® Tiber™ Trust Authority

- ✓ Independent Verification (Separates the provider of infrastructure from verifier of trust); adheres to Zero Trust principles.
- ✓ One set of APIs for Confidential workload attestation across Multiple Clouds.
- ✓ Unified Attestation with NVIDIA H100 and Intel TDX
- ✓ Chain of Trust extended to workloads.
- ✓ Uniform Attestation Token Format for different Relying Parties across CSPs.



Learn more at:
<https://intel.ly/45rmbEY>

CONFIDENTIAL COMPUTING SUMMIT 2025

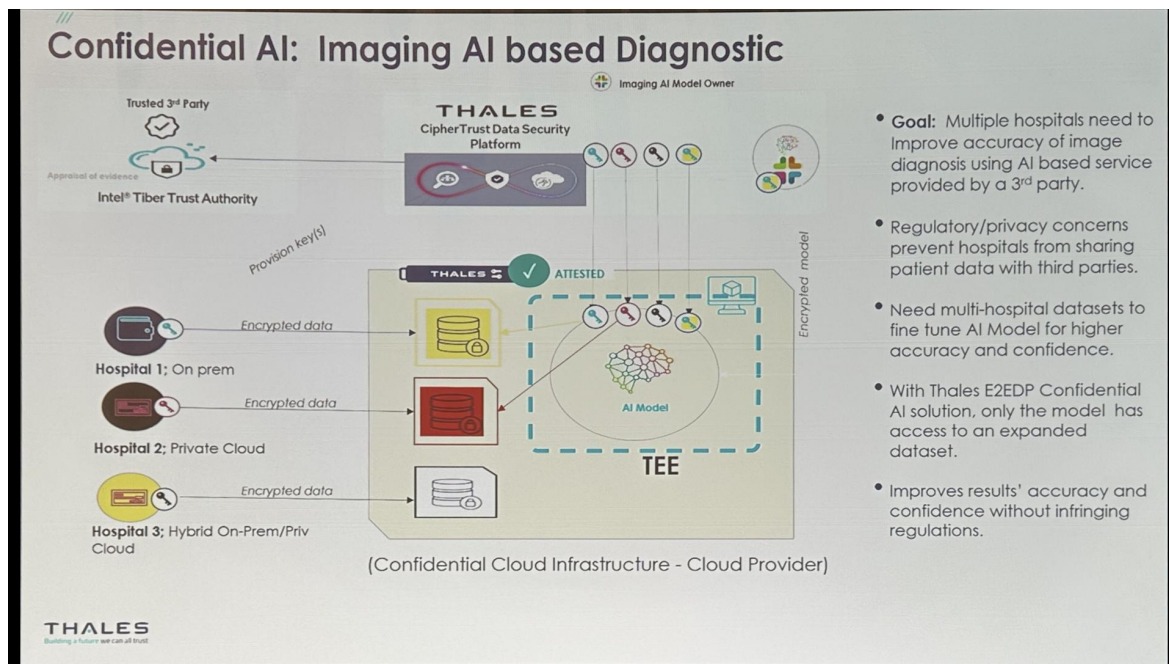
- **Google Cloud Attestation Service:** Google Cloud Confidential Space (built on Intel TDX)

can use either Google's attestation service or the Intel Tiber Trust Authority to verify the environment before releasing protected resources.

- **Use Case: Imaging AI for Diagnostics (with Thales):**

- **Goal:** Multiple hospitals want to use a third-party AI service to improve diagnostic accuracy but cannot share patient data due to regulations.

- **Solution:** The model is fine-tuned inside a TEE in a confidential cloud environment. Each hospital provides encrypted data. The Thales platform manages the keys, which are only released to the attested TEE. This allows the model to be trained on the combined dataset without any party (not even the cloud provider or the model owner) seeing the raw patient data.



Title: summit insights day 2

Speaker: (panel)