



Confidential AI for the Masses

Julian Stephen & Michael Le



Agenda

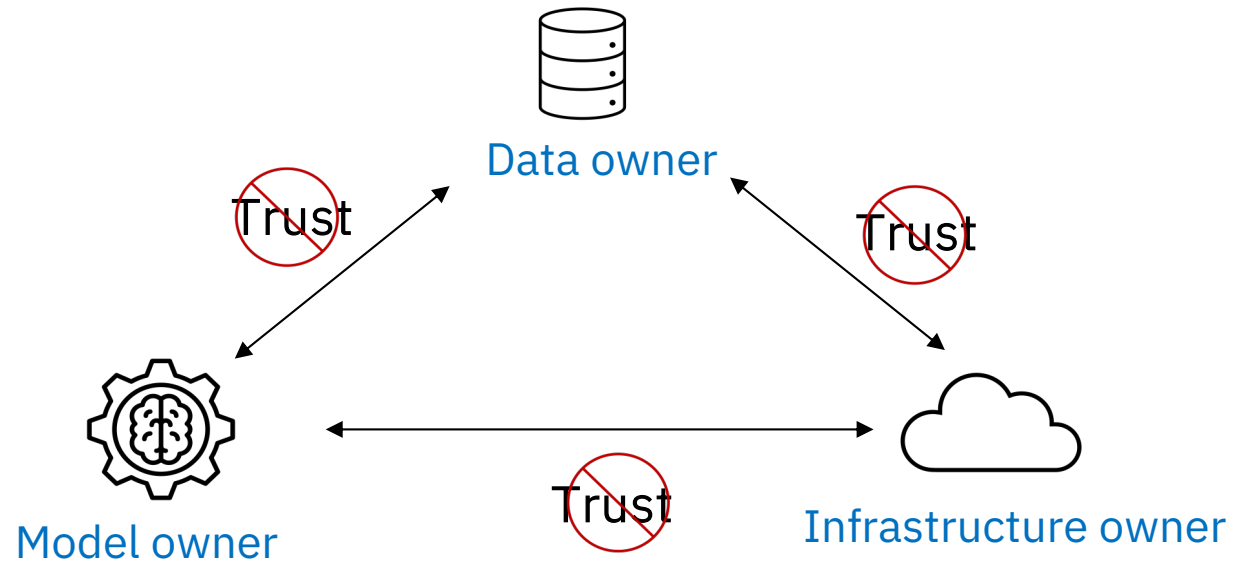
- AI use cases and the reshaping of compute
- The case for confidential AI
- Who you trust changes everything
- Key elements of confidential AI
- A practical confidential AI solution
- What is still missing



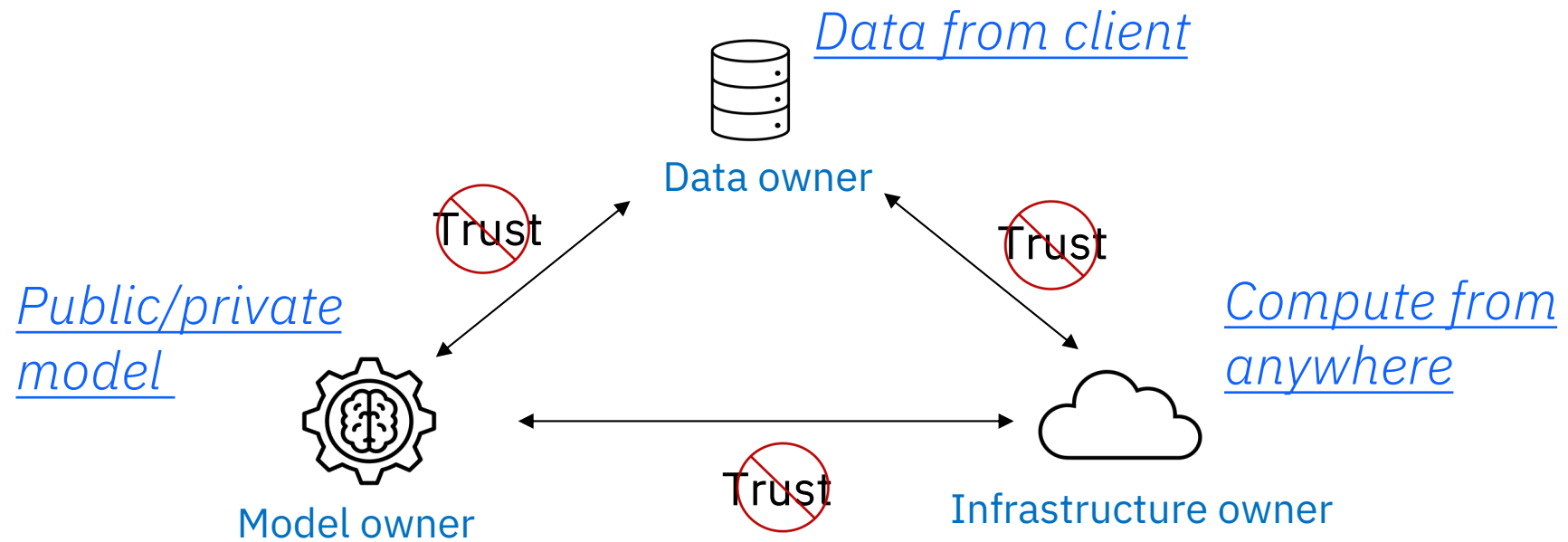
The **new calculus** of enterprise AI

Enterprise AI = Public/private model +
Data from client +
Compute from anywhere

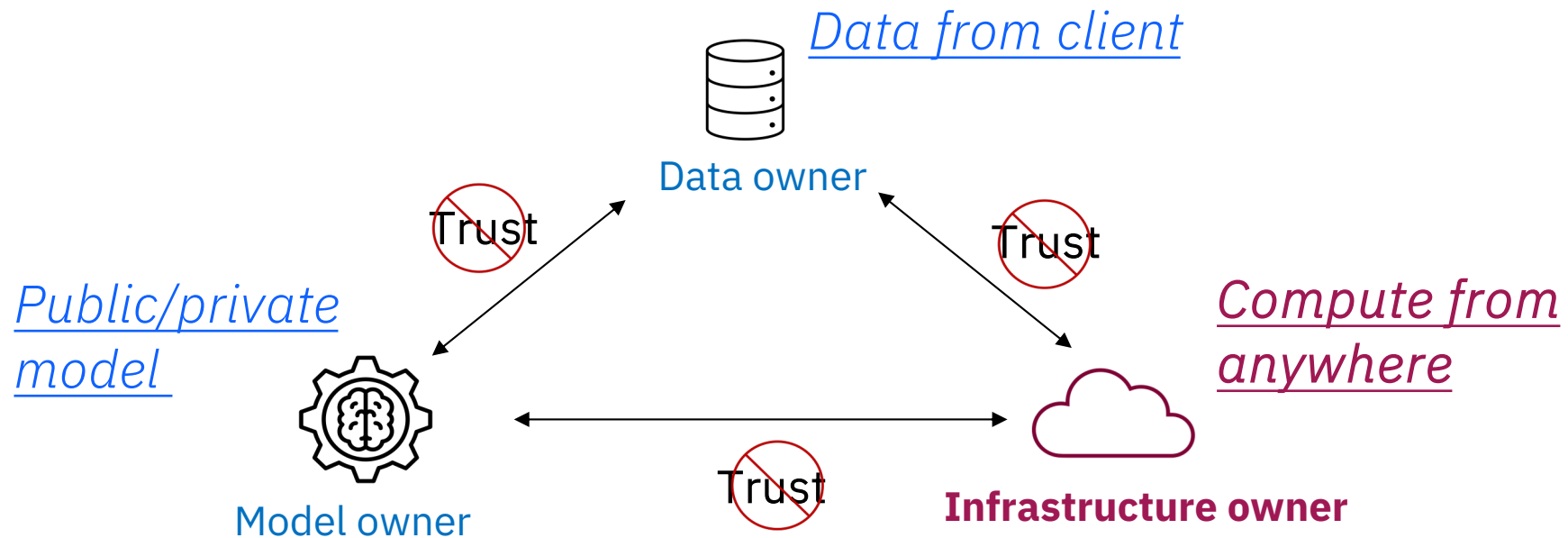
Multiple parties, mutual distrust



Multiple parties, mutual distrust

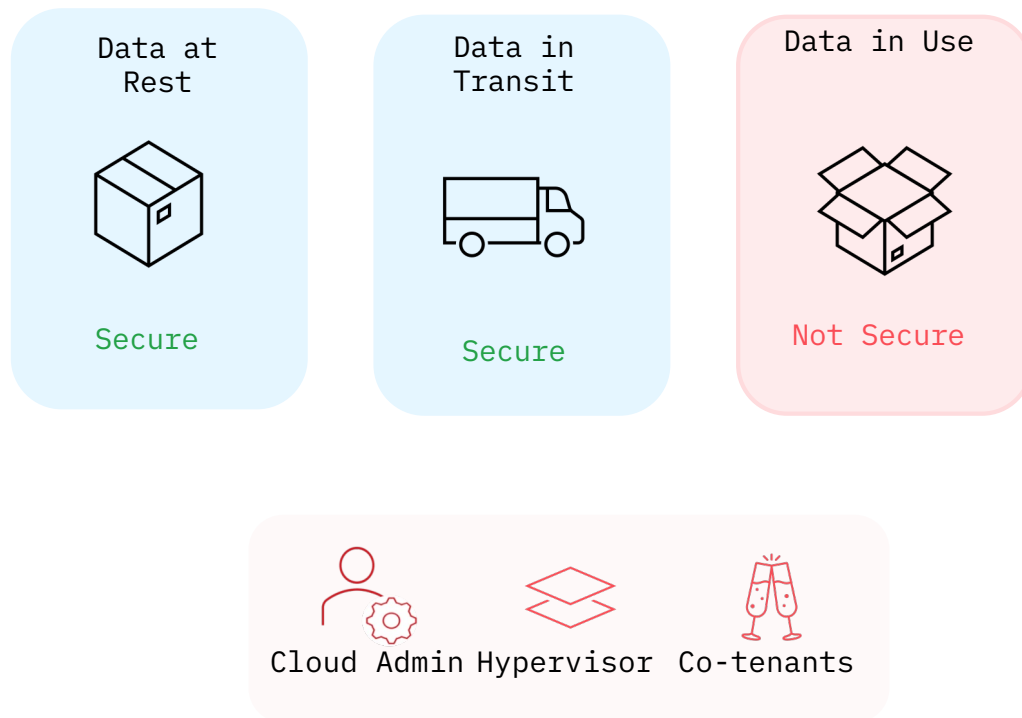


Multiple parties, mutual distrust



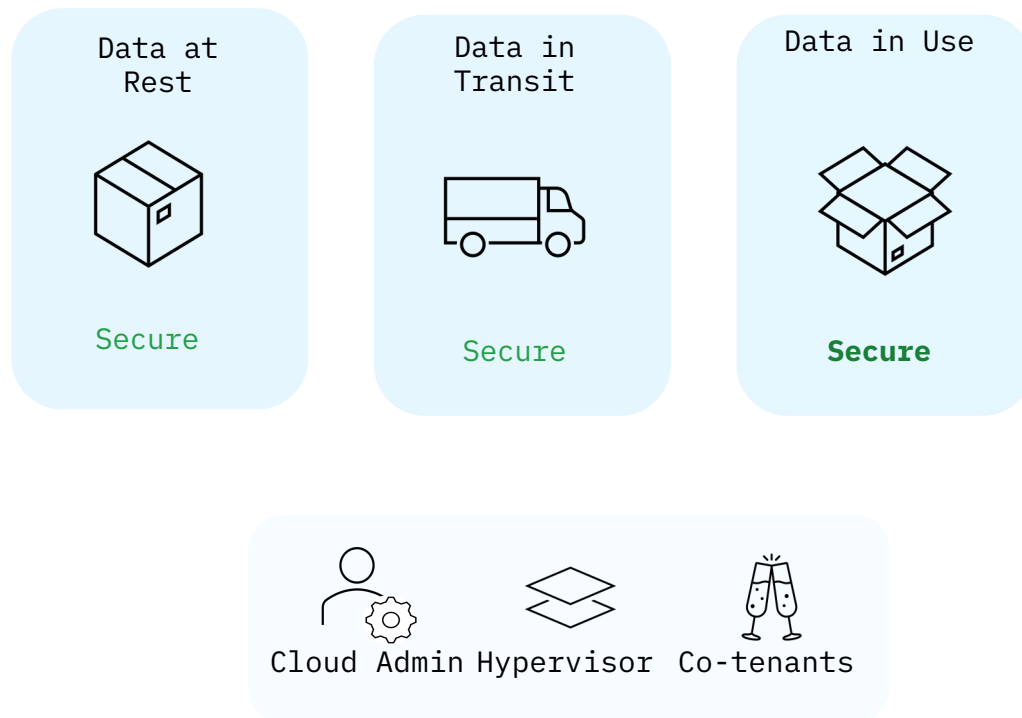
Clear, unencrypted data and code
during processing inside the VM

Why do we need confidential computing?



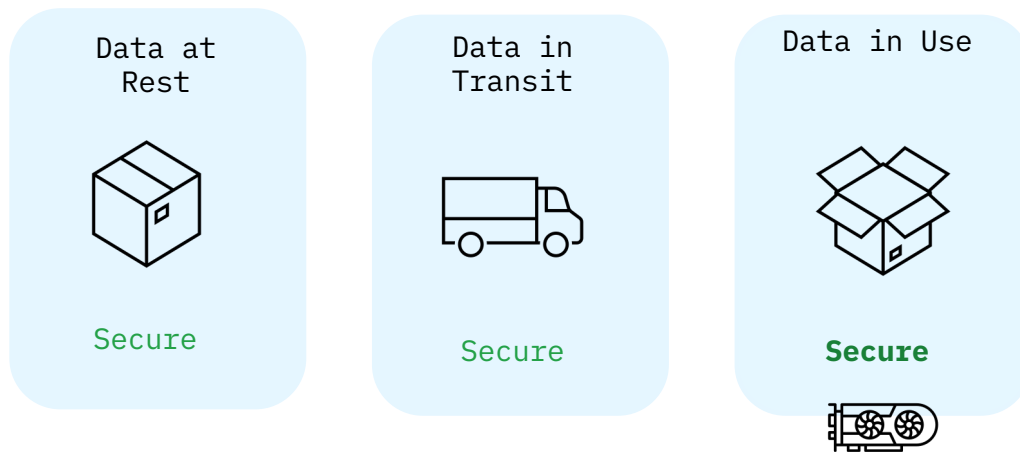
- Little/no protections against weaknesses in
 - Hypervisor
 - Host OS
 - Infrastructure owner
 - Physical access
- Cloud users need to trust
 - Co-tenants
 - Cloud provider
 - Hypervisor
 - Hardware vendors

Confidential computing



- Protect data in use by performing computation in a hardware-based, attested TEE
- Supported by:
 - AMD EPYC, Intel TDx, IBM Z Secure Execution, ARM CCA, ...
- No need to trust
 - Co-tenants
 - Cloud provider/admin
 - Hypervisor
- Need to trust
 - Hardware vendor

GPU confidential computing



- Protect CPU to GPU channel
 - Security Protocol and Data Model (SPDM) session to securely connect the GPU to the GPU driver in a CPU TEE
- Chain of trust established through GPU boot sequence, with secure and measured boot
- Cryptographically signed measurements
- Supported by:
 - NVIDIA Hopper, Blackwell series GPU

A new way to **securely deploy** LLMs

Confidential AI Workflows Anywhere

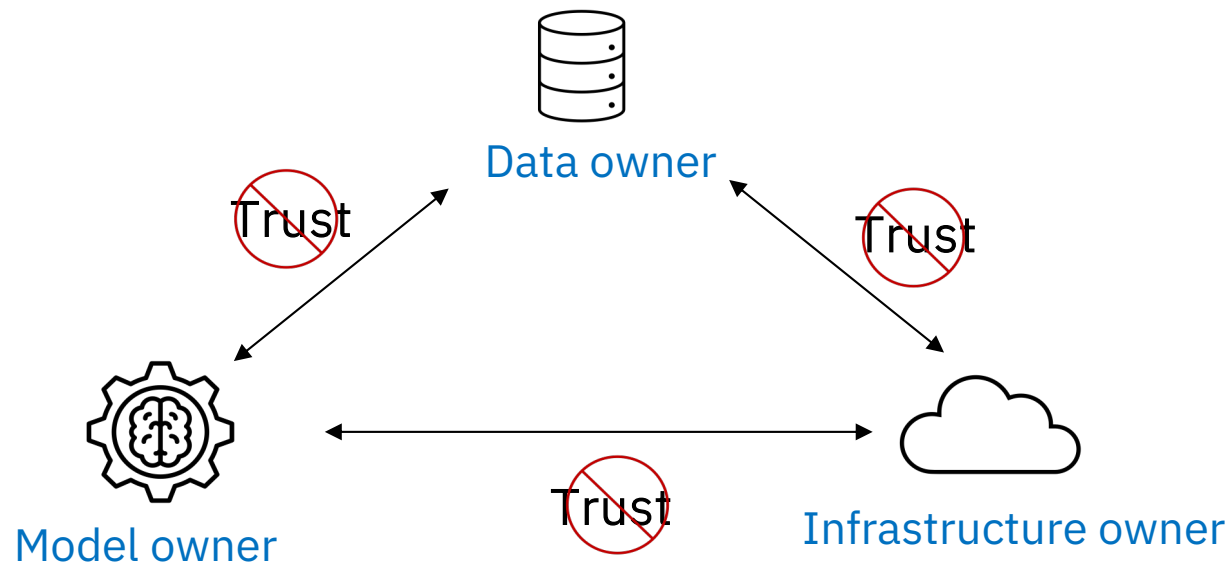
AI Workflows: Large-scale inference / training / fine-tuning with multiple confidential GPUs

Confidential AI Workflows Anywhere

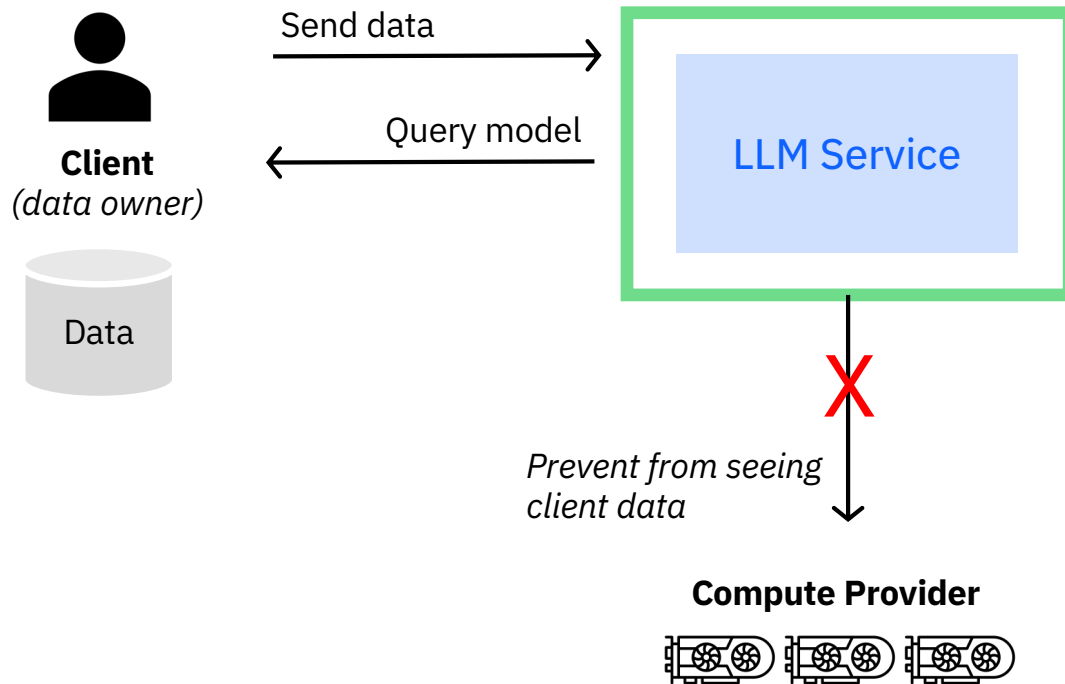
Confidential: Harness confidential computing for secure AI among mutually distrusting parties

Anywhere: on-prem, on the cloud, and anywhere in between

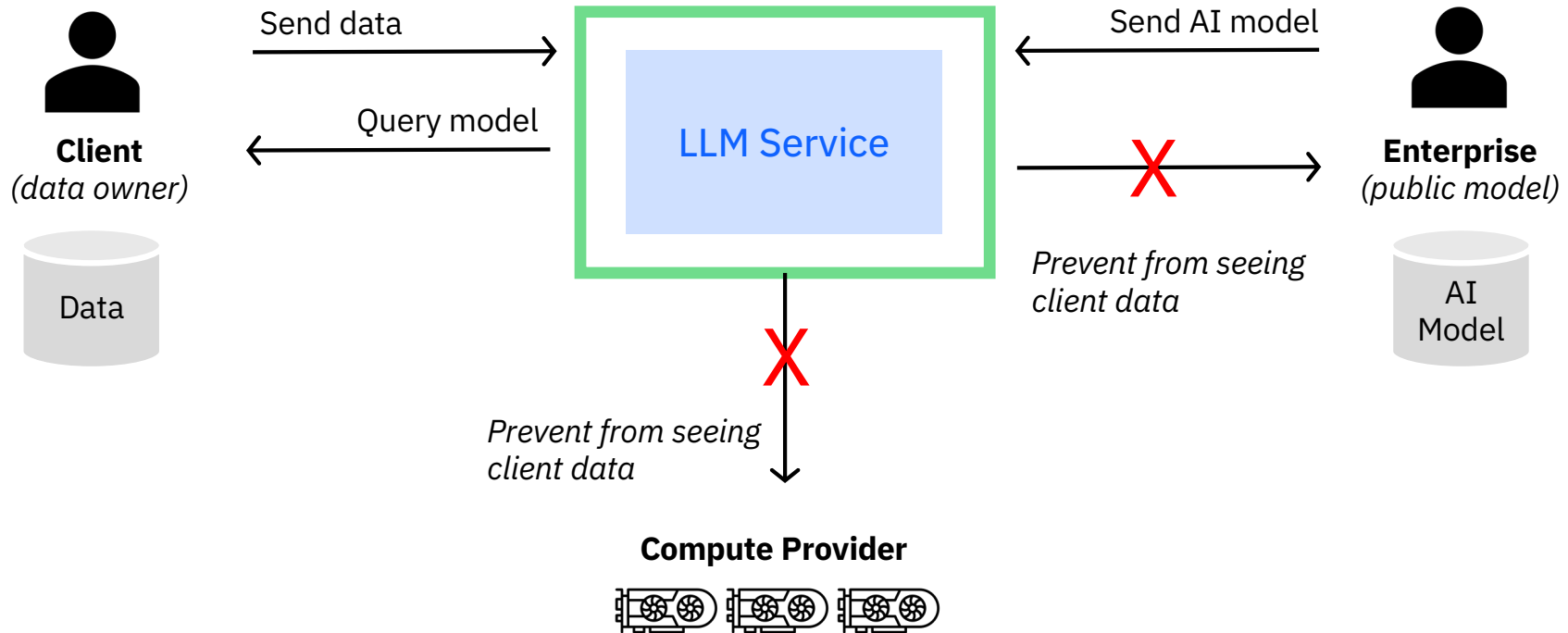
Multiple parties, mutual distrust



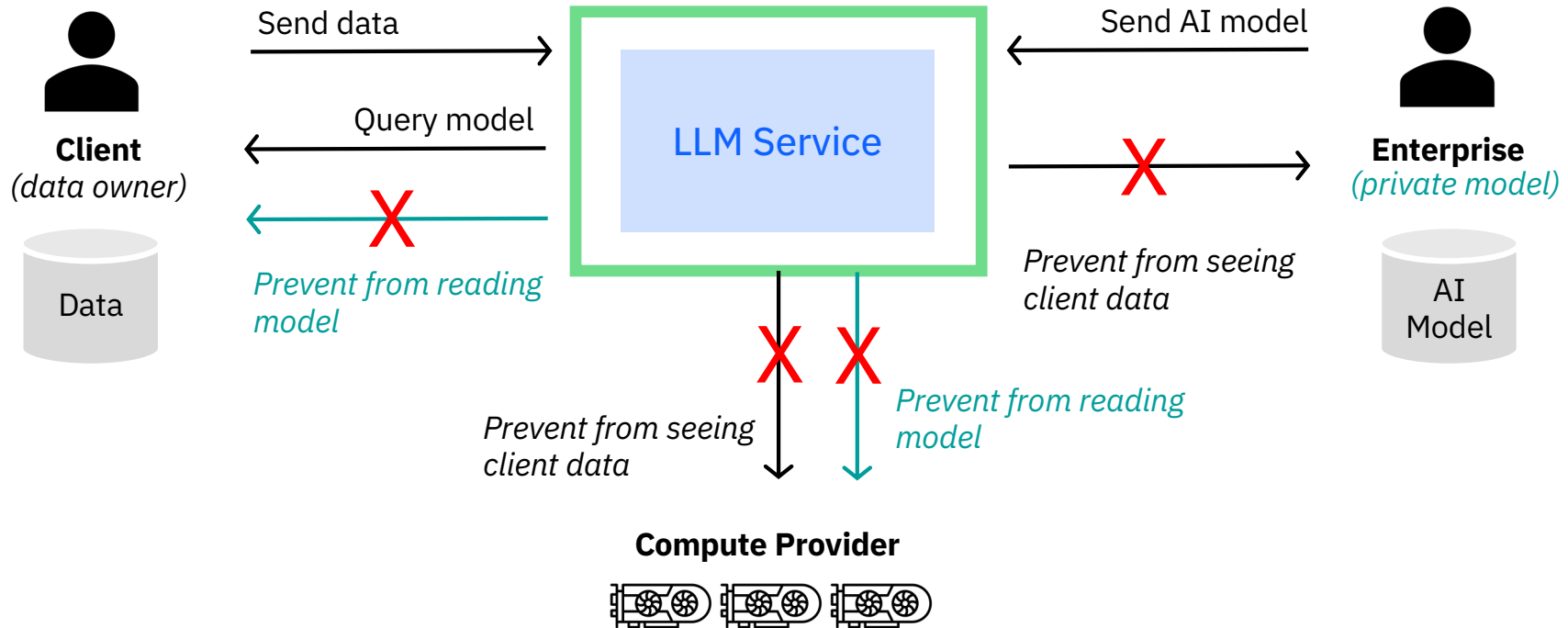
Examples: Growing **distrust** by everyone: client & cloud



Examples: Growing **distrust** by everyone: client, cloud & model vendor



Examples: Growing **distrust** by everyone: client, cloud, & confidential model



Is confidential AI this simple?



Lease CVM

Instantiate confidential VM
with confidential GPU from
cloud provider



Deploy Inference Server

Login to CVM, install your
favorite LLM server (e.g.,
vllm)



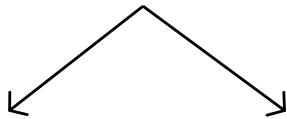
Celebrate

?

Is confidential AI this simple? Not quite!



Performance tradeoffs



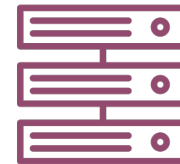
Model loading



GPU-GPU transfer

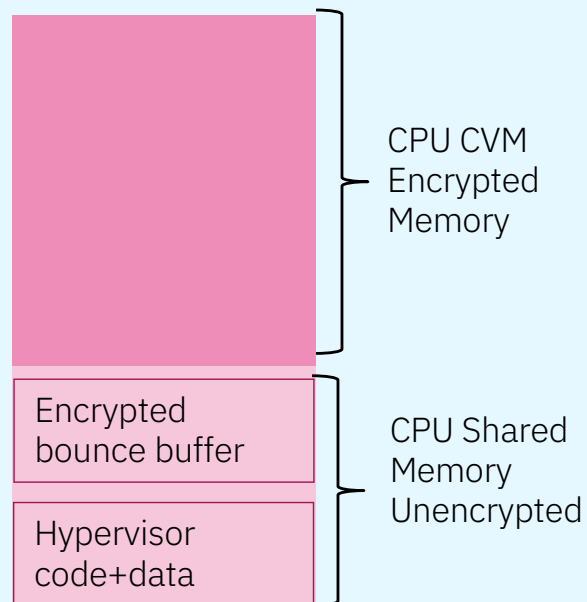


Attestation



Distributed trust

Model loading - 1



Bounce buffers

- NVIDIA Driver allocates bounce buffers in the Shared Memory area and encrypts data in those buffers with the session key

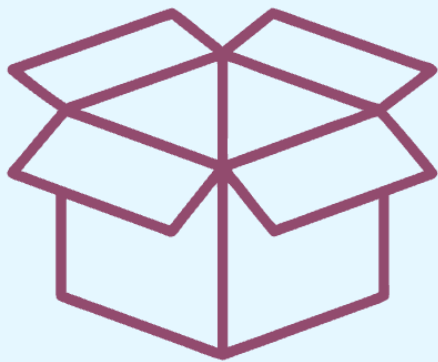
Host to device data transfers

- CPU encrypt by driver to bounce buffer in shared memory
- DMA engine reads from bounce buffer and decrypts to protected GPU memory

Device to host data transfers

- DMA engine encrypts and writes across PCIe to bounce buffer
- CPU decrypt by driver into confidential VM memory

Model loading - 2



Bounce buffers

- CPU and GPU in different trust zone
- Disk to GPU load time higher than no-CC

Serving multiple models with frequent loads/unloads?

- Performance impact

TDISP (TEE Device Interface Security Protocol) to the rescue?

- AMD SEV-TIO
- Intel TDX Connect

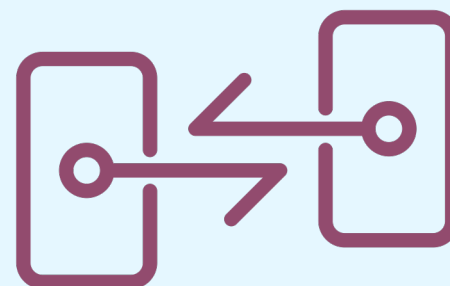
GPU-GPU transfer

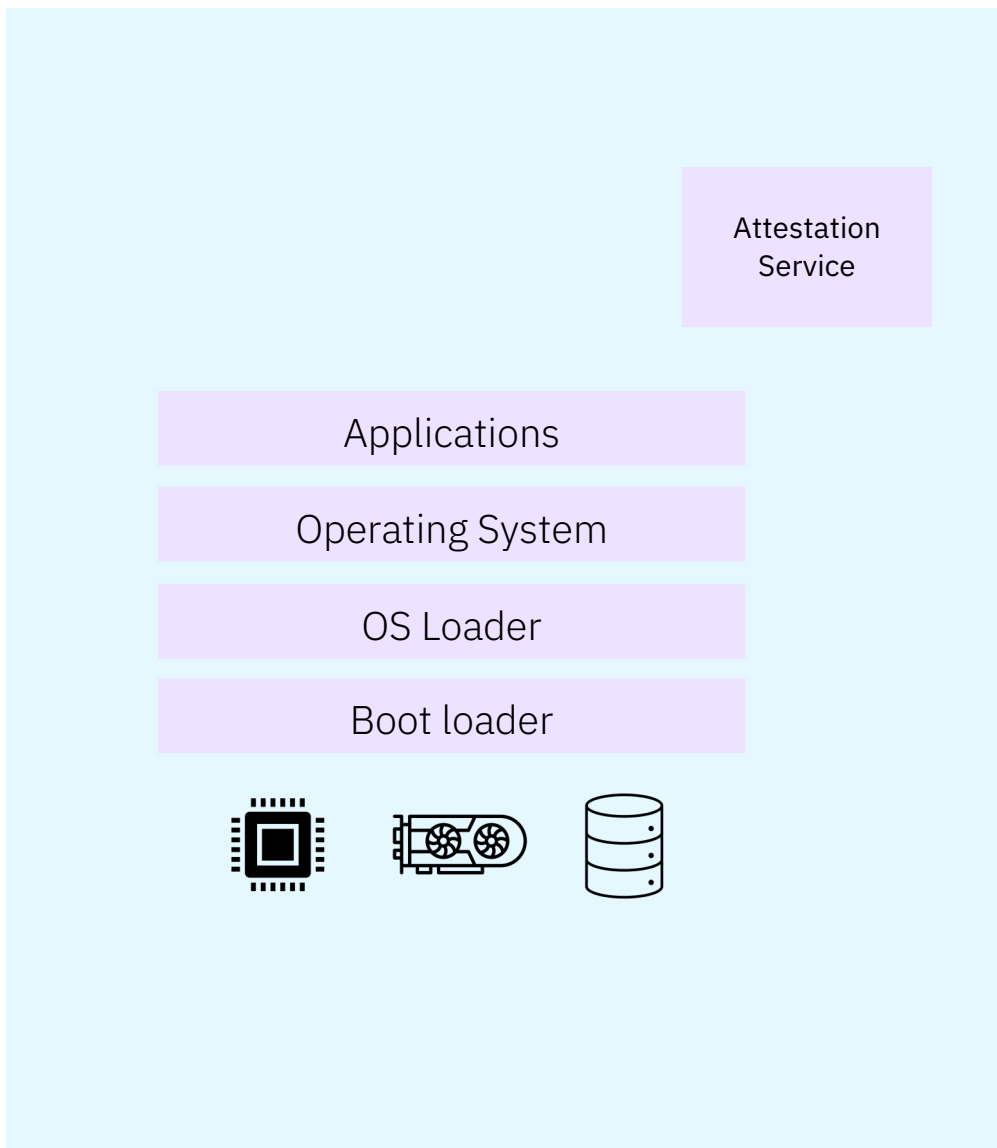
Trust model for protected PCIE (Hopper)

- No NV Link encryption
- Attestation needs to measure both GPUs and switches

Interconnect architecture

- NVLink encryption (Blackwell)
- TEE-I/O-capable





Secure boot and Attestation

Confidential data should be moved to the VM only after attestation

- Establish chain of trust -> lower layer component verified before executing next level
- Attestation by verifying trustworthiness of component/state to a remote verifier
- Full stack: hardware to applications
- Encrypted volumes

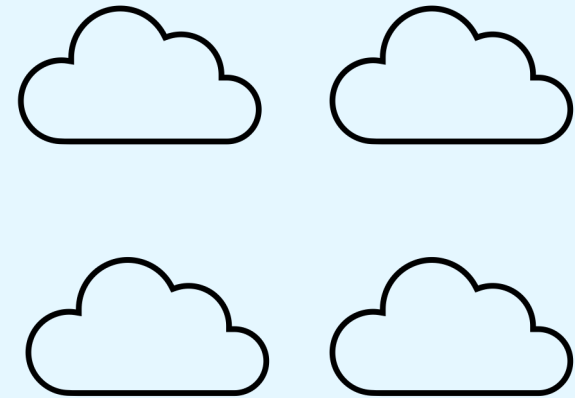
Distributed Trust

- AI applications not single node
- Compute requirements are vast
- CC resources are still scarce (beginning to change)

Needs:

- Flexibility
- Scale
e.g., attestation cross cluster and cloud
- CC aware orchestration

=> Manage using K8



Is confidential AI this simple? Summary



Secure boot & Attestation

- Confidential data should be moved to the VM only after attestation
- E.g., disk encryption keys for volume mounts



Model loading

- Higher load time (bounce buffers)
- Impact on services serving multiple models



Other performance tradeoffs

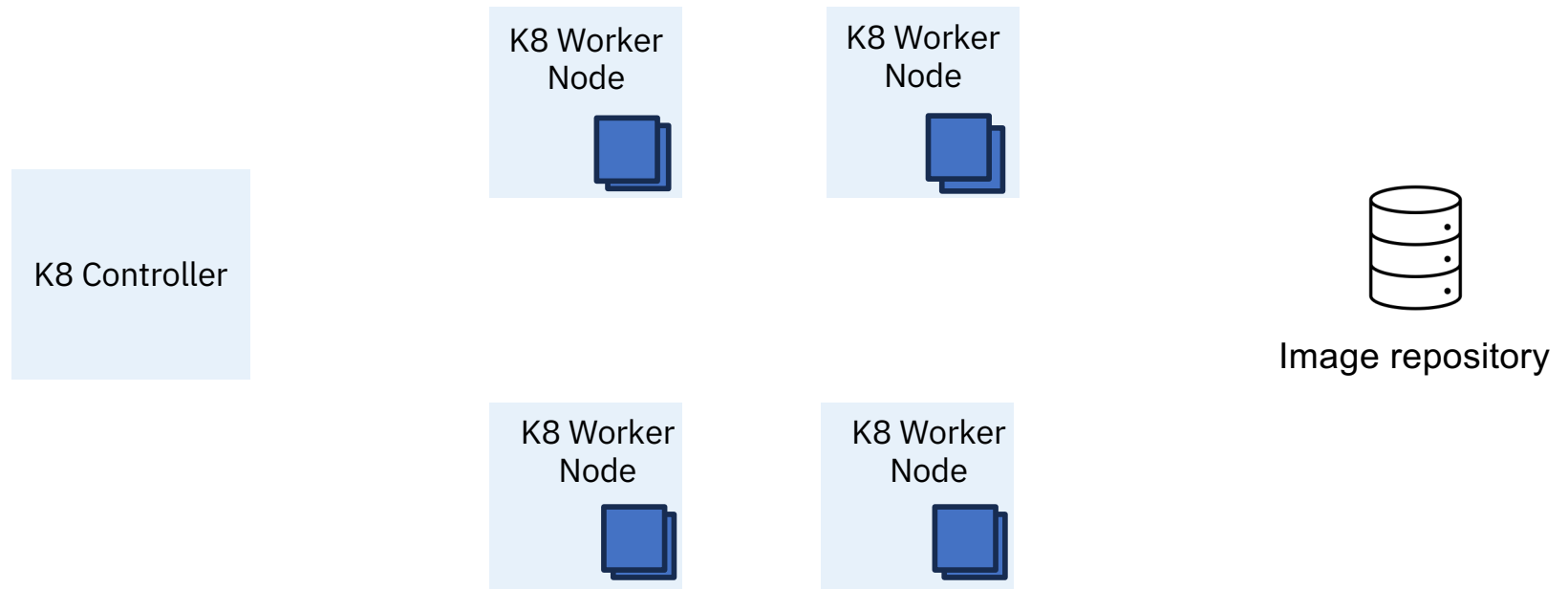
- GPU-GPU interconnect speed
- Deploying on K8, CoCo



Distributed Trust

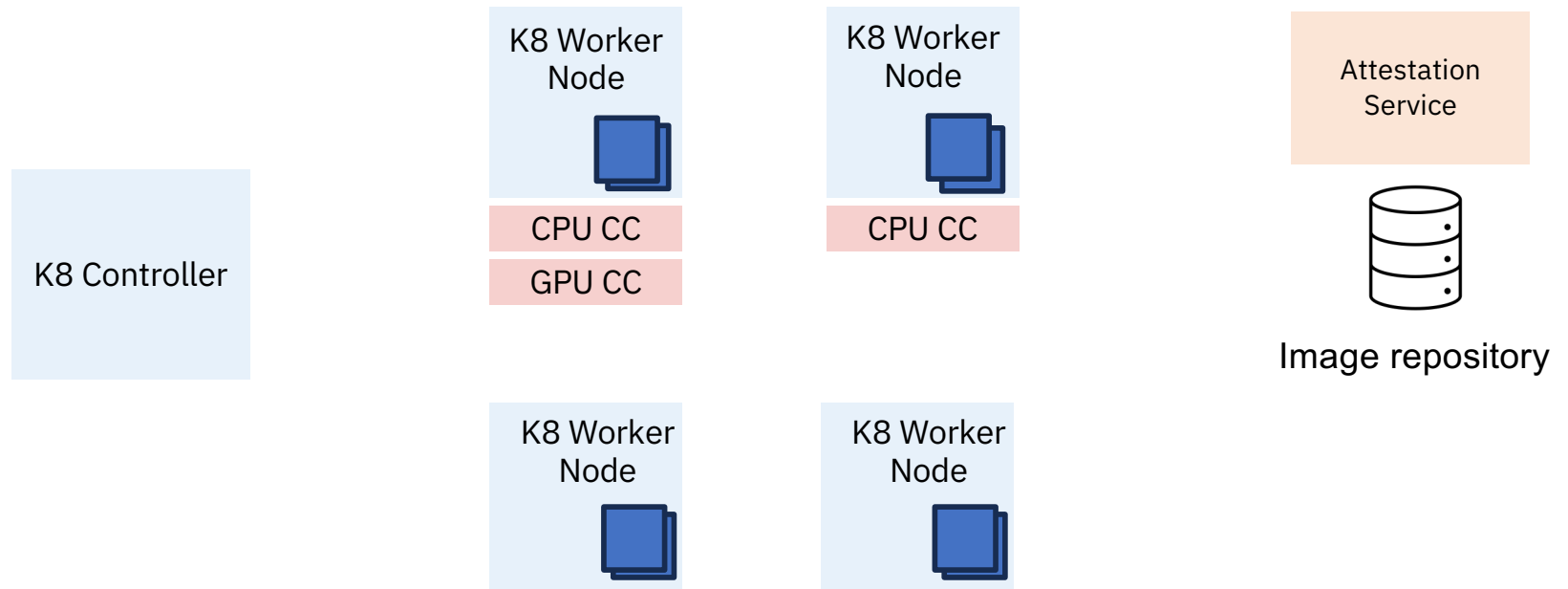
- Cluster wide attestation
- CC aware distributed platform

Microservices on the cloud



Microservices on the cloud

CoCo, K8 w/ Kata w/ CPU CC and GPU CC



Deployment policy: sample

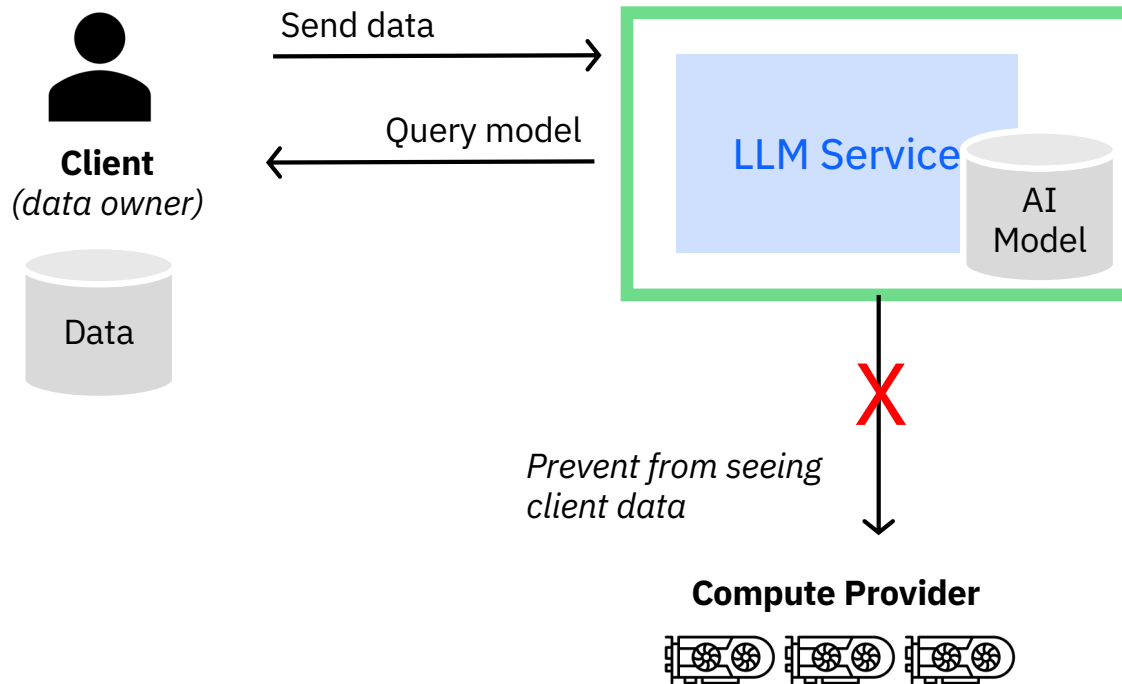
```
kind: k8s
k8s_config:
  config_file: ~/.kube/config
  replicas: 1
Agents:
- name: HR Agent
  image_repo: <repo name>
- name: BoxTool
  runtime:
    env_from_file: .env
  image_repo: <repo name>
  policies:
    - name: security-rag
      provider: [watsonx, Sanitizer]
      milvus_uri: https://some-uri.com/milvus
      CC: true
      env_from_file: .env
- name: BoxTool
  image_repo: <repo name>
  runtime:
    name: security-SPI
    env_from_file: .env
  policies:
    - name: security-SPI
      provider: Sanitizer
      CC: true
- name: Sanitizer
  policies:
    - name: security-no-network
    - name: security-exec-python-only
    - name: security-no-write-files
    CC: true
```



K8 Deployment yaml files

Growing **distrust** by everyone...

Client and model owner doesn't trust provider



Bootstrap confidential LLM inference

1 init

- GPU-CC Attestation
- CPU-CC Attestation
- Mount Encrypted Disk

2 download

- Download encrypted model

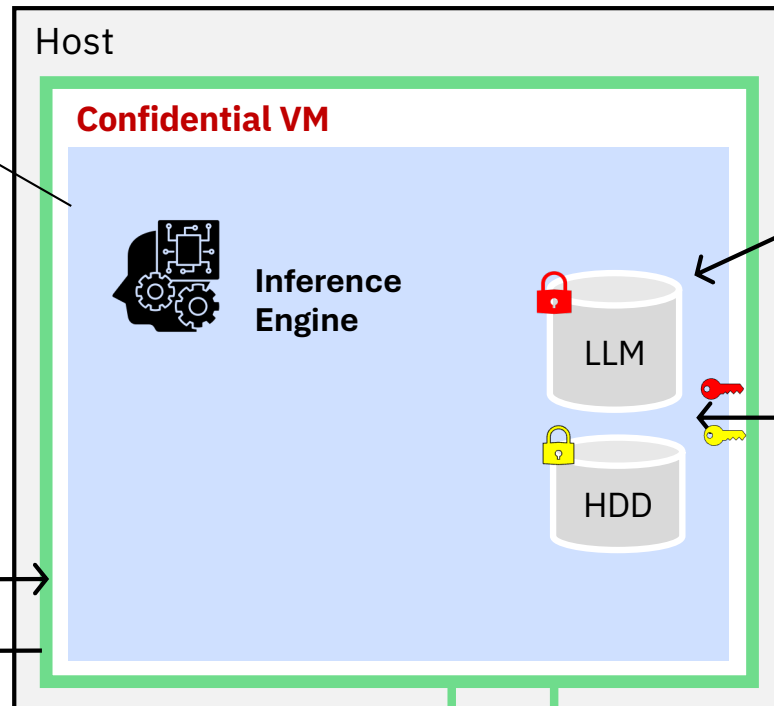
3 serve

- Decrypt model
- Establish inference service

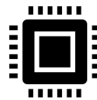

Client
(data owner)

Send data

Query model



AMD SEV-SNP



NVIDIA H100



Hugging Face

Trustee



Disk Encryption Key



Model Encryption Key



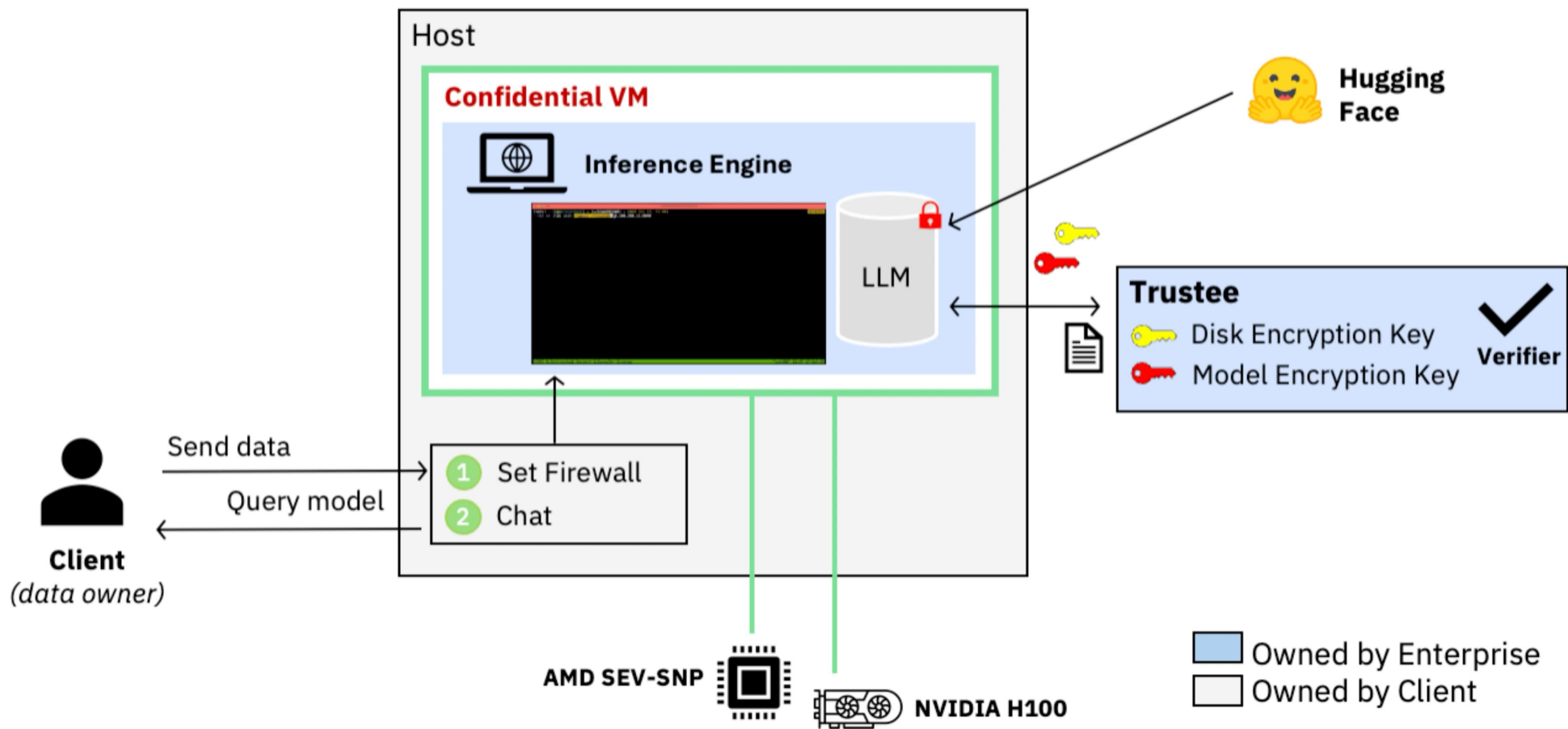
Verifier



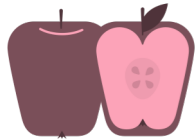
Owned by Client



Owned by Cloud Provider



Confidential computing alone insufficient



Split K8 Control Plane

- Control plane owned by cloud provider
- Split control plane API to be owner and host owned ¹



Scalable attestation and verification services

- High frequency of agents, tools, and services coming and going
- Multiagent systems



VM's TCB too big?

- Sandbox service to enforce security invariant, e.g., reduce interface, data tainting, network firewall

¹Valdez et al.: *Crossing Shifted Moats: Replacing Old Bridges with New Tunnels to Confidential Containers*

New DevOps Tooling: Ensure confidentiality is enforced

- ✓ Confidential VM memory is not accessible
- ✓ GPU memory is not accessible
- ✓ Software in CVM and GPU is attested
- ✓ AI model is downloaded into an encrypted disk
- ✓ AI model is encrypted
- ✓ Communications between CVM and Trustee/KBS are secured
- ✓ Disk passphrase and model decryption key owned by appropriate owners

Summary

- AI need sensitive data
- CC crucial to delivering real world confidential AI
- K8 + CoCo + GPU-CC as confidential AI platform
- Confidential AI solutions are bespoke
- Better user/developer experience
 - CC GPU still growing
 - More vendor support for GPU-CC



Questions?



Michael Le



Julian Stephen

Thank You!

Audience take-away

- One size does not fit all. Trust model is key
- OSS components to build Confidential AI stack
- Performance considerations