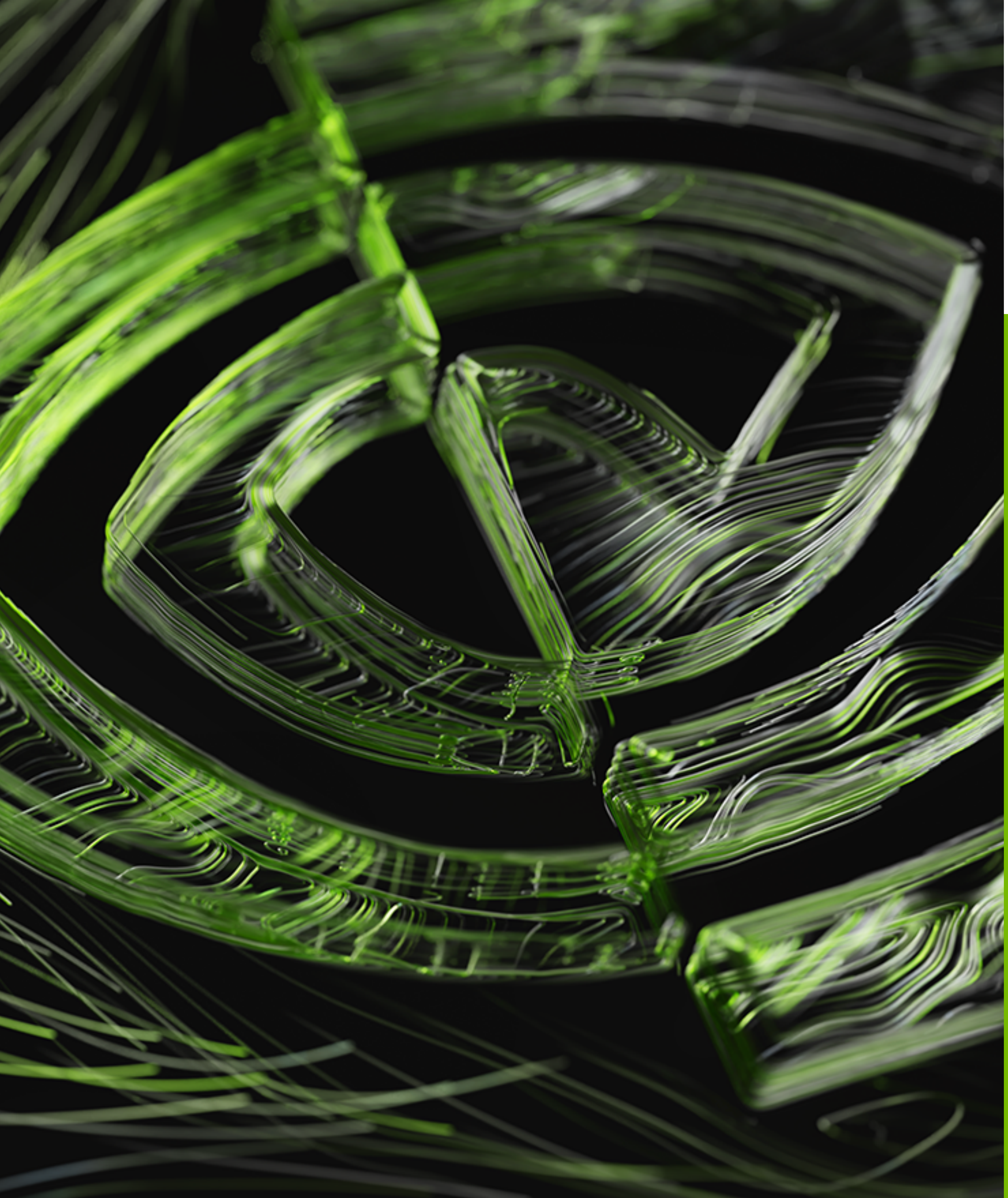
The Developer's View to Secure an Application and Data on NVIDIA H100 with Confidential Computing

GTC Spring 2023

Rob Nertney CUDA Technical Product Manager



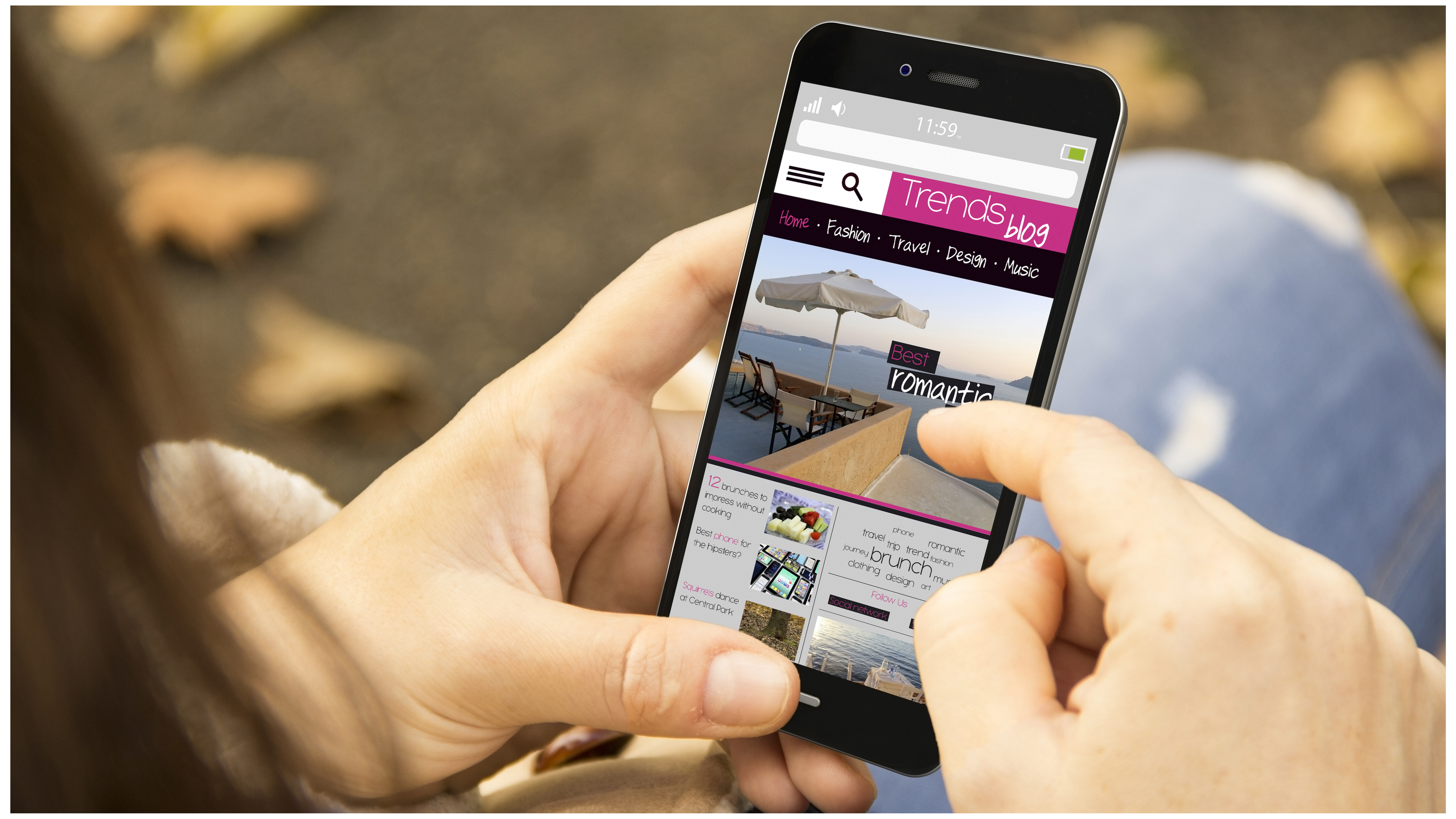
Agenda

- Why Confidential Computing Matters
- How Confidential Computing Protects Data In-Use
- Expanding Trusted Execution Environments to Hopper H100
- CUDA Coding Considerations

AI Driving Transformation Across Industries

Driving Deeper Insights I Accelerating Product Innovation I Powering Greater Efficiencies

Financial Services
Banking I Insurance I Payments



Consumer Internet
Ecommerce I Social Media I Video Conferencing

Healthcare and Life Sciences
Medical Devices I Smart Hospitals I Genomics

96% of organizations have AI initiatives in pilot[1]

92% of organizations are increasing their AI investments[1]

67% see increase in revenue[2]

79% see significant cost savings[2]

[1]NewVantage Partners, "Data and AI Leadership Executive Survey", 2022; [2]McKinsey, "State of AI in 2021", 2021

Data Privacy and Security Can be a Barrier to Deriving Value from AI

Attacks Follow Value… and Data is Valuable

Sensitive I Regulated I Private
• Protected Health Information (PHI)
• Credit Card Holder Information
• Credit Reports & Credit Card Transactions
• Banking / Financial Transactions
• Enterprise Operational Data
• Intellectual Property (Including AI Models)

Growing Need for Regulatory Compliance
• HIPAA - Health Insurance Portability and Accountability Act
• PCI DSS - Payment Card Industry Data Security Standard
• GLBA – Gramm-Leach-Bliley Act
• GDPR - General Data Protection Regulation
• CCPA - California Consumer Privacy Act

45% of Data Breaches were Cloud-Based[1]

$4.35M Global Average Total Cost of a Data Breach[1]

$10.10M Average Cost of Breach in Healthcare[1]

$9.23M Average Cost of Breach in Finance[1]

[1]*Cost of a Data Breach 2022 Report, IBM : https://www.ibm.com/reports/data-breach*

Confidential Computing Overview

The Security Gap in Data Protection

Data at Rest	Data in Transit	Data in Use
SECURE	SECURE	NOT SECURE

CC Enables A Holistic Data Security Model

- Protect data at rest, in transit, and <u>in use</u>
- Protect data and code from other users
- Protect and code from machine owners

Industry Standards Engagement

CONFIDENTIAL COMPUTING CONSORTIUM

https://confidentialcomputing.io/

Benefits:

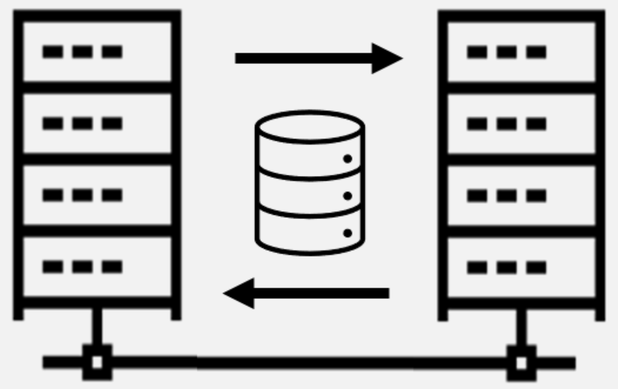
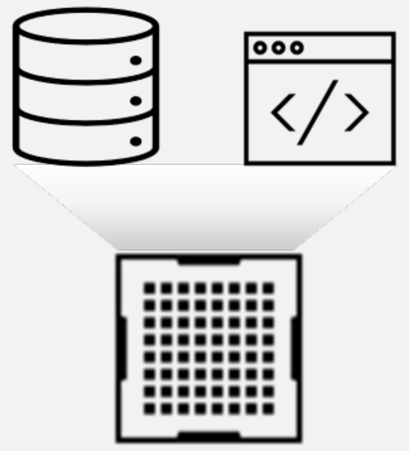
- Informed on industry direction
- Ability to influence specifications
- Alignment with customers

Other Members:

- Google
- AMD
- NVIDIA
- Microsoft
- Intel
- Meta

Confidential Computing Overview

The Security Gap in Data Protection

Data at Rest	Data in Transit	Data in Use
		
SECURE	SECURE	SECURE

CC Enables A Holistic Data Security Model

- Protect data at rest, in transit, and in use
- Protect data and code from other users
- Protect and code from machine owners

Industry Standards Engagement

CONFIDENTIAL COMPUTING CONSORTIUM

https://confidentialcomputing.io/

Benefits:

- Informed on industry direction
- Ability to influence specifications
- Alignment with customers

Other Members:

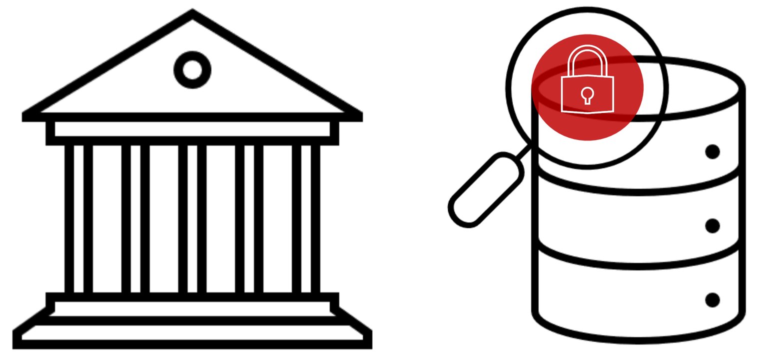
- Google
- AMD
- NVIDIA
- Microsoft
- Intel
- Meta

NVIDIA

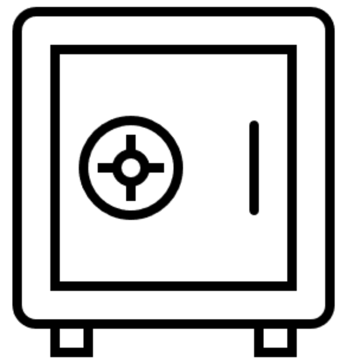
Key Market Uses

Data is Power

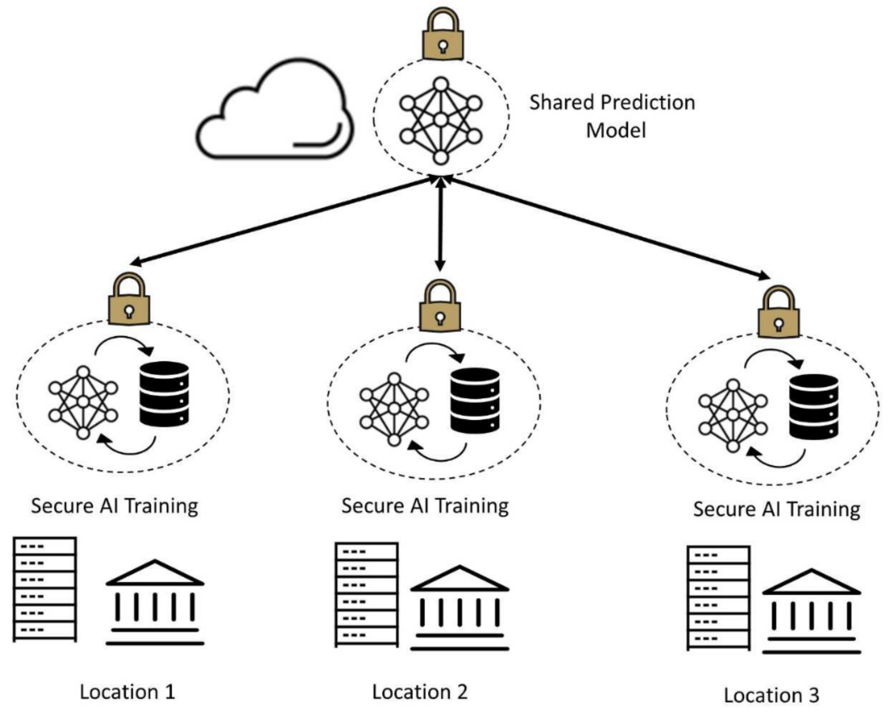
- Data Security and Privacy
  - Provides agility to quickly adhere to ever-evolving regulatory regimes requiring increasing levels of integrity and confidentiality
  - System owner-operators reduce their cost liability from potential bad-actor employees



- Digital Asset Protection
  - Hardware-based security features enable extension of deployment styles: edge devices can now include more sensitive data
  - ISVs or Model creators can require their solution work only on Confidential GPUs to prevent unauthorized access with end users



- Datacentric Trust Enables New Collaboration
  - Federated learning enables training of networks without providing access to sensitive datasets (e.g., healthcare data)
  - Internal groups –previously isolated from each other –within a single company may now be able to share data with each other without confidentiality requirements



NVIDIA Confidential Computing Goals

Protect data in use for accelerated computing

Run CUDA applications unchanged

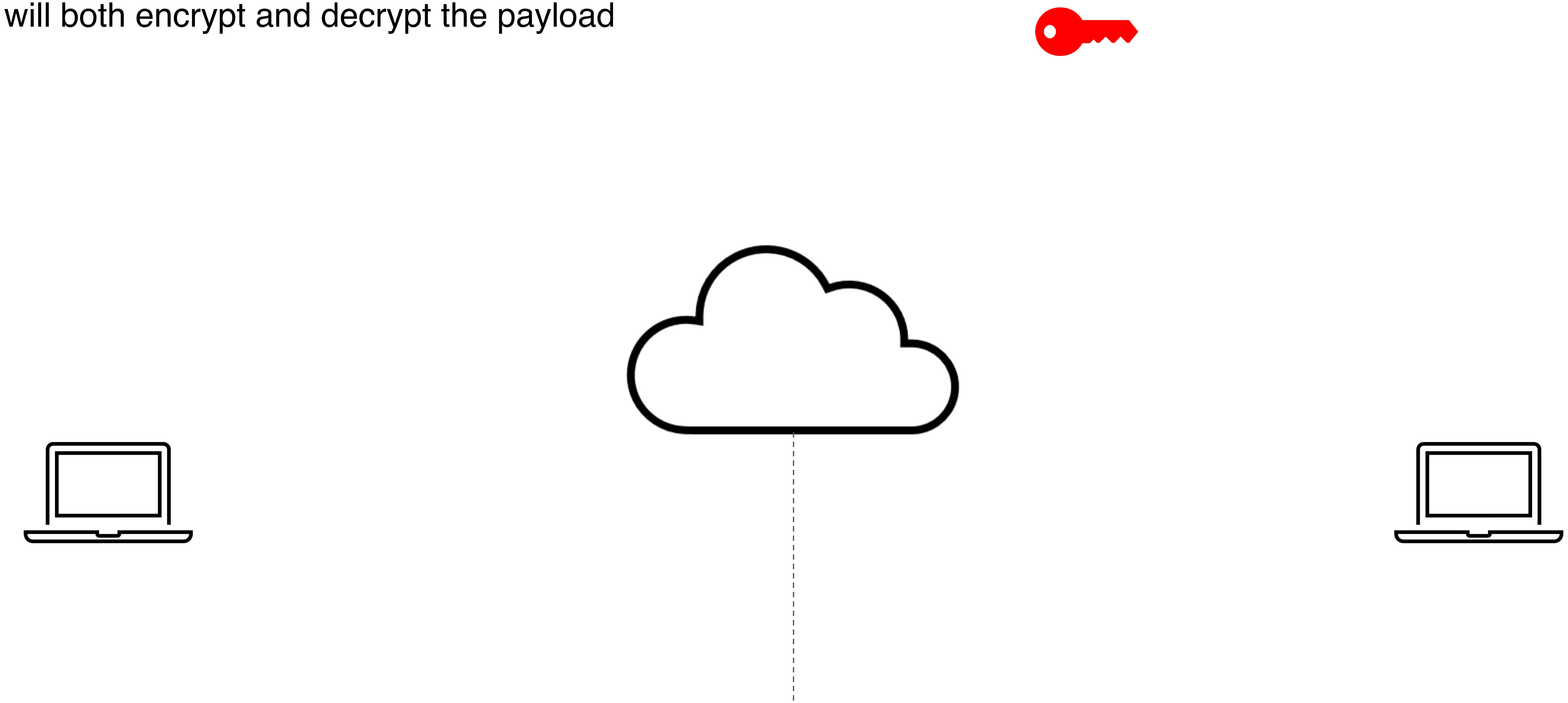
Offer scale from multi-instance GPUs to multi-node

A Brief Introduction to Encryption & Authentication

Key Based Encryption

Symmetric Keys

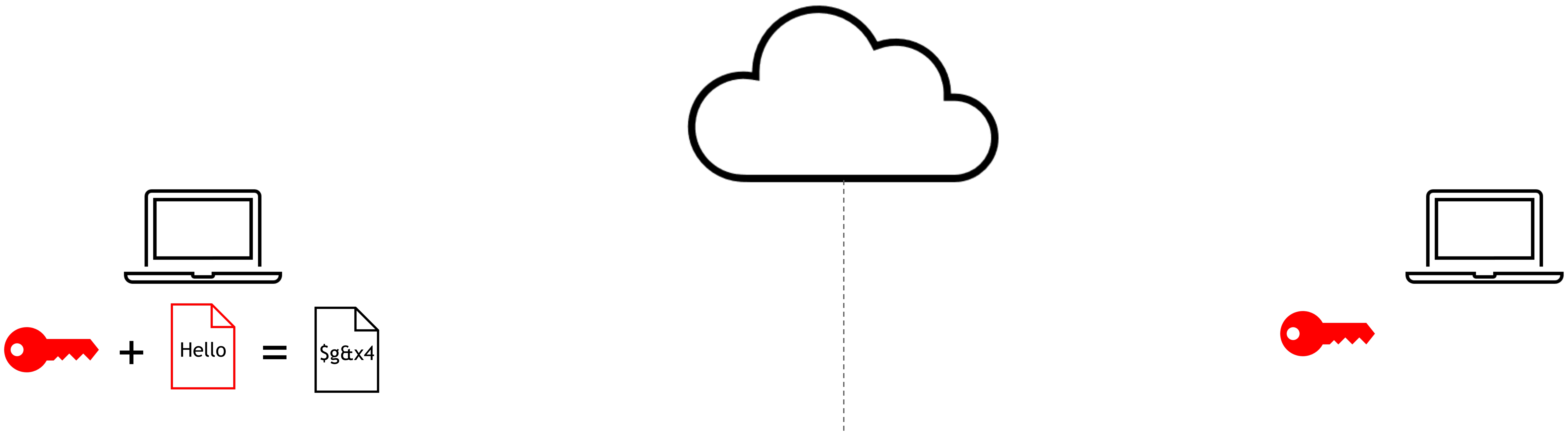
- Symmetric encryption utilizes a common password called a “Private Key”
- Both sender and receiver require the same key
- The key will both encrypt and decrypt the payload



Symmetric Key Encryption

E.g., AES

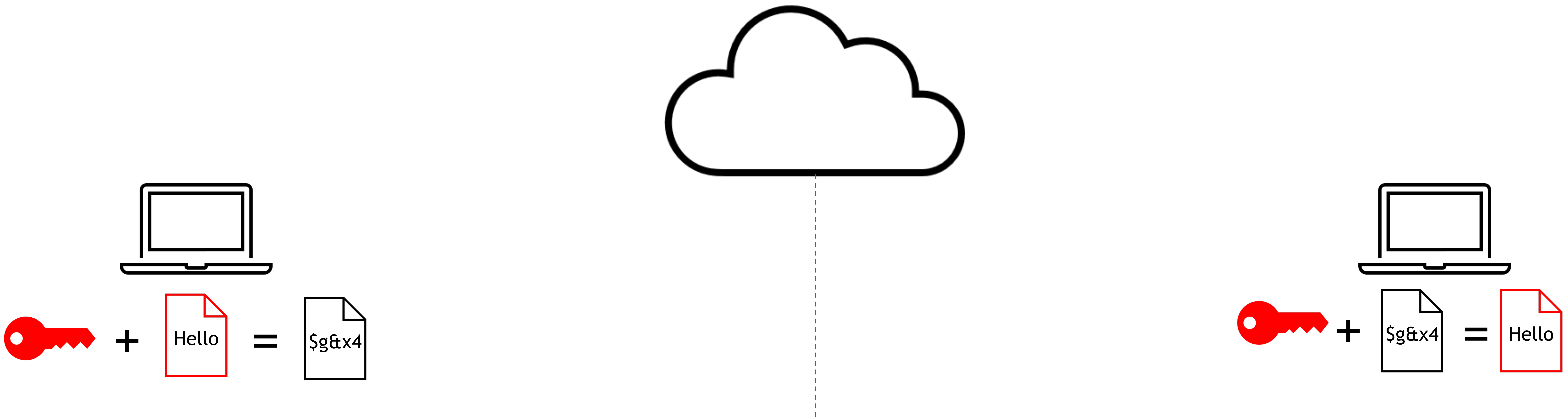
- Symmetric encryption utilizes a common password called a “Private Key”
- Both sender and receiver require the same key
- The key will both encrypt and decrypt the payload



Symmetric Key Encryption

E.g., AES

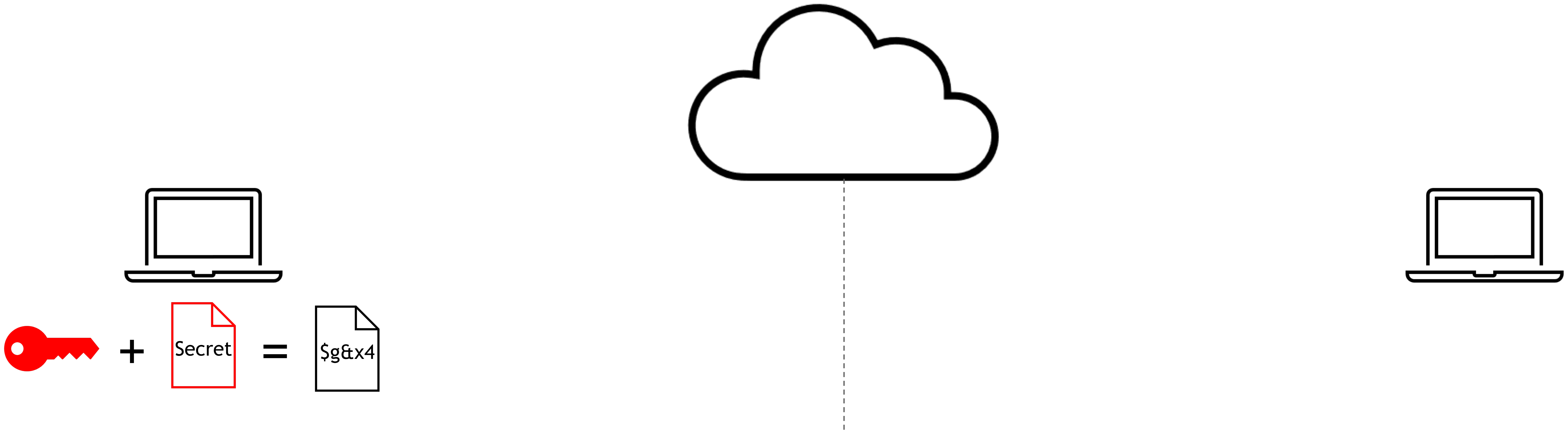
- Symmetric encryption utilizes a common password called a “Private Key”
- Both sender and receiver require the same key
- The key will both encrypt and decrypt the payload



Symmetric Key Encryption

E.g., AES

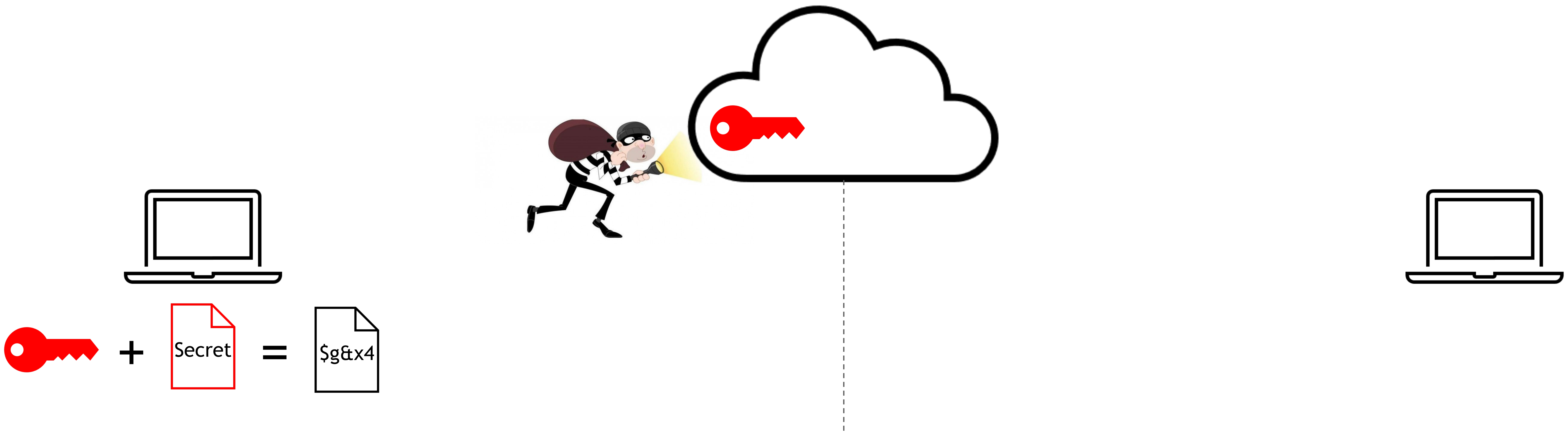
- The problem with symmetric keys is 2-fold:
  - How do I exchange the keys?
  - How do I know the receiving end is who I think it is?



Symmetric Key Encryption

E.g., AES

- The problem with symmetric keys is 2-fold:
  - How do I exchange the keys?
  - How do I know the receiving end is who I think it is?

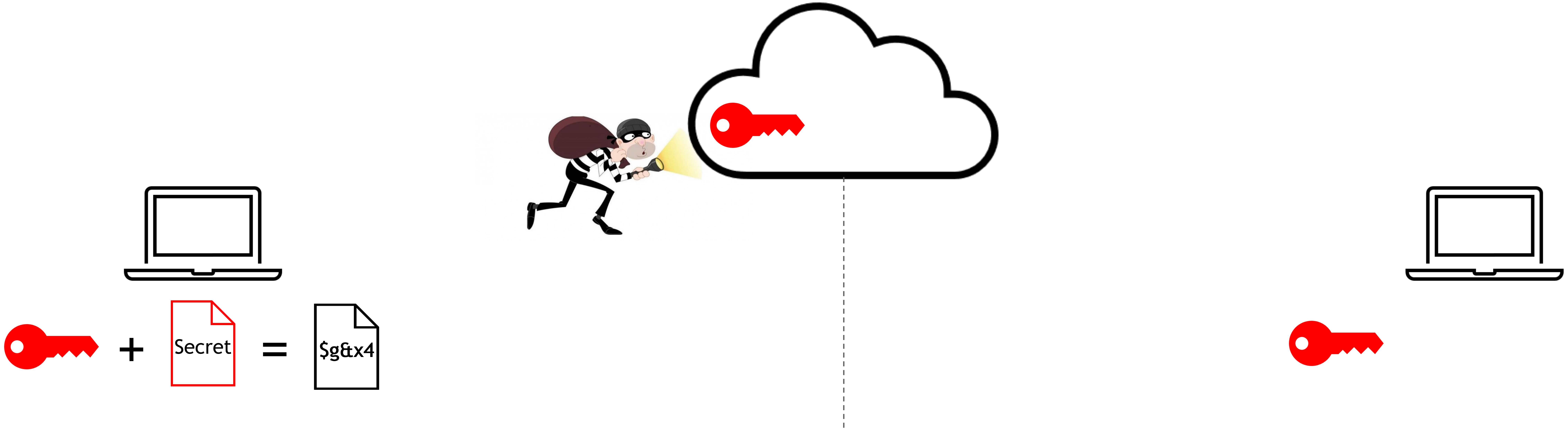


NVIDIA

Symmetric Key Encryption

E.g., AES

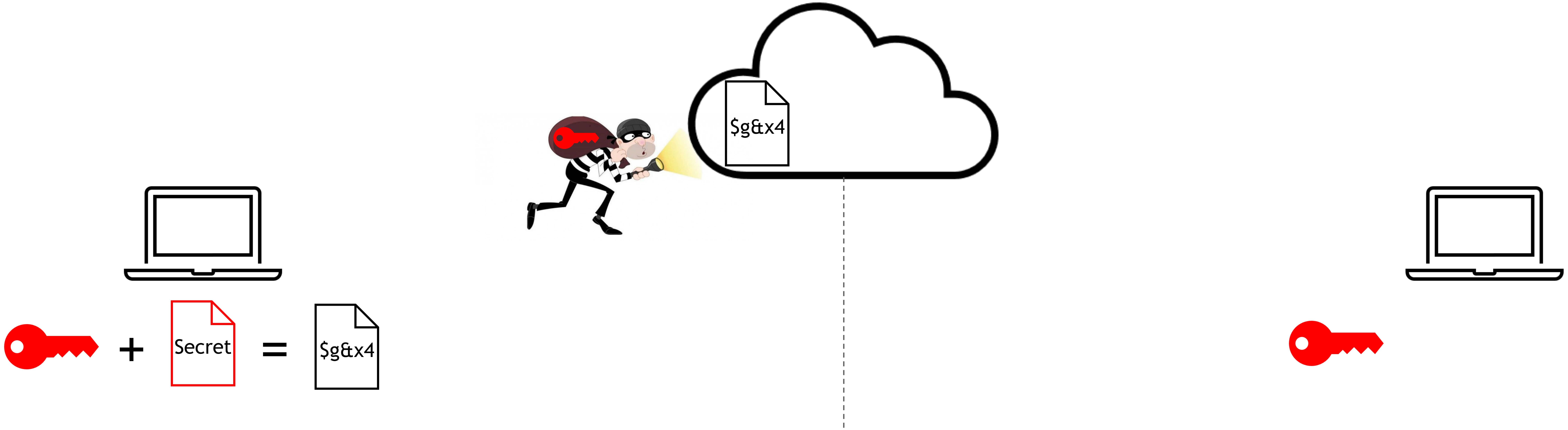
- The problem with symmetric keys is 2-fold:
  - How do I exchange the keys?
  - How do I know the receiving end is who I think it is?



Symmetric Key Encryption

E.g., AES

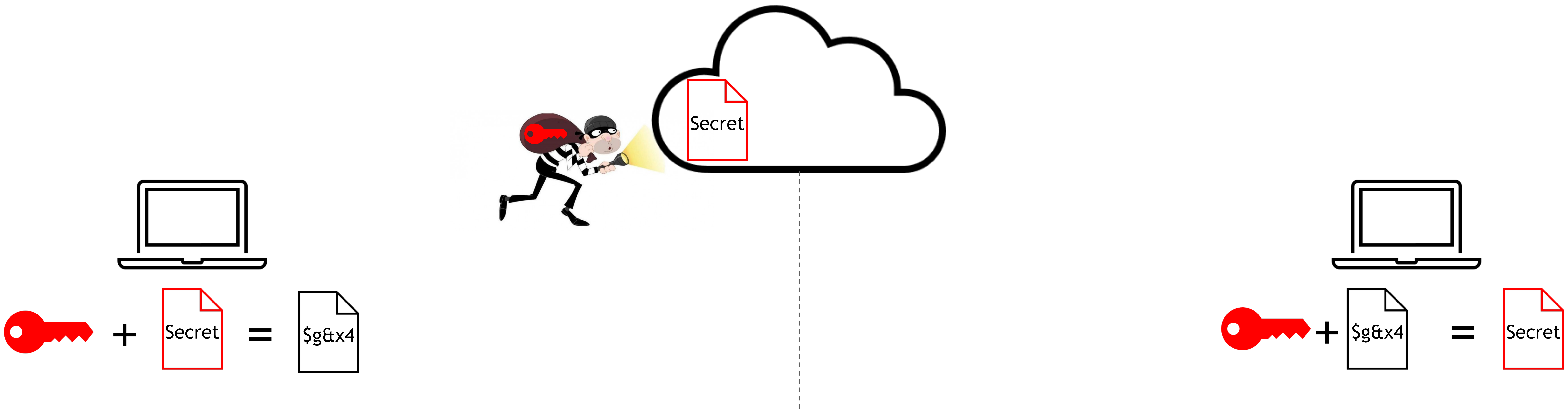
- The problem with symmetric keys is 2-fold:
  - How do I exchange the keys?
  - How do I know the receiving end is who I think it is?



Symmetric Key Encryption

E.g., AES

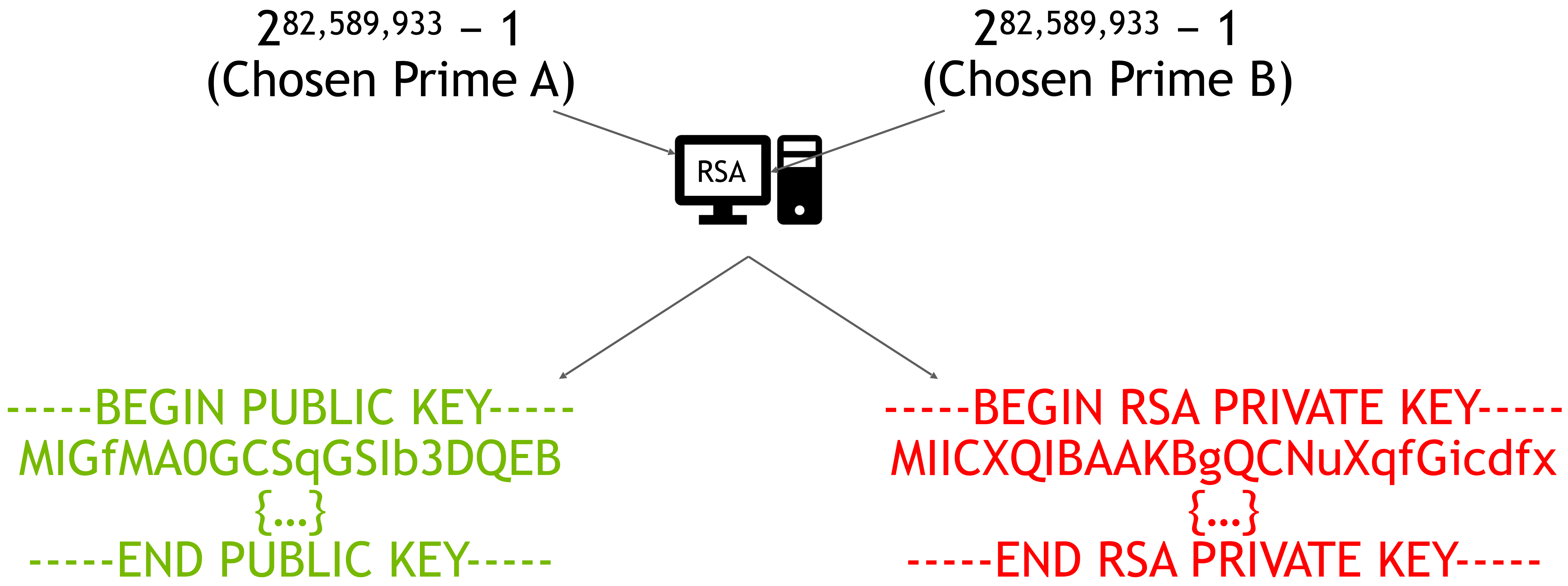
- The problem with symmetric keys is 2-fold:
  - How do I securely exchange the keys?
  - How do I know the receiving end is who I think it is?



Enter: Asymmetric Keys

A Public/Private Keypair

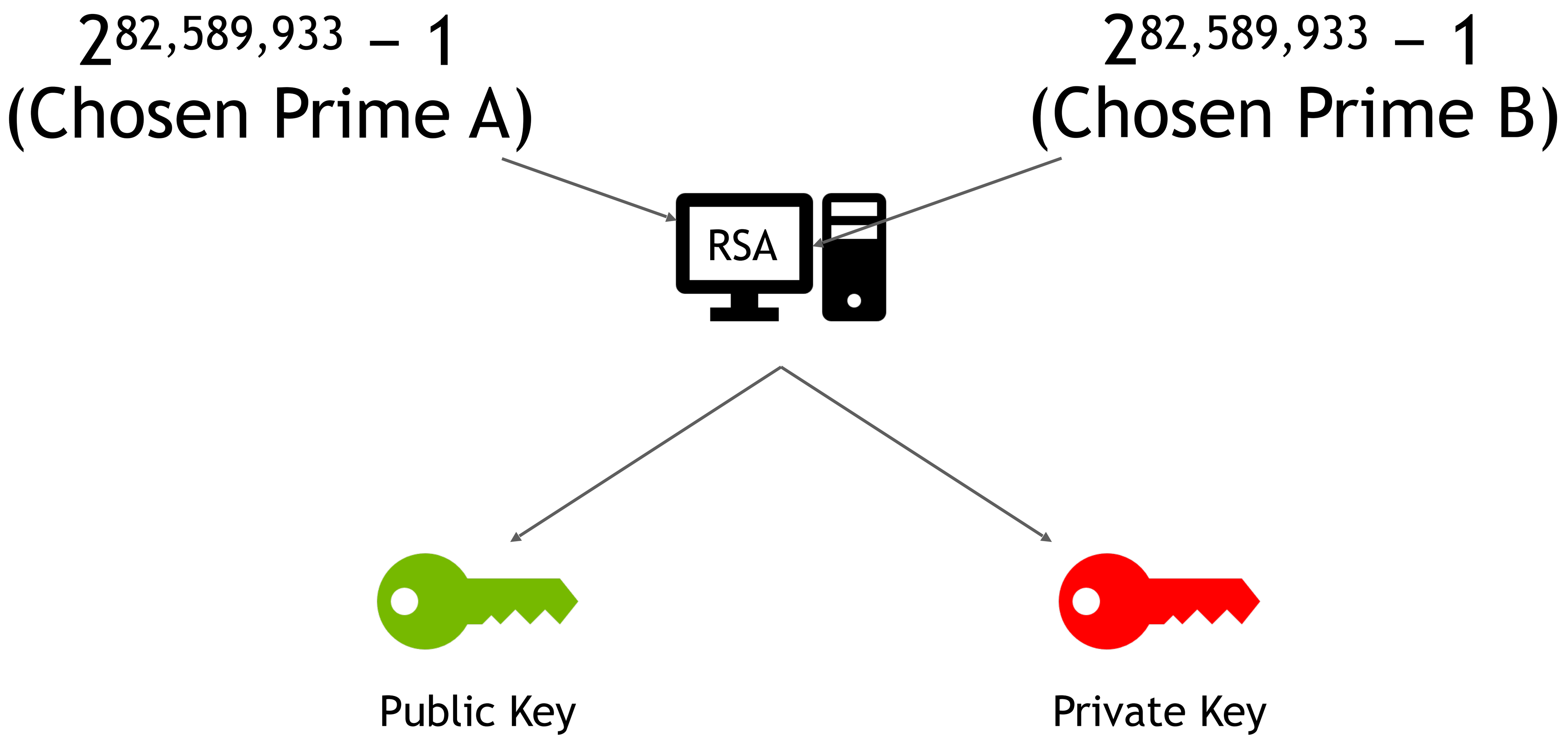
- Asymmetric Keys are created from a complex mathematical operation involving very large prime numbers
- The keys are relationally bound, however, (pragmatically) unable to directly derive one from the other
    - The keys are often called “Public” and “Private”
- Asymmetric keys can be used both for encryption and ‘authentication’, or verifying the integrity of a payload



Enter: Asymmetric Keys

A Public/Private Keypair

- Asymmetric Keys are created from a complex mathematical operation involving very large prime numbers
- The keys are relationally bound, however, (pragmatically) unable to directly derive one from the other
    - The keys are often called “Public” and “Private”
- Asymmetric keys can be used both for encryption and ‘authentication’, or verifying the integrity of a payload

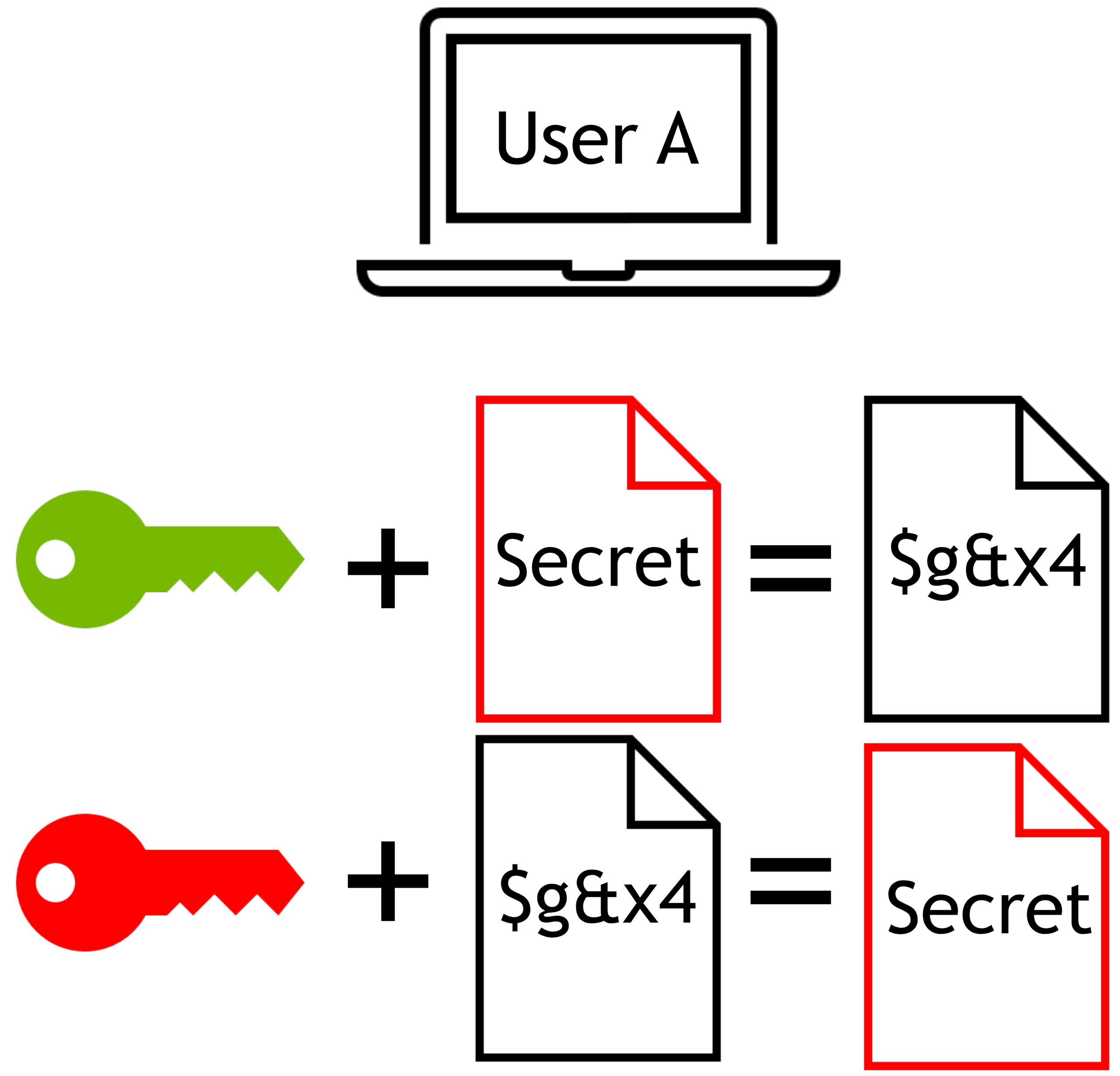


Encryption vs. Authentication

Encryption Keeps Your Information Secure

Encryption

- The Public key is used to encrypt data
- Only the Private key can be used to decrypt it

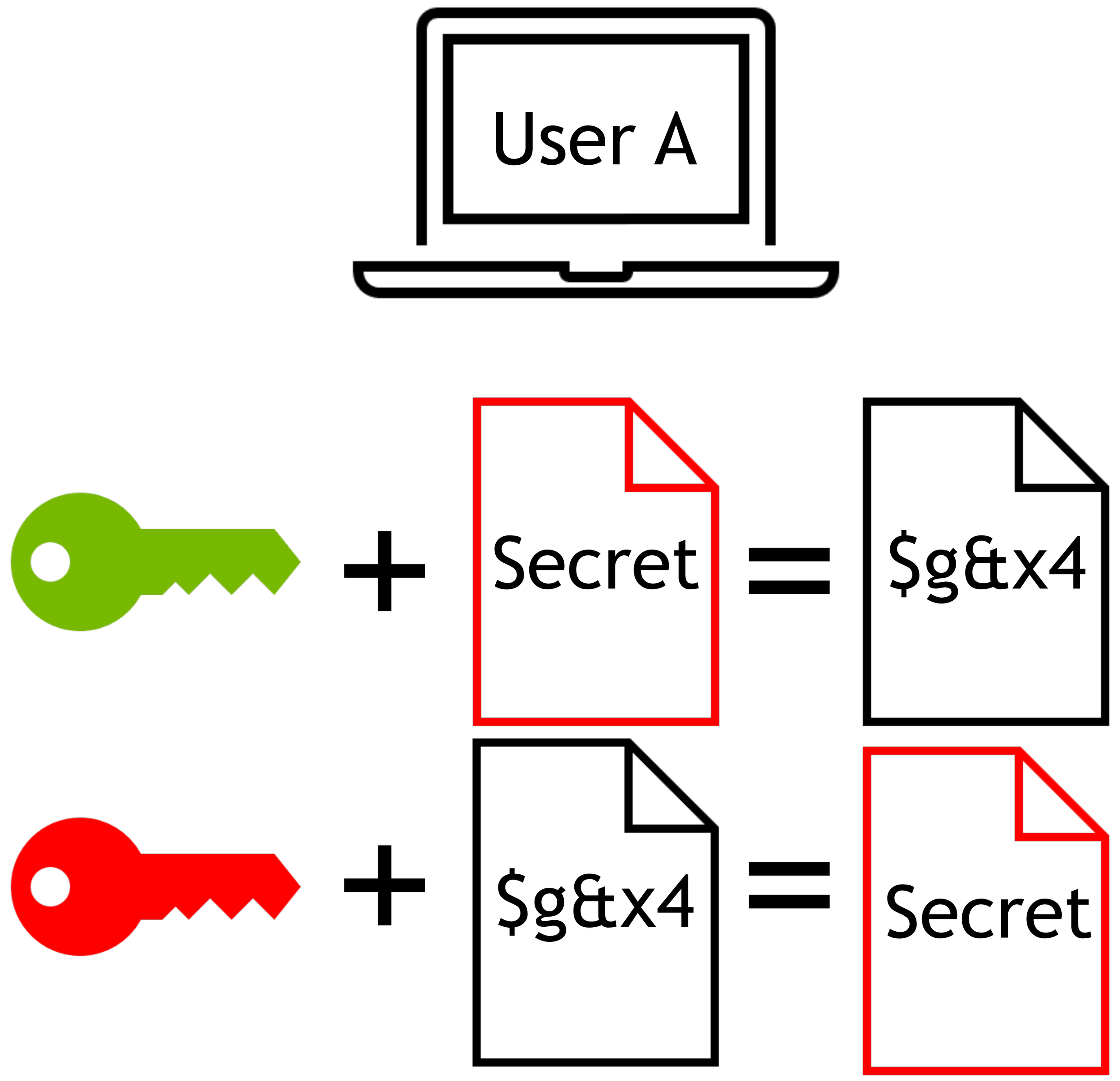


Encryption vs. Authentication

Authentication Keeps Your Information Accurate

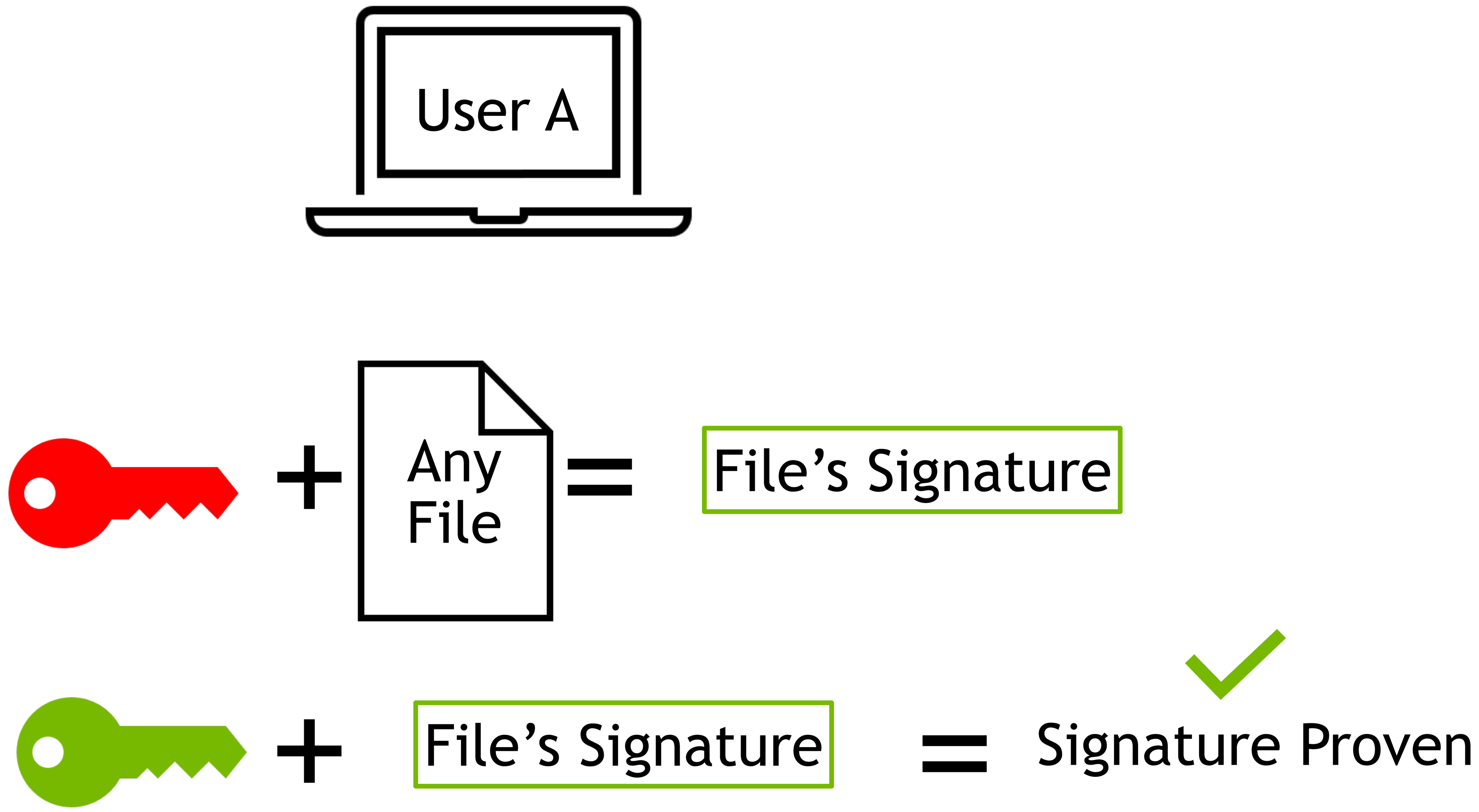
Encryption

- The Public key is used to encrypt data
- Only the Private key can be used to decrypt it



Authentication

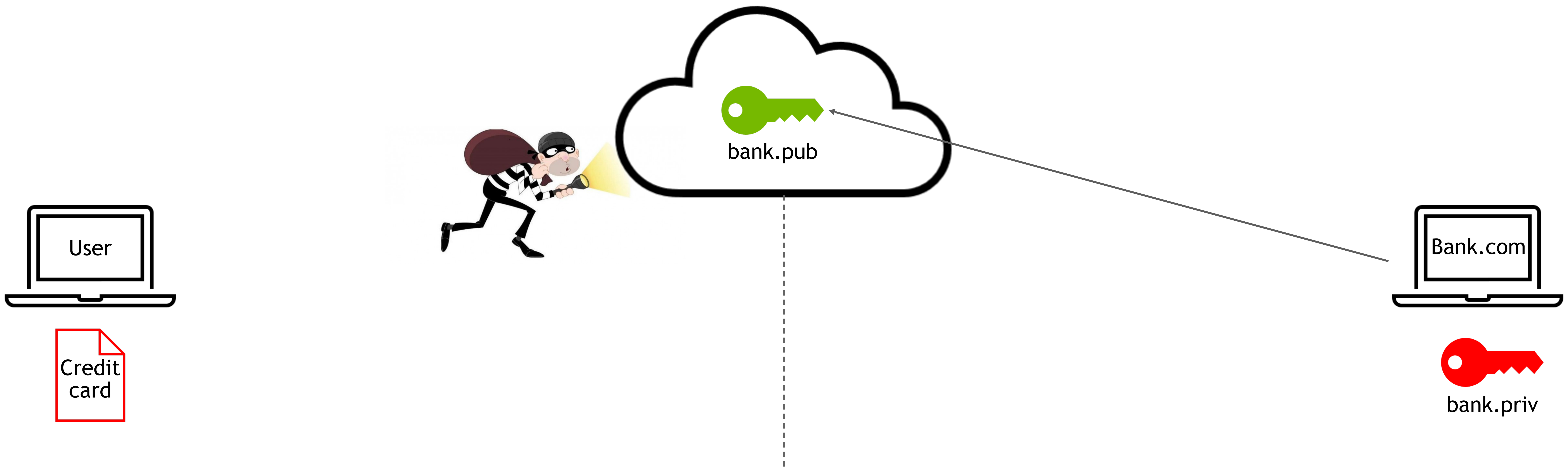
- The Private Key is used to create a “signature”
- Only the Public Key can be used to verify it is accurate



Asymmetric Key Encryption

How It Helps Supplement Security

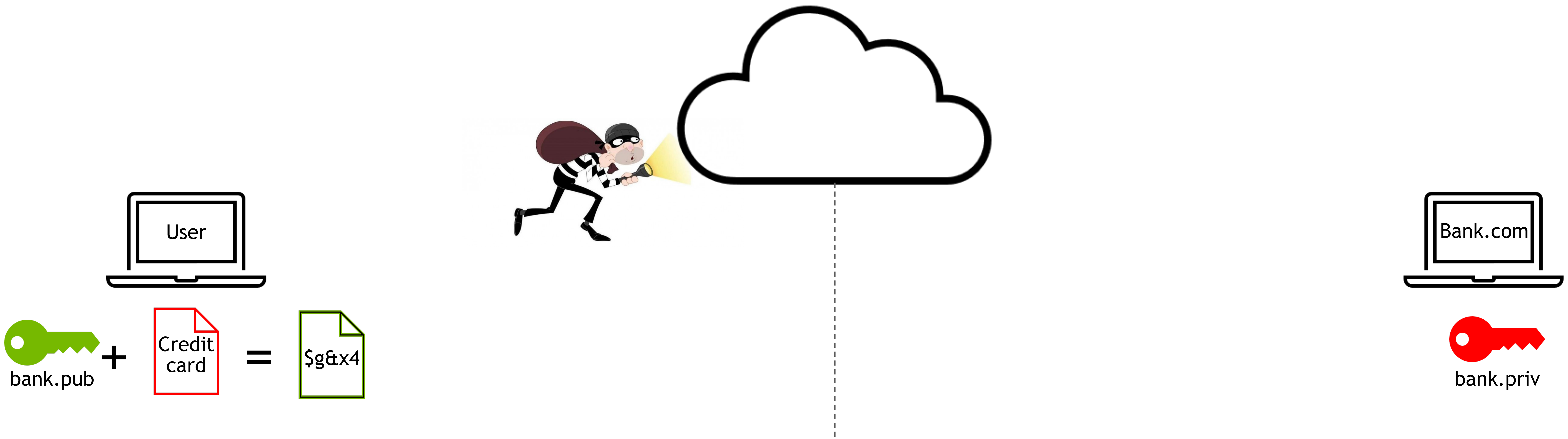
- Public keys can be published online, as they can only be used to encrypt data, not decrypt



Asymmetric Key Encryption

How It Helps Supplement Security

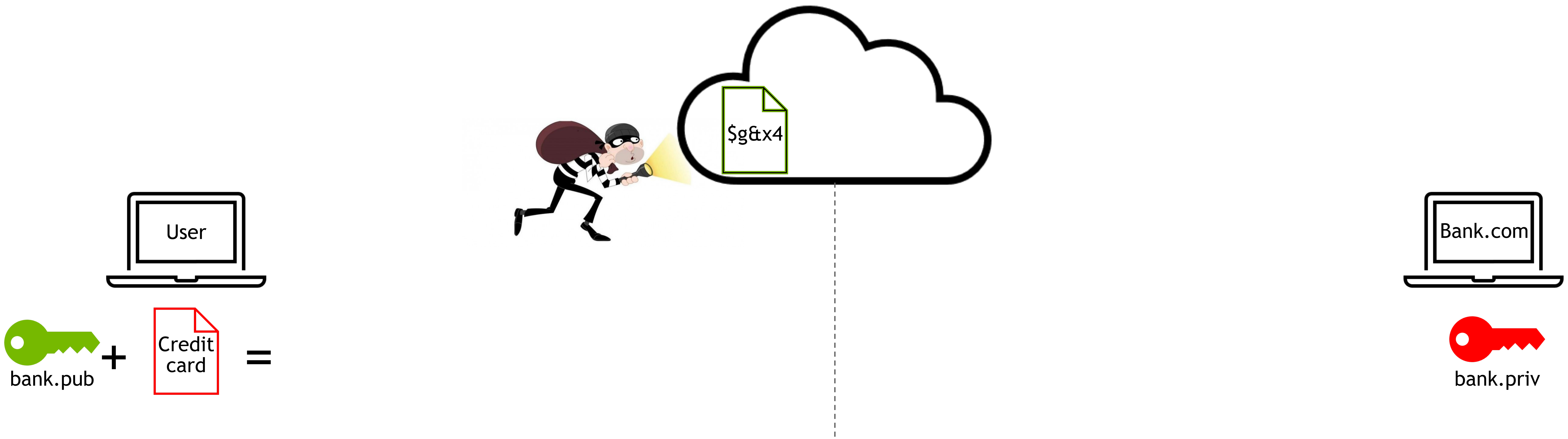
- Public keys can be published online, as they can only be used to encrypt data, not decrypt
- You use the published public key to encrypt your data



Asymmetric Key Encryption

How It Helps Supplement Security

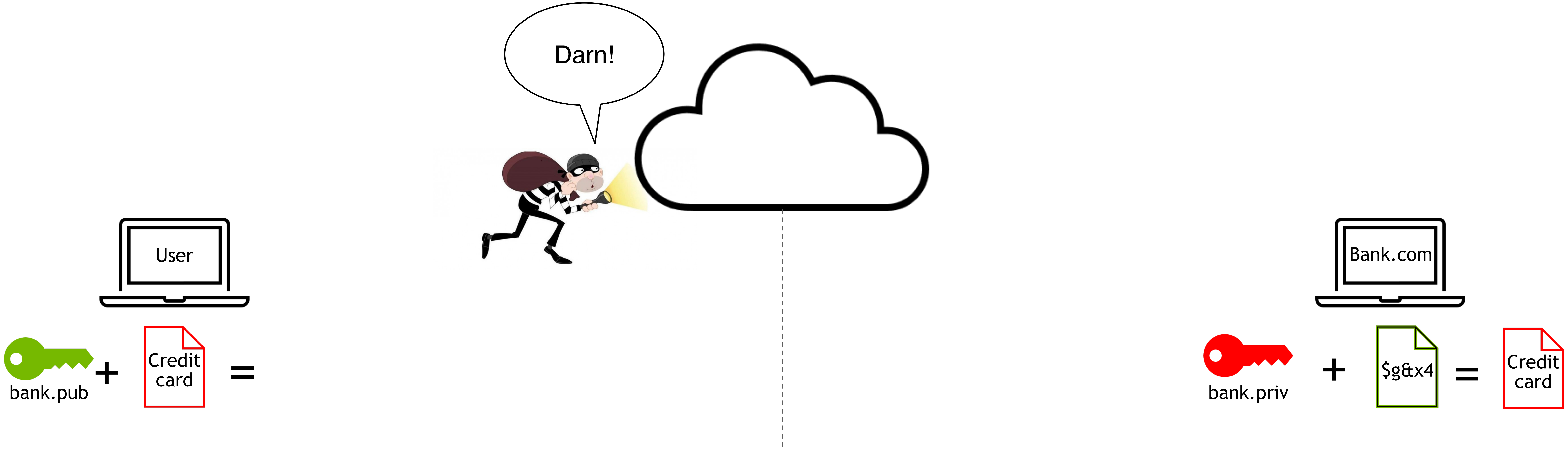
- Public keys can be published online, as they can only be used to encrypt data, not decrypt
- You use the published public key to encrypt your data



Asymmetric Key Encryption

How It Helps Supplement Security

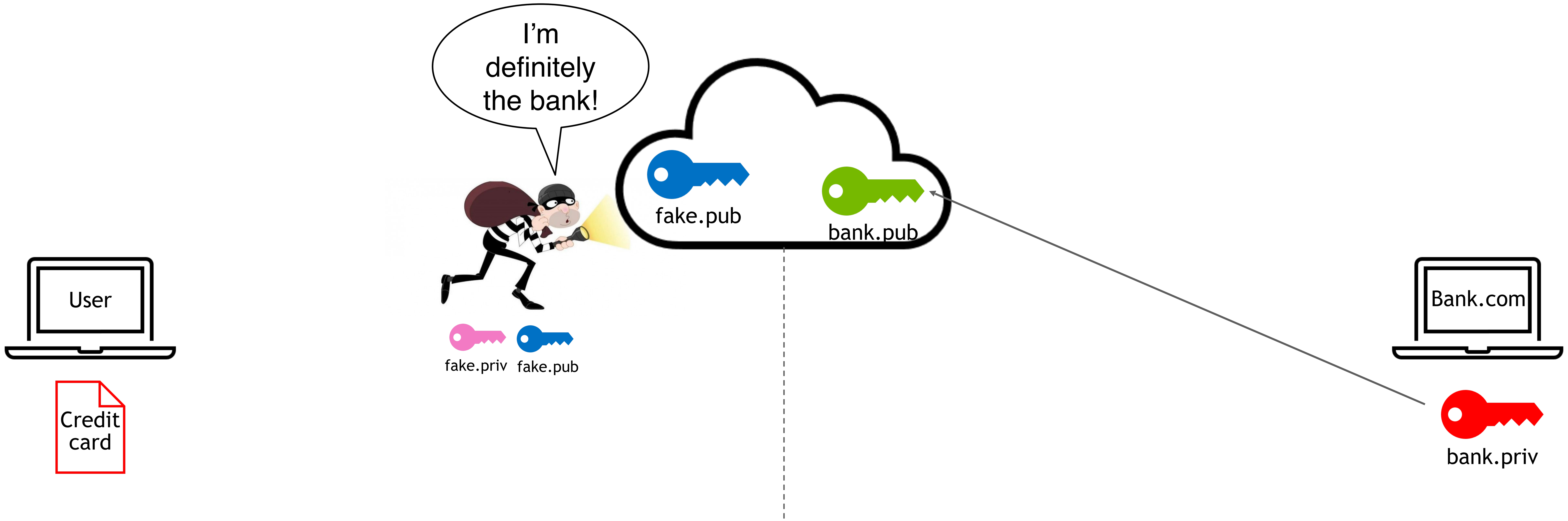
- Public keys can be published online, as they can only be used to encrypt data, not decrypt
- You use the published public key to encrypt your data
- Only the owner of the private key can decrypt it



Attestation

Public Keys Aren't Enough to Prove Identity

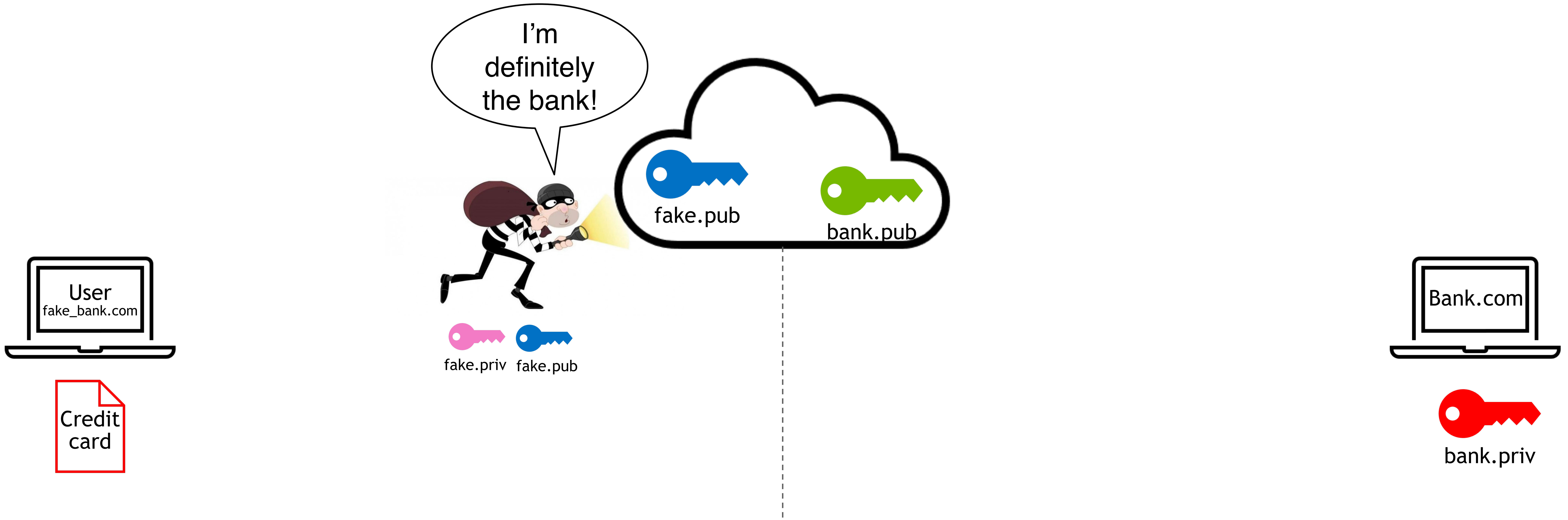
- What if a bad-actor acts as an impostor?



Attestation

Public Keys Aren't Enough to Prove Identity

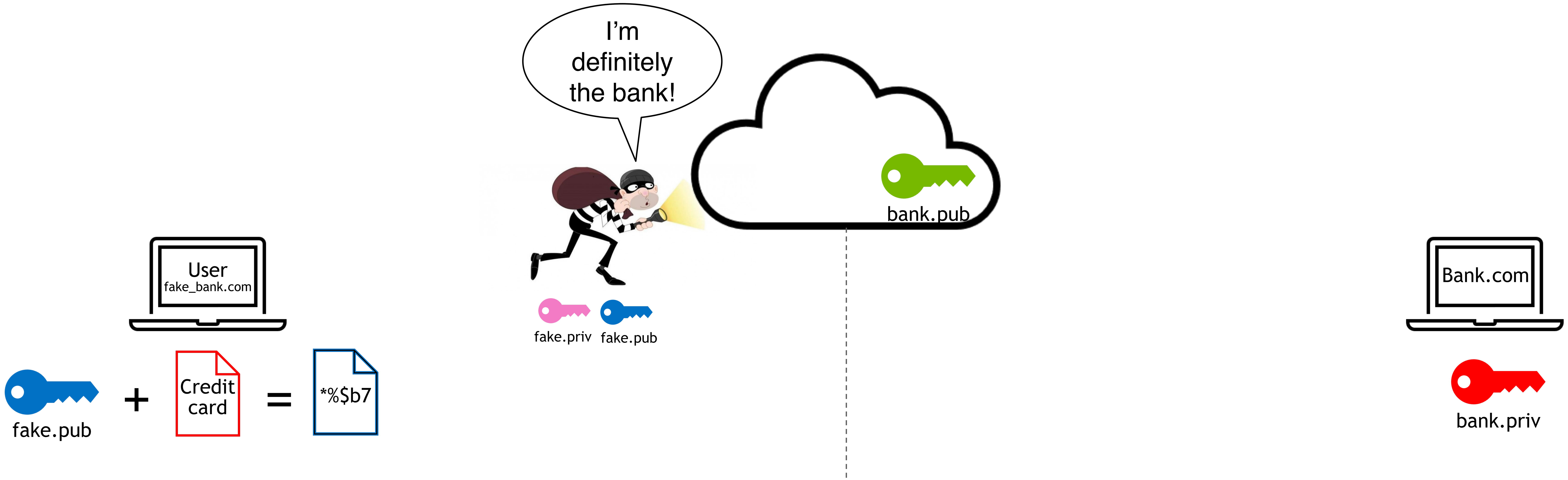
- What if a bad-actor acts as an impostor?
- They can pretend to be a legitimate public-key-provider



Attestation

Public Keys Aren't Enough to Prove Identity

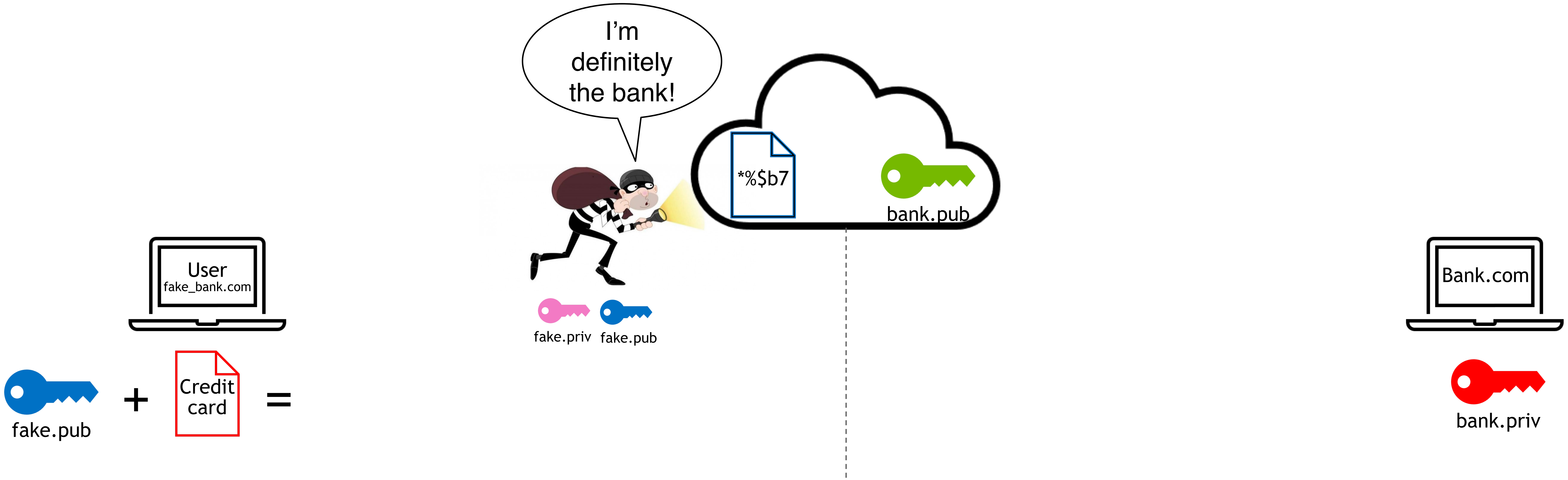
- What if a bad-actor acts as an impostor?
- They can pretend to be a legitimate public-key-provider



Attestation

How to Prove Identity

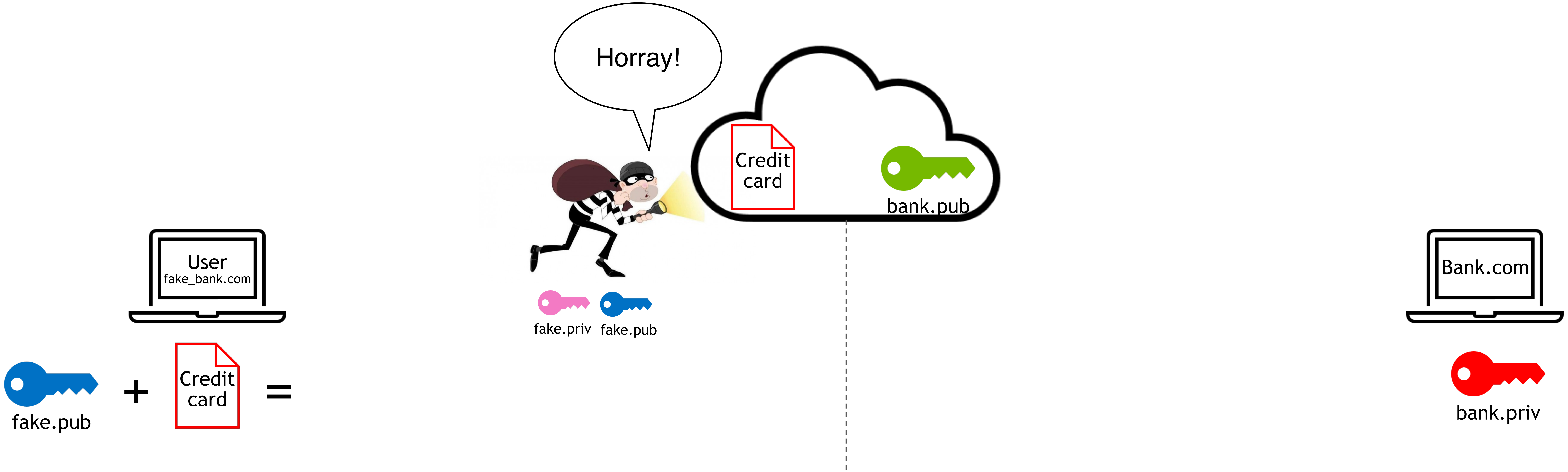
- What if a bad-actor acts as an impostor?
- They can pretend to be a legitimate public-key-provider



Attestation

Public Keys Aren't Enough to Prove Identity

- What if a bad-actor acts as an impostor?
- They can pretend to be a legitimate public-key-provider
- And can access your information!



Attestation

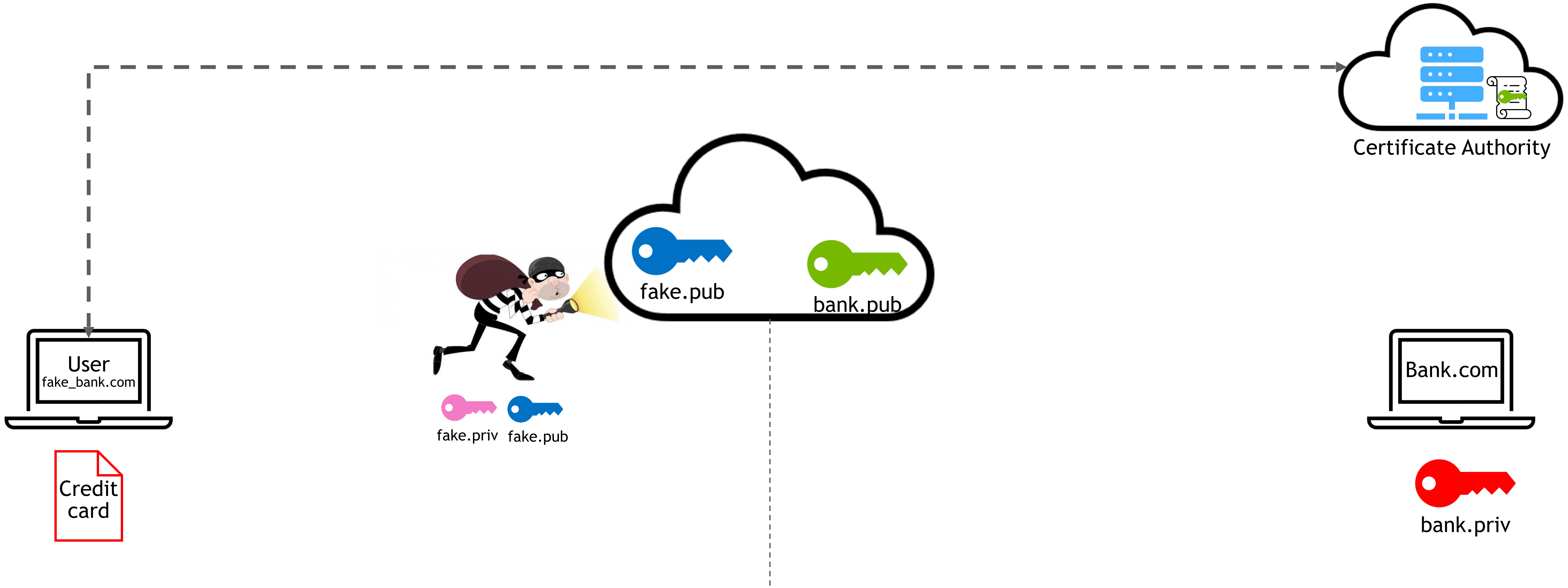
Enter: Certificate Authorities

- Publishing your Public Keys enables users to validate that information truly comes from YOU
- Certificate Authorities ("CA"s) exist to provide another layer of security to prove identity

Attestation

How to Prove Identity

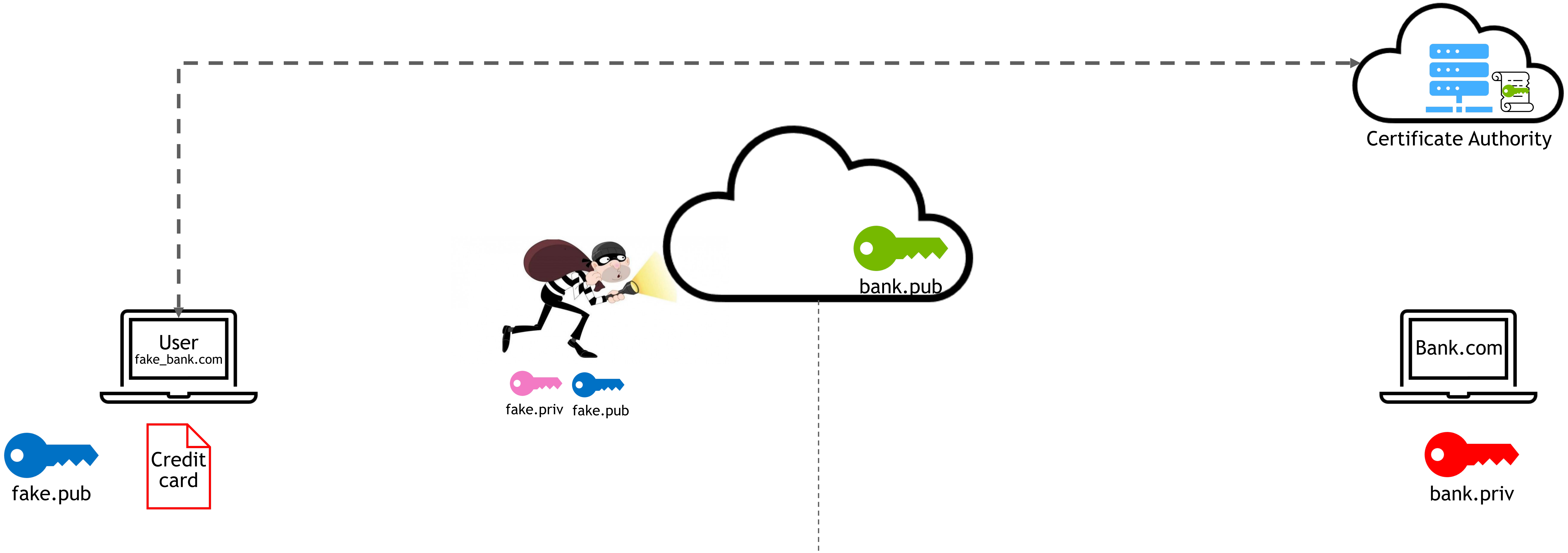
- Publishing your Public Keys enables users to validate that information truly comes from YOU
- Certificate Authorities ("CA"s) exist to provide another layer of security to prove identity
- They "sign"/certify that a given public key is authentic



Attestation

How to Prove Identity

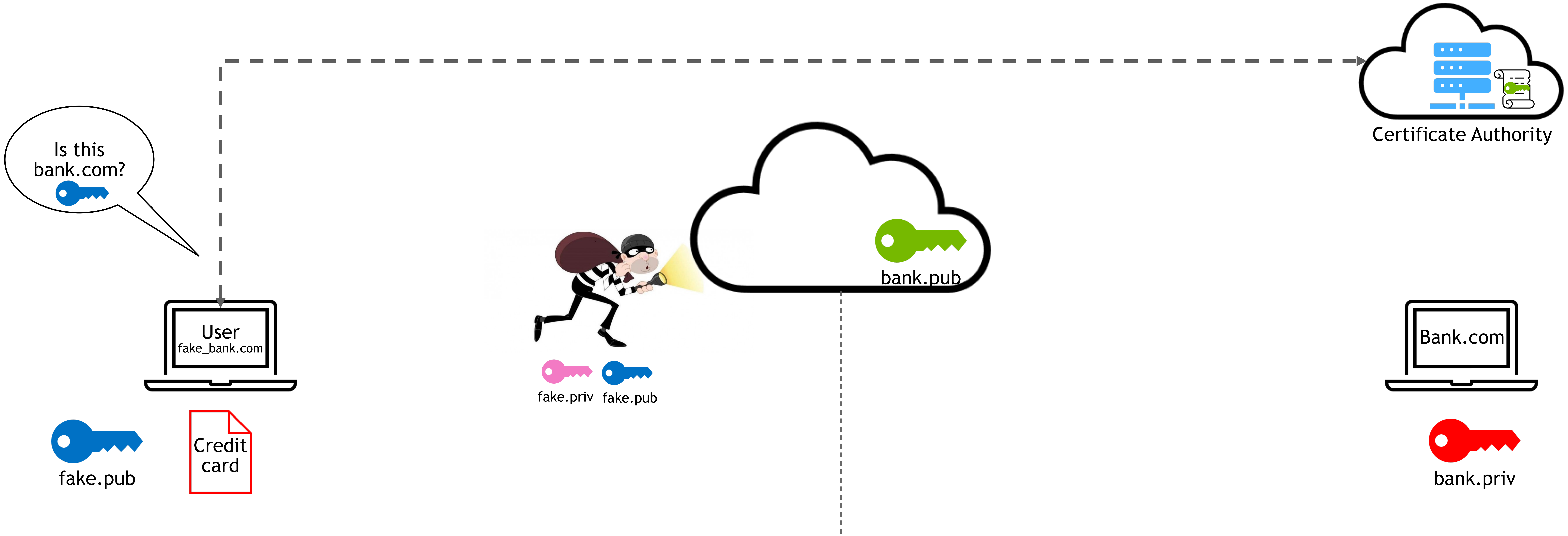
- Publishing your Public Keys enables users to validate that information truly comes from YOU
- Certificate Authorities (“CA”s) exist to provide another layer of security to prove identity
- They “sign”/certify that a given public key is authentic



Attestation

How to Prove Identity

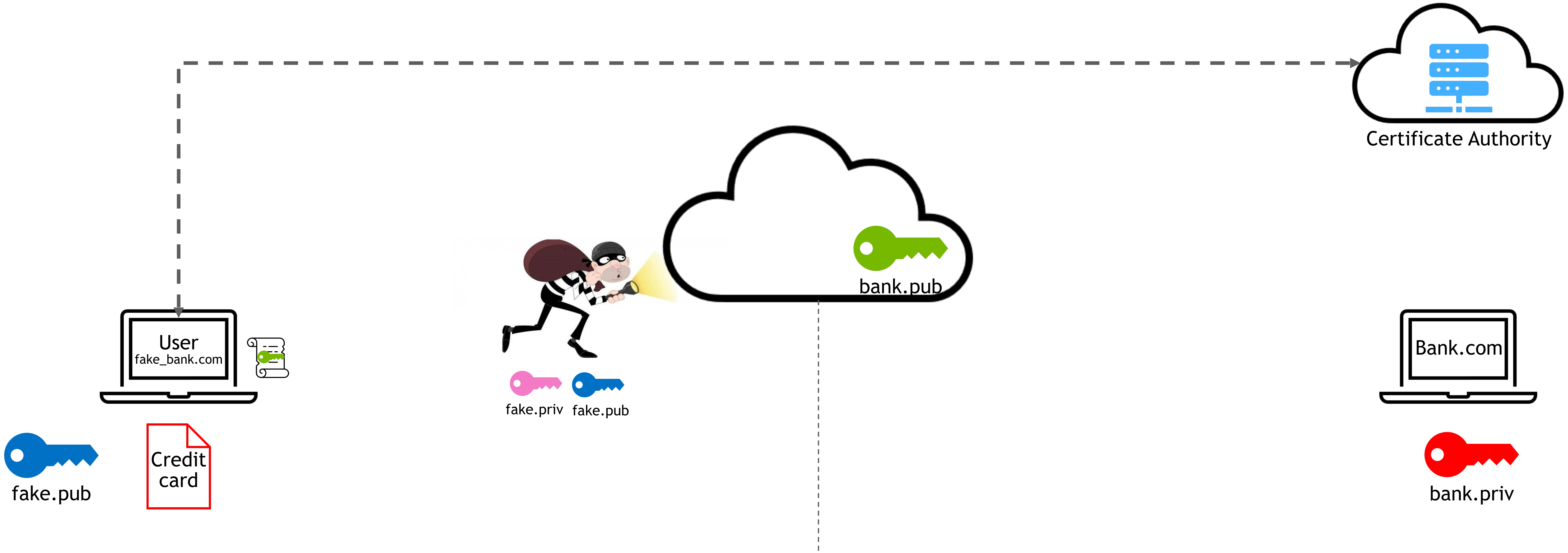
- Publishing your Public Keys enables users to validate that information truly comes from YOU
- Certificate Authorities ("CA"s) exist to provide another layer of security to prove identity
- They "sign"/certify that a given public key is authentic



Attestation

How to Prove Identity

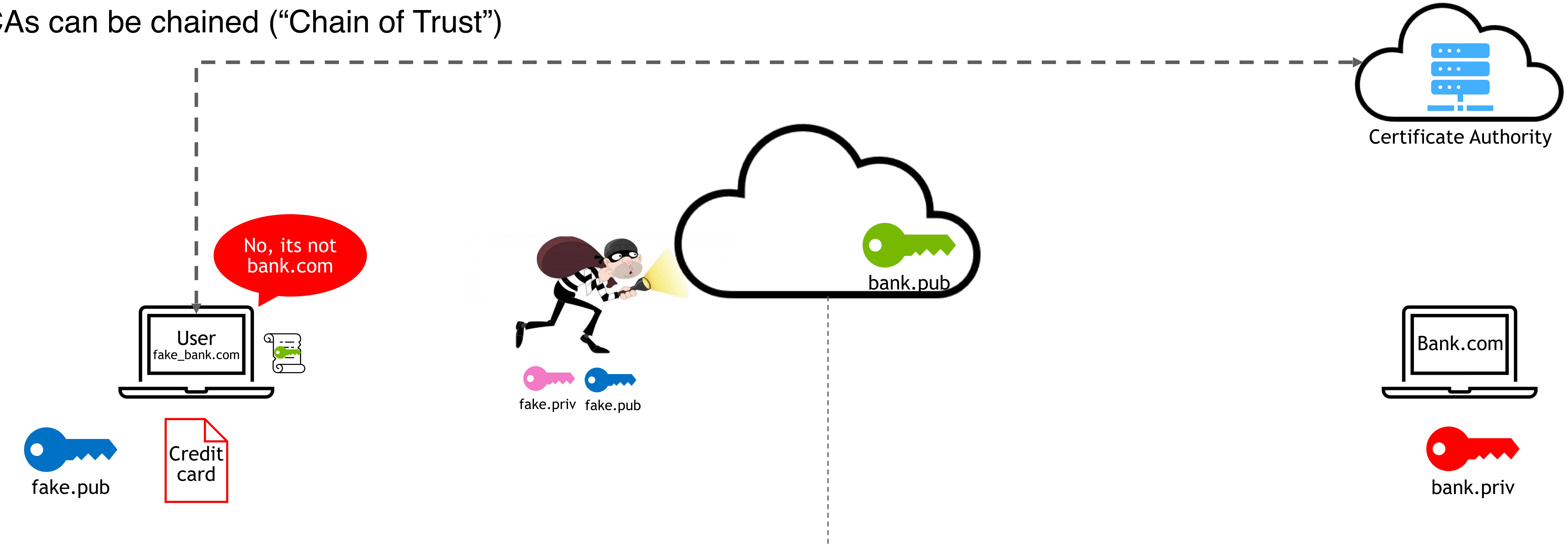
- Publishing your Public Keys enables users to validate that information truly comes from YOU
- Certificate Authorities ("CA"s) exist to provide another layer of security to prove identity
- They "sign"/certify that a given public key is authentic



Attestation

How to Prove Identity

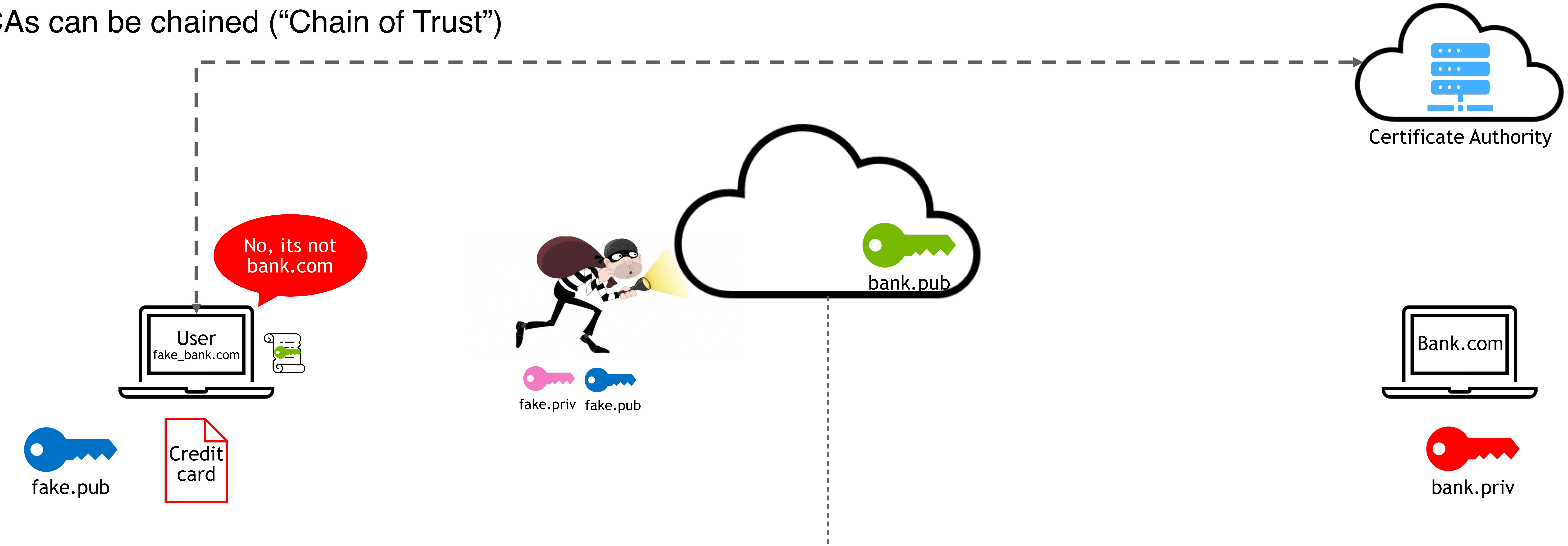
- Publishing your Public Keys enables users to validate that information truly comes from YOU
- Certificate Authorities ("CA"s) exist to provide another layer of security to prove identity
- They "sign"/certify that a given public key is authentic
- They provide another layer of confirmation
- CAs can be chained ("Chain of Trust")



Attestation

How to Prove Identity

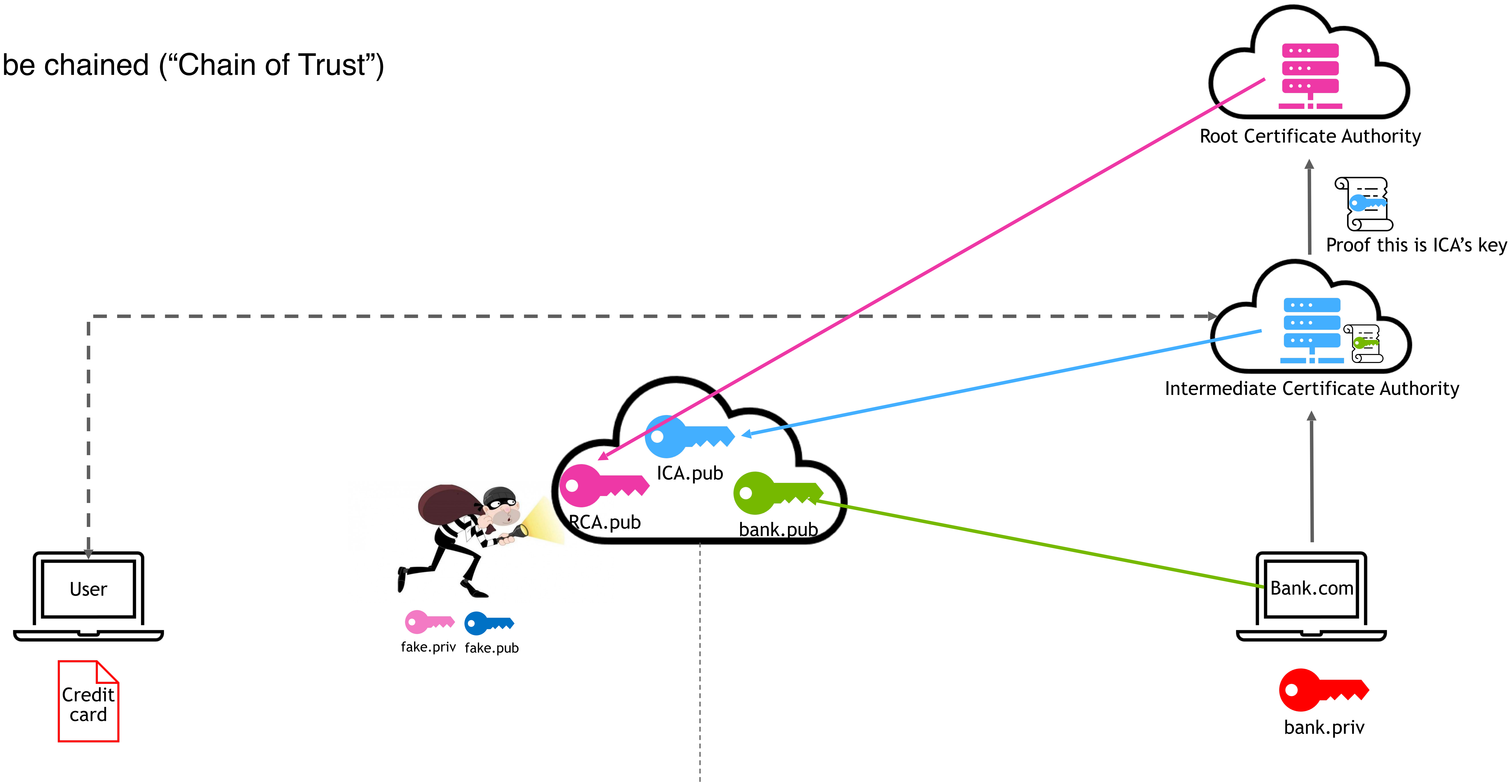
- Publishing your Public Keys enables users to validate that information truly comes from YOU
- Certificate Authorities ("CA"s) exist to provide another layer of security to prove identity
- They "sign"/certify that a given public key is authentic
- They provide another layer of confirmation
- CAs can be chained ("Chain of Trust")



CA Signed Certificate

How to Prove Identity

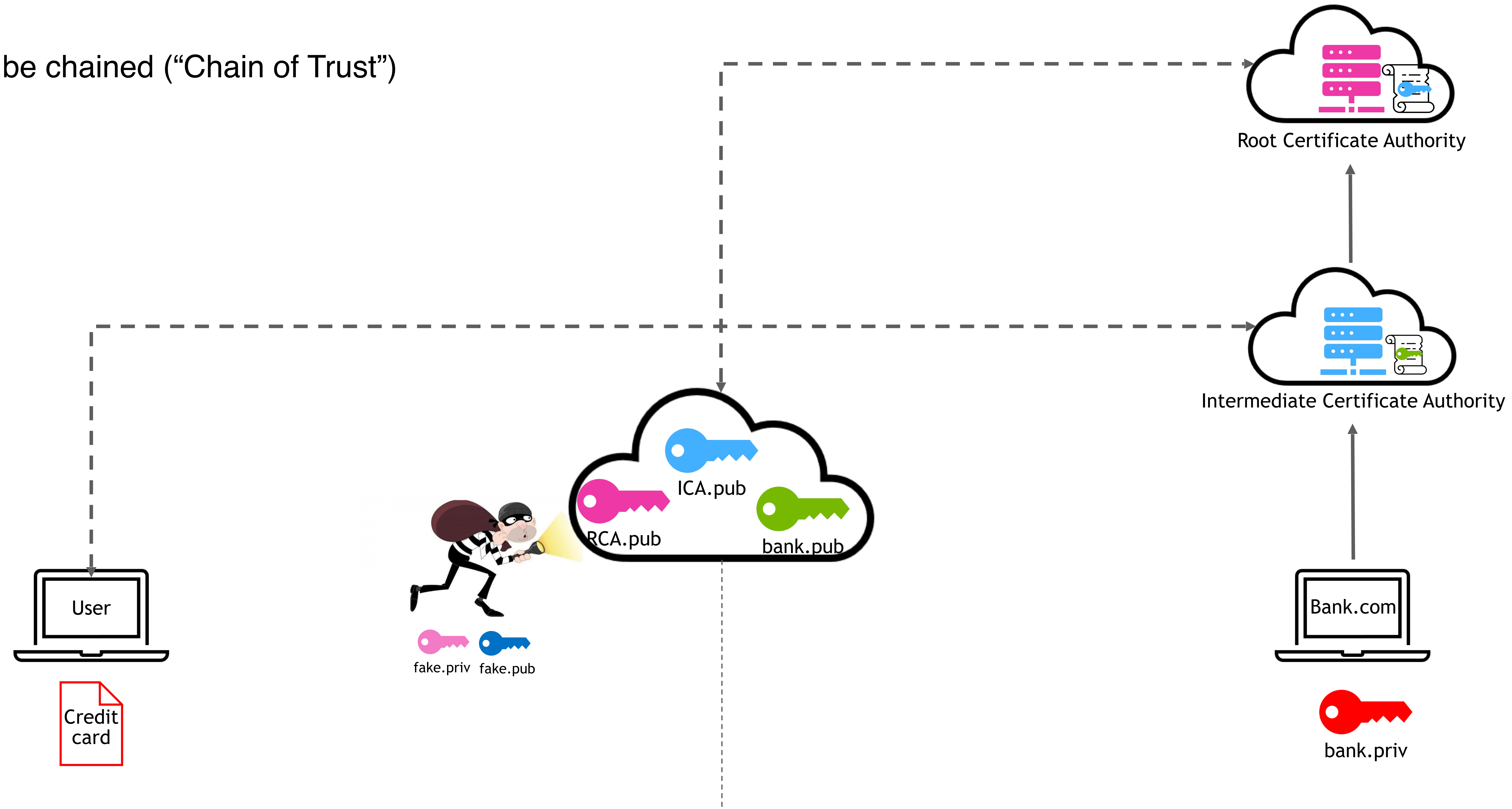
- CAs can be chained ("Chain of Trust")



CA Signed Certificate

How to Prove Identity

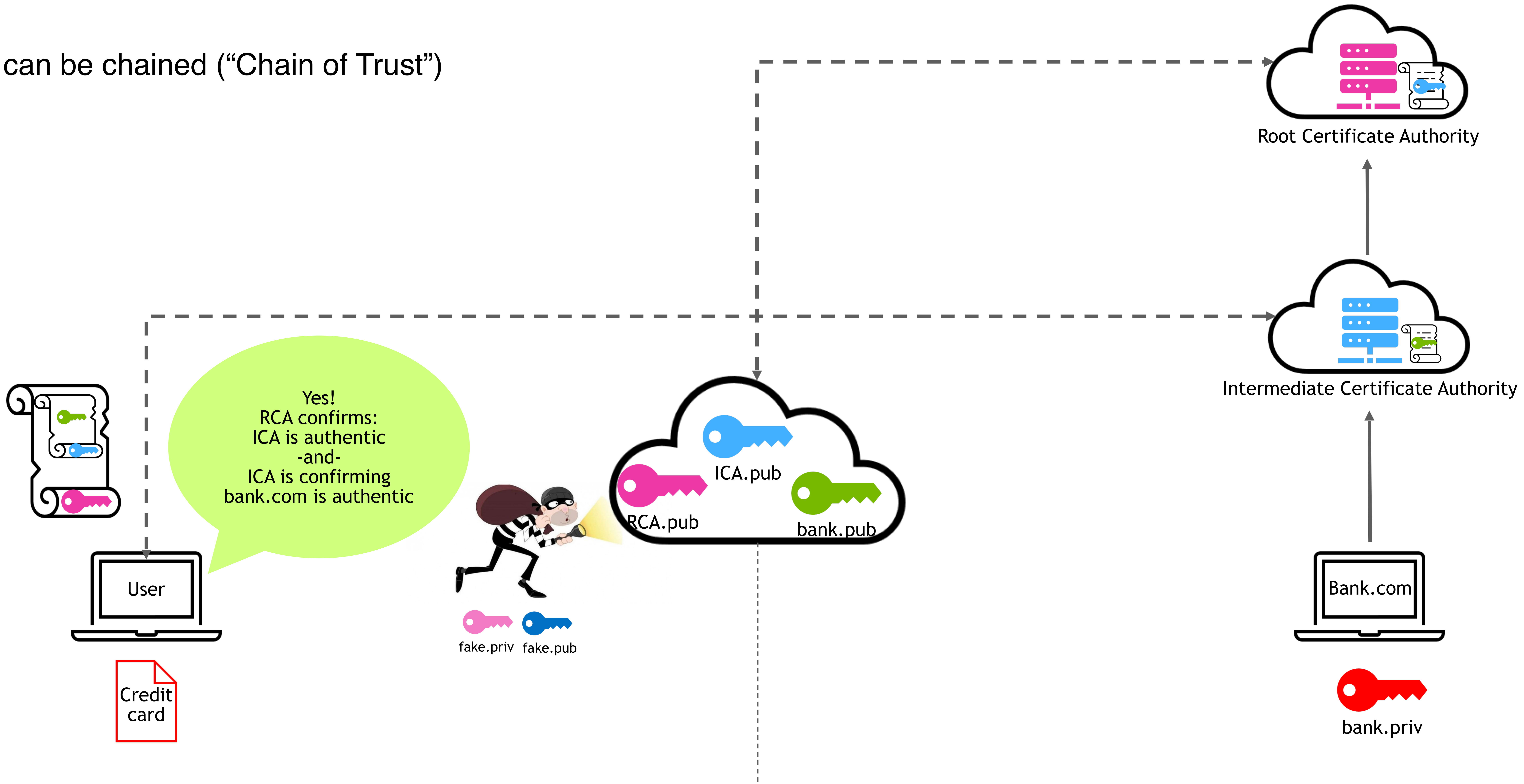
- CAs can be chained ("Chain of Trust")



CA Signed Certificate

How to Prove Identity

- CAs can be chained ("Chain of Trust")



Applying to Confidential Computing

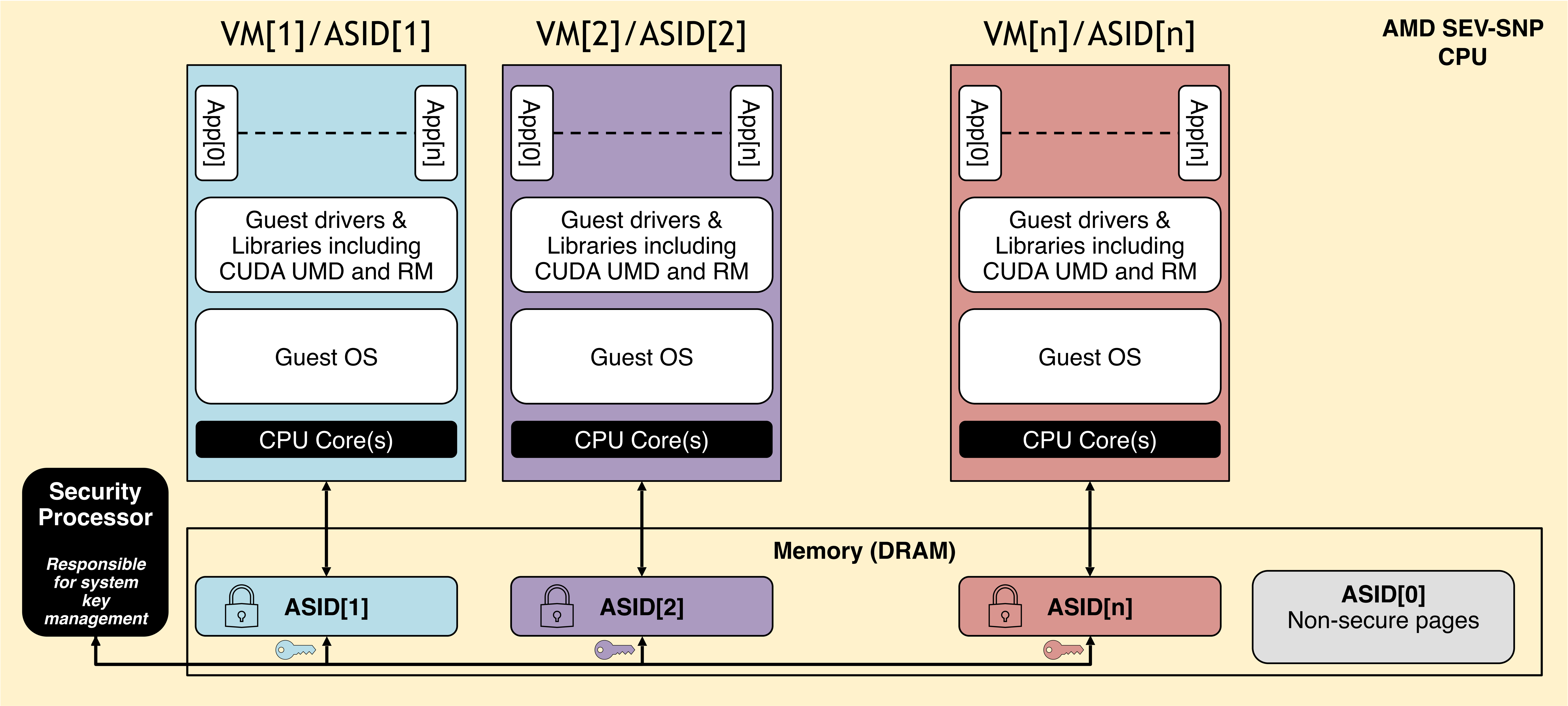
Users Have Sensitive Data

Increased Liability for the Developer, GPU Operator, etc.

- Developers and users have sensitive data. This may be due to privacy-based regulatory requirements, trade secrets, or other liability-prone reasons
- These users are reluctant to utilize CSPs for the above reasons
  - They don't or can't trust the Technicians
- CPU vendors (Intel, AMD) have already begun providing confidential modes to prevent unauthorized access to data-in-use

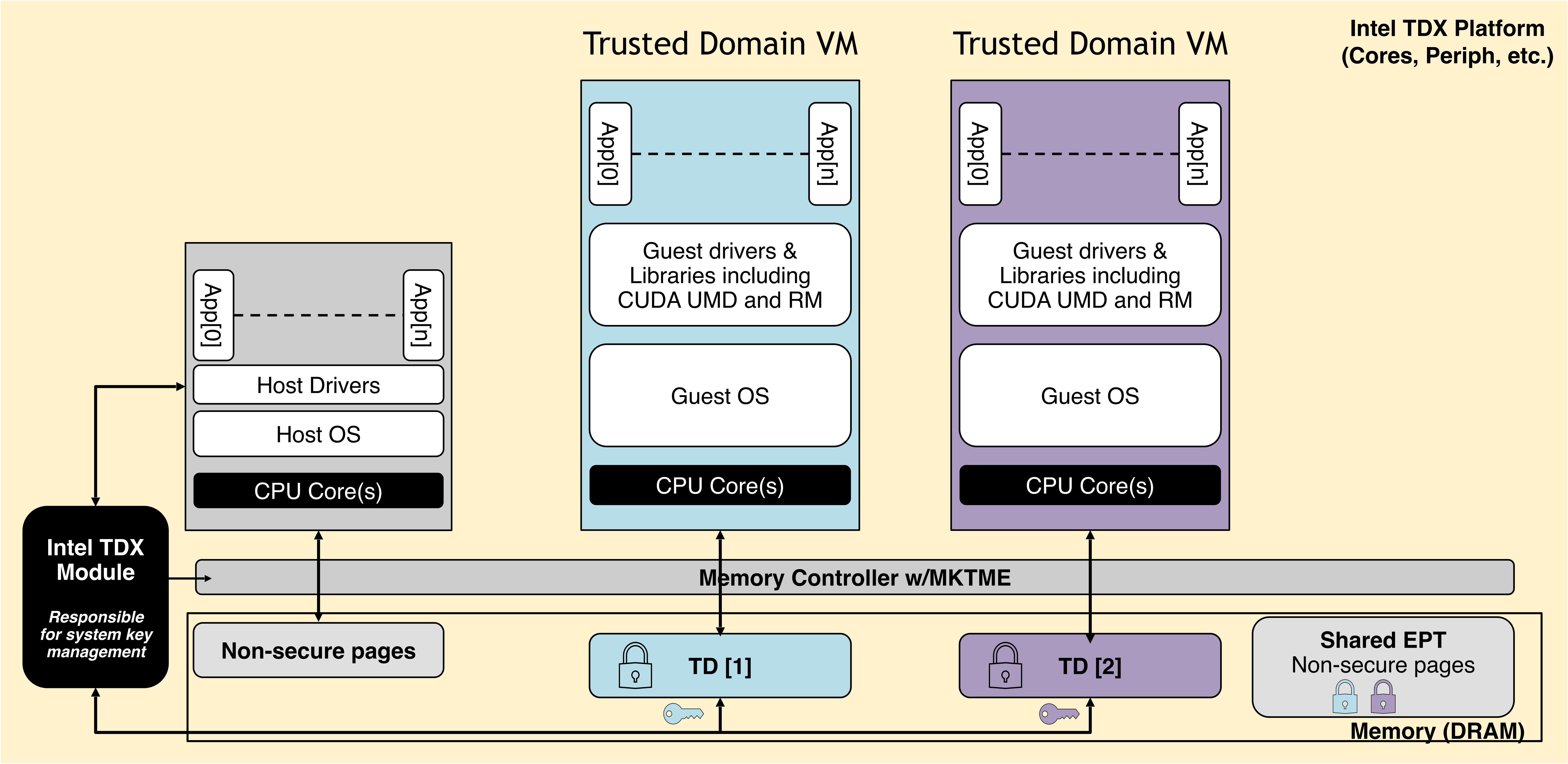
AMD SEV-SNP TEE

How AMD Handles Trusted Execution Environments



Intel TDX

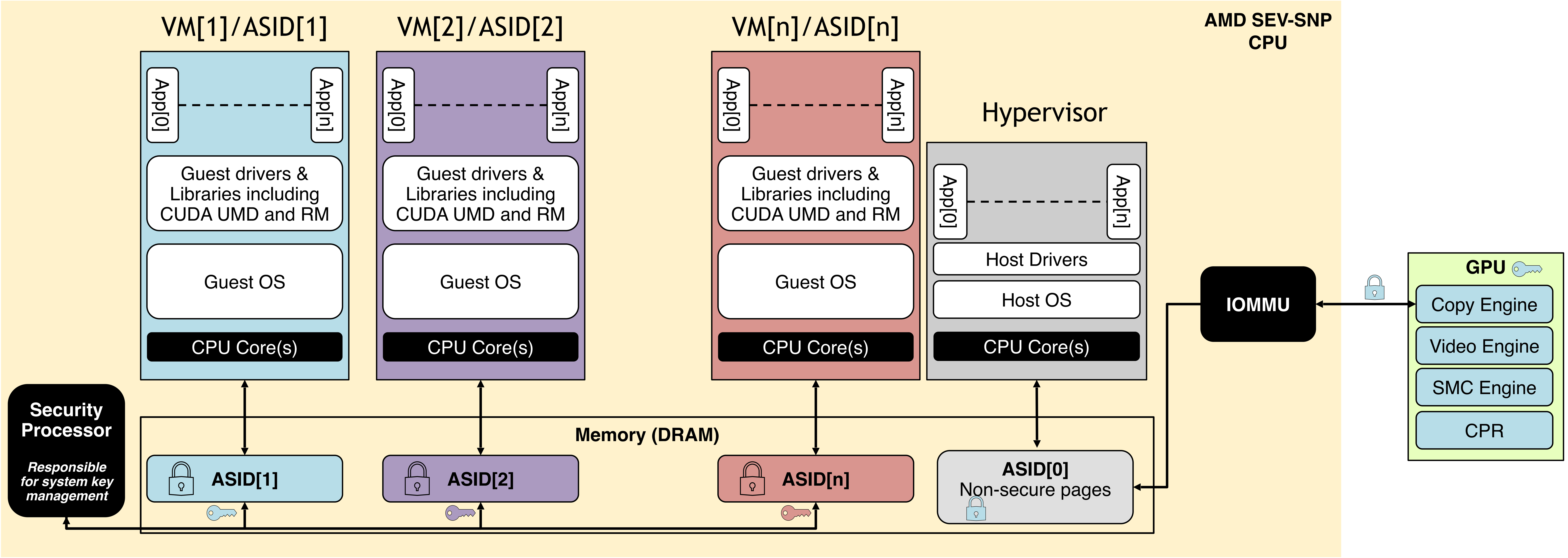
How Intel Handles Trusted Execution Environments



Users Have Sensitive Data

Increased Liability for the Developer, GPU Operator, etc.

- The CPU solutions are currently insufficient for PCIe based accelerators
- Hopper Confidential Computing extends the trust boundary to the GPU
    - Requires hardware based Trusted Execution Environment (TEE)
- Prevents physical bus-snooping, or Hypervisor access to GPU



Confidential Computing with Hopper H100

NVIDIA Hardware Roadmap to Confidential Computing

V100

AES firmware authentication

2018

Tesla T4

Encrypted firmware

NVRISCV introduction

2019

A100/ A30

PKC firmware authentication

External RoT

Firmware revocation

2020

A40/A10/A2

Secure Boot (all firmware auth'ed)

HW fault injection countermeasures

SR-IOV

2021

H100

Confidential Compute Capable

On-Die RoT

Measured/Attested Boot

FIPS 140-3 Level 2

2022

Mapping Core Use Cases to Hopper-CC

	Confidential Inference	Confidential Training	Confidential Federated Learning	Use Cases Beyond AI
Confidential MIG	✓			✓
Single GPU	✓	✓	✓	✓
Multi GPU		✓	✓	✓

Confidential Computing Roadmap

Single GPU Passthrough

Non-virtual server with exclusive ownership of the GPU

↓

Single Confidential GPU direct-attached to a single Trusted VM.

Inference, HPC, lightweight training.

Multi GPU Pass-Through

GPU is directly & exclusively assigned to a VM

↓

Multiple Confidential GPUs direct-attached to a single Trusted VM.

Training workloads, NVLink Support, configurable topologies.

Multi-Instance GPU/MIG

Physical GPU is split into per-VM partitions

↓

Confidential Computing for multiple tenants per GPU

Uses SR-IOV & vGPU: Microservice inference, edge aggregation HPC

Threats Addressed by Confidential Computing

Protecting Data in Use from the owner of the Compute Infrastructure

In a cloud environment where users rent VM instances from a CSP, the following threats are mitigated:

- Data and Code Confidentiality – protect all application code and data in the VM instance from being read by the host
- Data and Code Integrity - protect all application code and data in the VM instance from being altered by the host
- Physical Attacks with everyday tools – interposers on buses such as PCIe and DDR memory cannot leak data or code

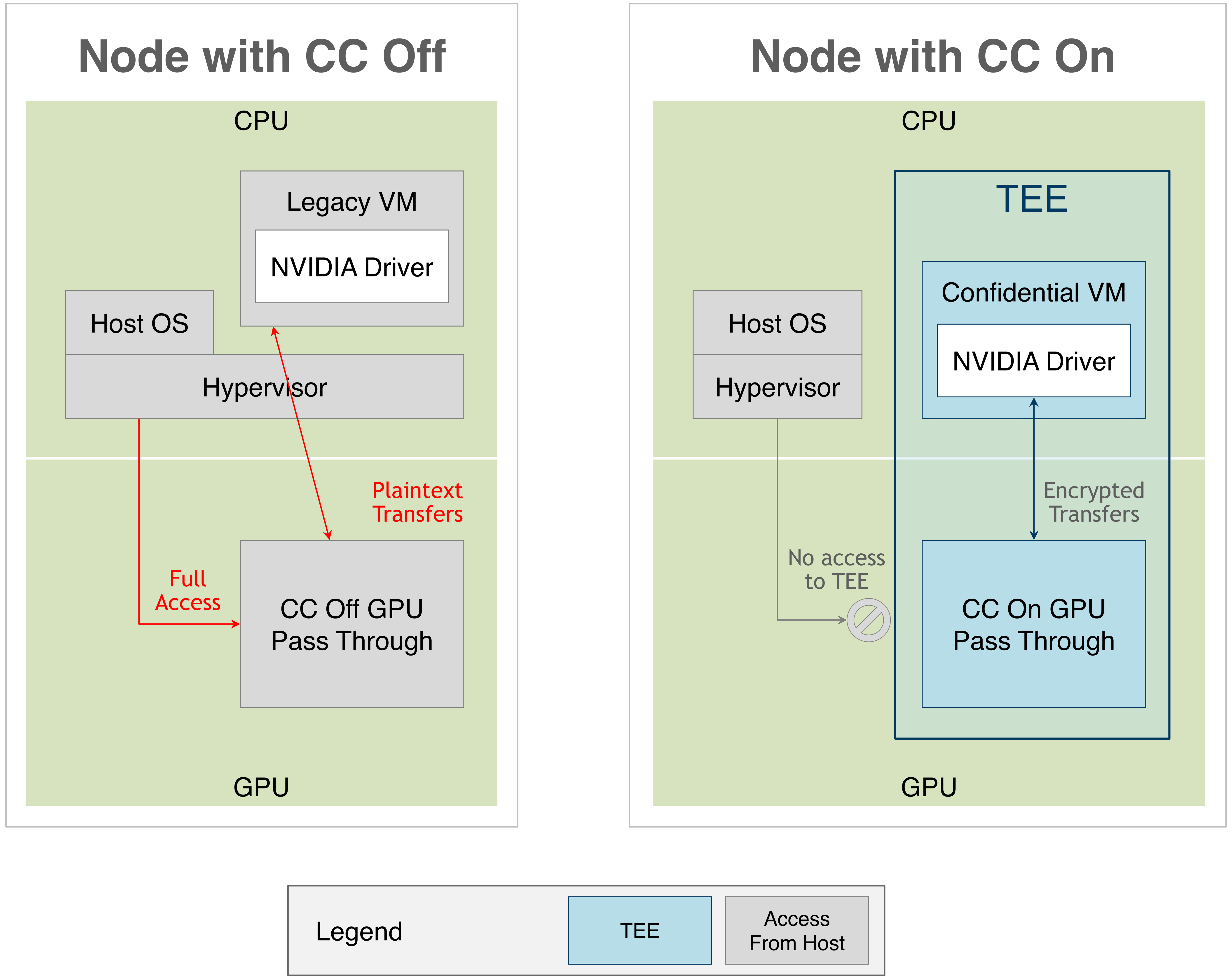
Threats & Mitigations

Category	Threat	Mitigation
Confidentiality	Use PCIE/NVLINK to read tenant data (e.g. Hypervisor, another VM, PCIE interposer)	✓
	Use Out-of-band management/debug channels to read tenant data (e.g. SMBus, JTAG)	✓
	Use memory remapping to read tenant data	✓
	Use GPU Cache/Memory based side channels to read tenant data	✓
	Use GPU TLB based side channels to read tenant data	✓
	Use GPU Performance Counters to read tenant data or fingerprint tenant	✓
	Read tenant data via hypothetical physical attacks (physical side channels / DPA / EM, HBM interposer)	✗
Integrity	Use PCIE/NVLINK to modify tenant data (e.g. Hypervisor, another VM, PCIE interposer)	✓
	Use Out-of-band management/debug channels to modify tenant data (e.g. SMBus, JTAG)	✓
	Corrupt tenant data by replaying previous data or MMIO transactions (replay attacks)	✓
	Corrupt tenant data via hypothetical physical attacks (fault injection, HBM interposer)	✗
Availability	Denial of Service to hypervisor by tenant	✓
	Denial of Service to tenant by another tenant	✓
	Permanent denial of service of GPU by tenant	✓
	Denial of Service to tenant by hypervisor	✗
General	Use a spoofed, non-genuine, or known vulnerable TCB component	✓
	Use hardware side channels (e.g. DPA) to extract persistent device keys	✓
	Use hardware side channels (e.g. DPA) to extract tenant ephemeral session key	✗

NVIDIA Confidential Computing Introduction

Protecting Data and Code from Hypervisor and Physical Attacks

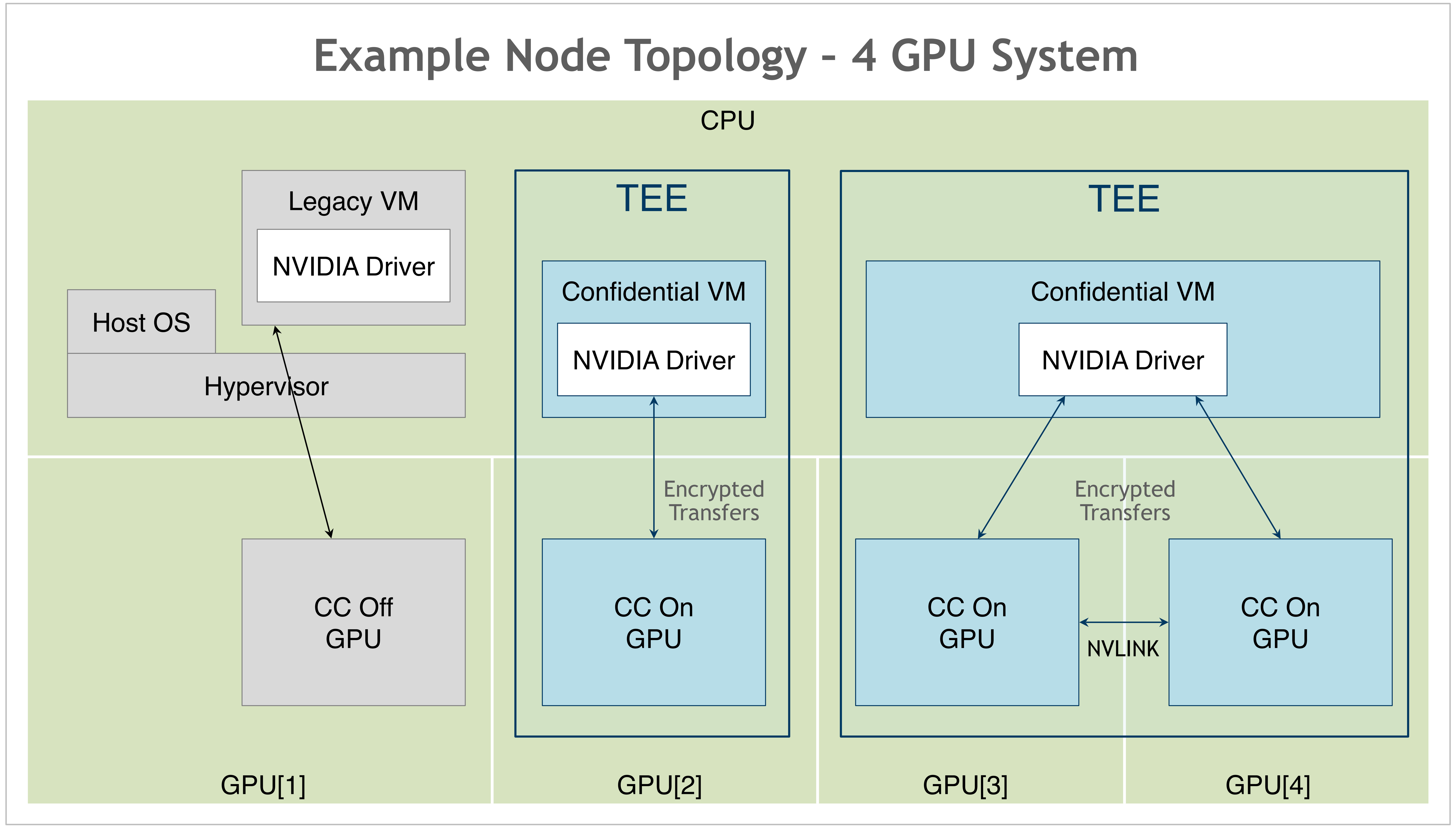
- Prerequisites:
  - CPU with support for a Virtualized-based TEE (“Confidential VM”)
  - Supported variants are AMD Milan or later, or Intel SPR and later.
- Capabilities:
  - Trusted Execution Environment
    Isolated environment providing confidentiality & integrity
  - Virtualization-based
    Applications can run unchanged and do not have to be partitioned
  - Secure Transfers
    High performance HW acceleration for encrypted CPU/GPU transfers
  - Hardware Root of Trust
    Authenticated firmware; measurement & attestation for the GPU



Secure Passthrough

Confidential Computing for Exclusive Assignment of a GPU to a VM

- VM-based TEE extended to protect entire workload:
  - NVIDIA drivers, libraries, APIs run in the VM
  - No application changes are needed to run in this mode
- Exclusive assignment to VM for each physical GPU:
  - VM may have multiple GPUs
- Isolation from host for confidentiality and integrity:
  - Encrypting the data over PCIE
  - Encryption transparent to applications



CC On Mode Initialization Sequence

Mode enablement and session establishment

Mode Enable

1 BMC issues out-of-band request to persistently enable CC mode

NVIDIA OOB Specification will provide APIs to integrate into customer tools and OpenBMC

Host triggers GPU reset for mode to take effect

2 GPU firmware scrubs GPU state & memory

Device Boot

3 GPU firmware configures firewall to prevent unauthorized access, then enables PCIE

4 GPU PF driver uses SPDM for session establishment & attestation report

Tenant Initialization

5 Tenant attestation service gathers measurements, device certificate using NVML APIs

6 Verification done locally or transmitted to remote service

CUDA programs allowed to use GPU

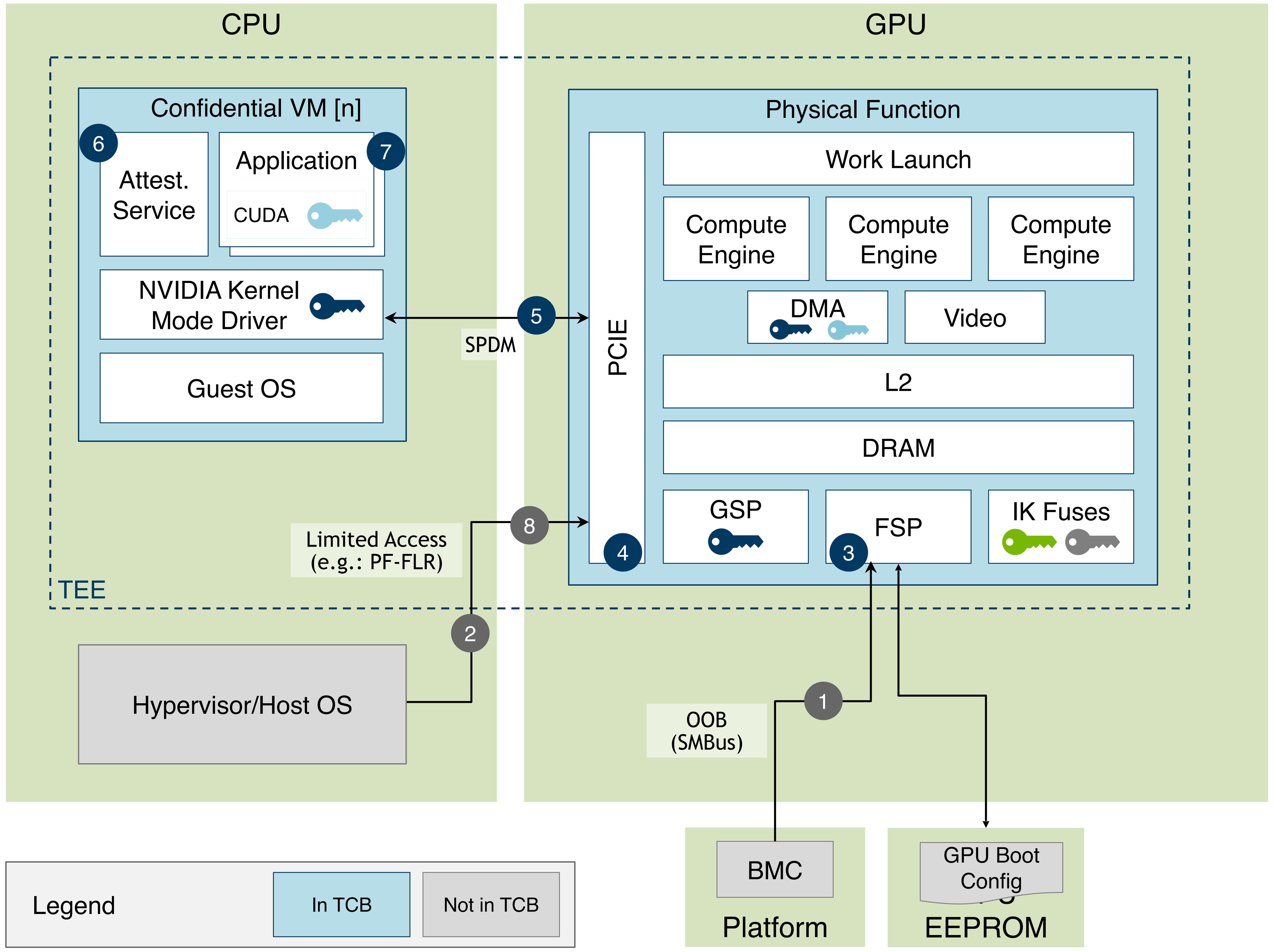
Host triggers PF-FLR to reset GPU; returns to device boot for scrubbing GPU state & memory

7

Tenant Shutdown

8

3



Attestation: *"I Am Who I Say I Am"*

Preventing a Spoofed GPU

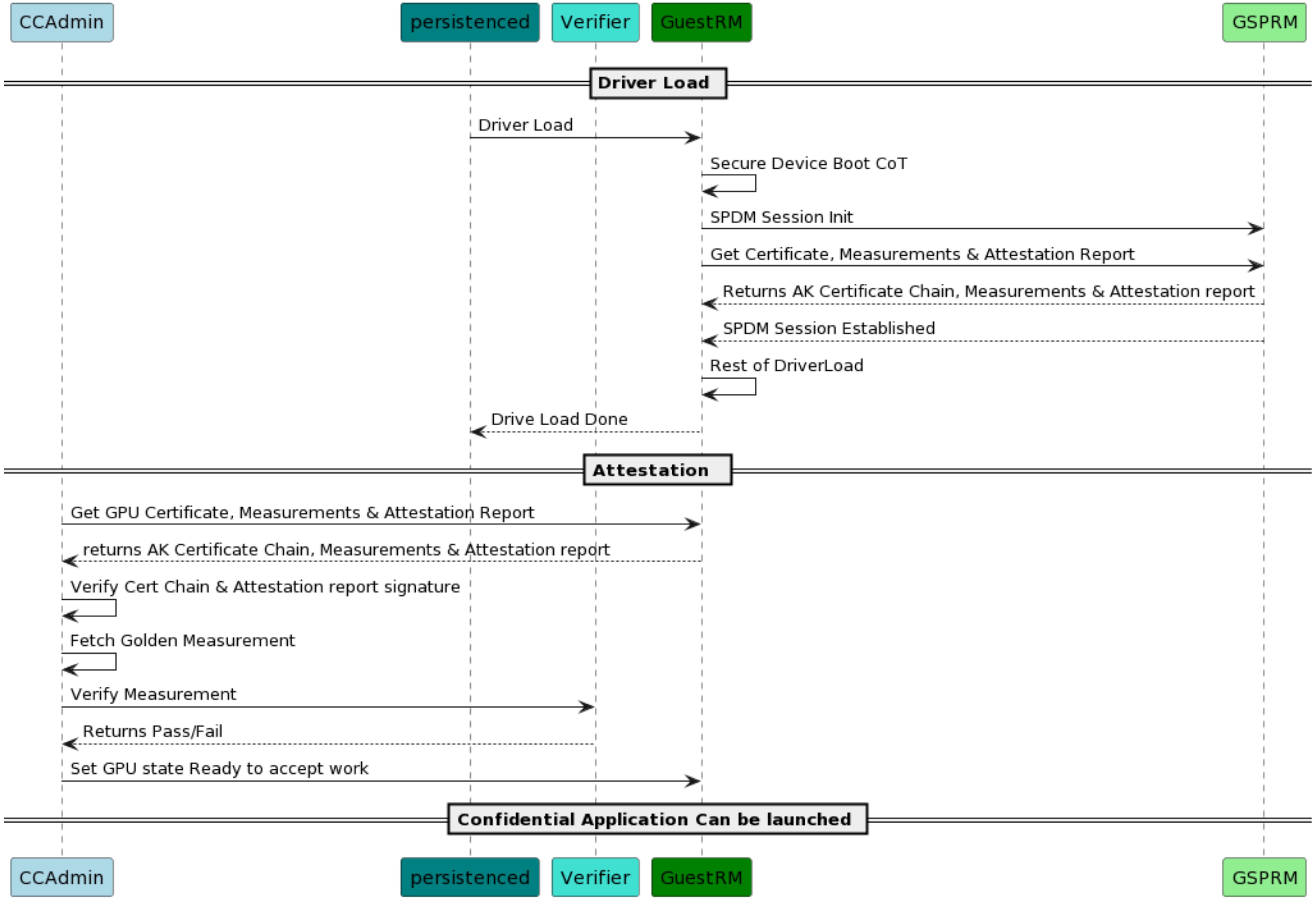
- Since all the local-attack surfaces are covered, how do we stop physically 'spoofing' a GPU?
- Attestation reports are generated with device-specific measurements, public/private key pairs, etc.

The Attestation Verifier Application

How to Verify an Authentic GPU

- An NVIDIA provided, open-source application, will be provided to developers to verify the GPU's readiness to accept Confidential workloads
- The GPU, after initialization, will be ready to provide Evidence that it is authentic
  - Static and dynamic device measurements, firmware versions, driver microcode, signed/Endorsed by the device
- The Verifier application has a Reference set of measurements, which is used to compare to the Endorsed Evidence provided by the GPU
- The Verifier application reviews and compares the Endorsed Evidence against the Reference measurements
  - If anything does not match, it will fail
- If all Endorsed Evidence matches expected Reference values, the Verifier will inform the developer the GPU is authentic
- Verifier will also return whether the GPU is in a Confidential Computing mode, or if it is in standard mode.

Tenant Initialization



Confidential MIG on vGPU

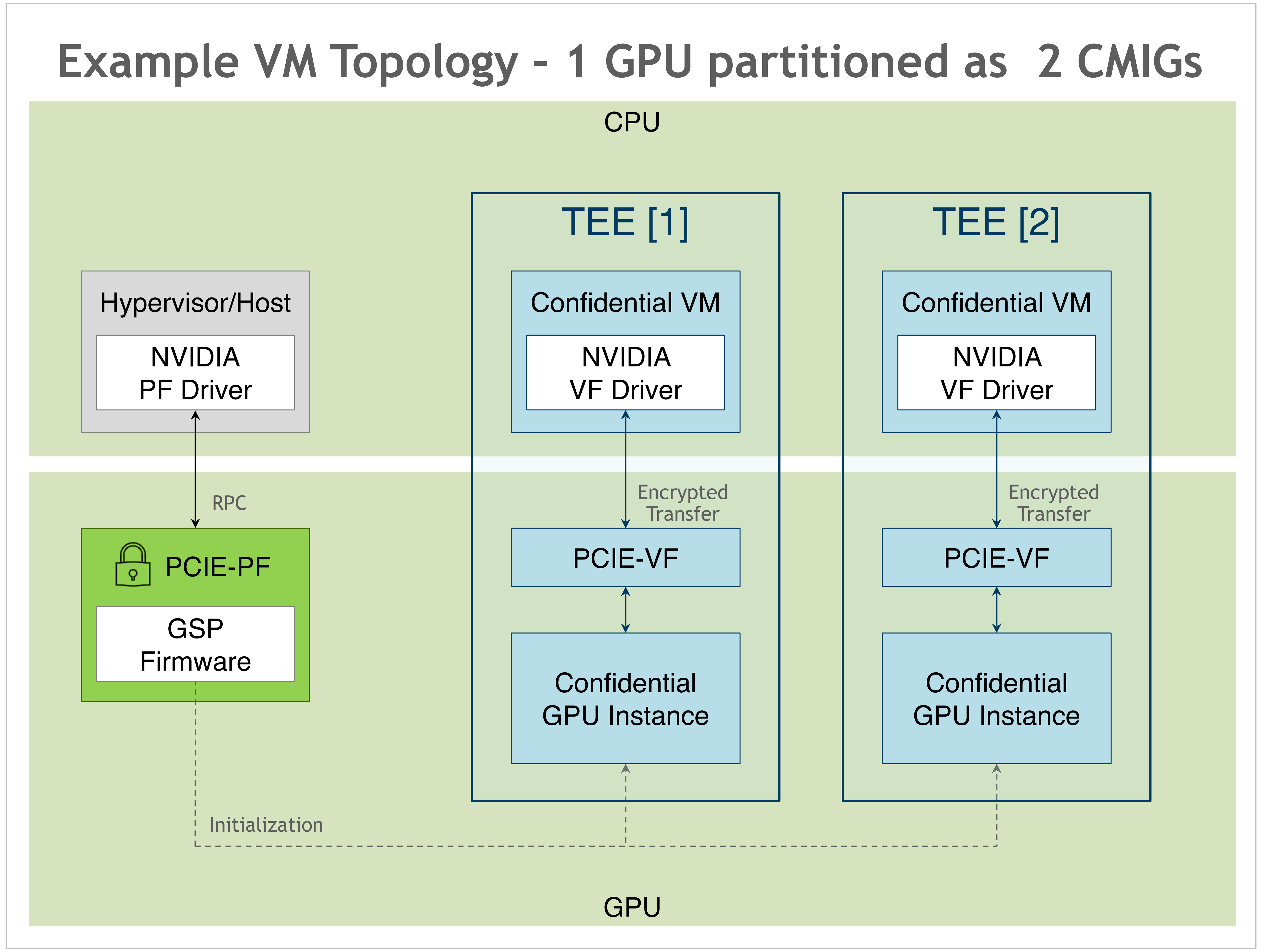
Confidential Computing for a GPU partitioned across VMs

vGPU with Confidential MIG provides multi-tenant CC:

- TEE = Secure GPU Instance + Confidential VM
- TEE isolated from hypervisor/host and other tenants
- Hypervisor/Host creates TEE instances but cannot access state

Feature is based on:

- NVIDIA vGPU – enables multiple VMs sharing a single NVIDIA GPU
- Multi-Instance GPU (MIG) – partitioning of GPU into instances
- SR-IOV – PCIE devices expose VF for direct control by VM



NVIDIA

Hopper CC Performance

- Hopper Confidential Computing runs GPU side compute at full performance

- Applications with high compute vs data transfer will run at close to non-CC Hopper performance
- This includes AI Training for most AI models with large batch sizes
- This includes AI Inference for large modes (BERT, Transformer, etc.)

- AI Inference on smaller models, e.g., Resnet50, will see slowdowns on Hopper CC, but would still be significantly faster than CC on CPU.

- Some CUDA APIs have different performance characteristics
  - Execution control APIs (e.g. kernel launch, CPU side synchronization) can take longer under HCC
  - Allocations made via cudaMallocHost/cudaHostAlloc may have increased access latencies
  - Data transfers via cudaMemcpy or cudaMallocManaged may have lower performance under HCC

Confidential Compute Security Standards

Overall	Device Identity	Session
FIPS 140-3 Level 2 (All implementations of crypto operating on tenant data)	Online Certificate Status Protocol/OCSP (Device Identity Certificate Revocation; NVIDIA managed)	DMTF Security Protocol and Data Model/SPDM (Key Exchange, Retrieve Device Attestation Report)
	X.509 (Device Identity Certificate)	AES-GCM 256 (Confidentiality/Integrity of tenant traffic)
	ECC-384 (Device Private Key)	ECC-384 (Device Public/Private Key via SPDM)
		SHA2-384 (Measurements, HKDF)
		TCG Reference Integrity Manifest/RIM (Reference Measurements)

CUDA Considerations With Hopper Confidential Computing

CUDA Applications Will Run Unmodified

- NVIDIA has worked hard to ensure that a transition to Confidential Computing mode is as transparent to our developers as possible.

- Due to the architecture of CPUs' Trusted Execution Environments, developers should be aware of a few areas where changes may be required
  - Host memory cannot be directly accessed by the GPU, as it is blocked by the CPU's IOMMU:
    - Registering a host-pointer to the GPU will be blocked
    - CUDA stream batch MemOps targeting pinned CPU memory are not supported

- As the trust is built between the CPU and the GPU only, RDMA-based applications can't access either CPU or GPU memory directly
  - This applies to inter-device peer-to-peer and multinode NVLink as well
  - Multi-device can still perform cudaMemcpyD2D for multi-GPU environments

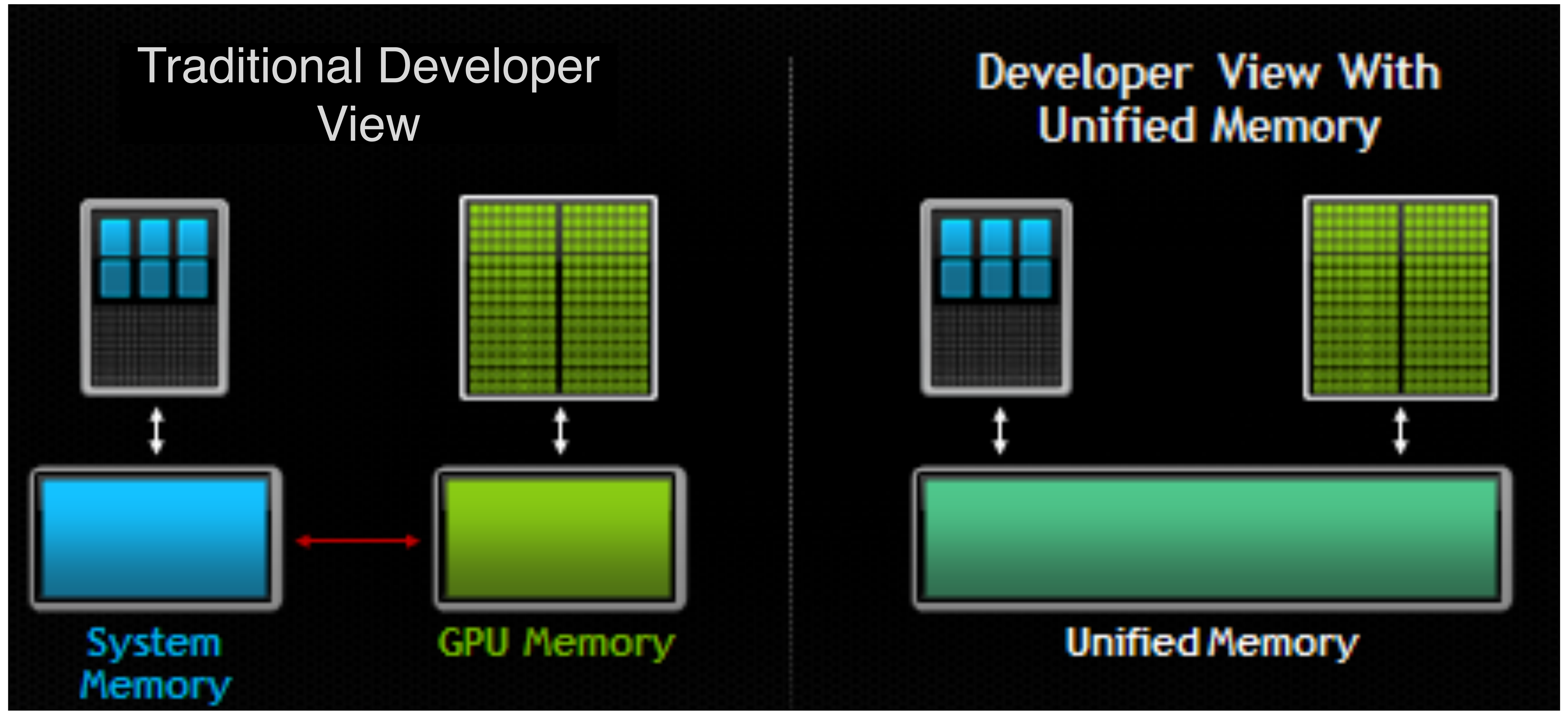
- Good coding practices of checking for supported functionality of the GPU with simple guardrails should be sufficient to catch and fall-back on any corner-case code

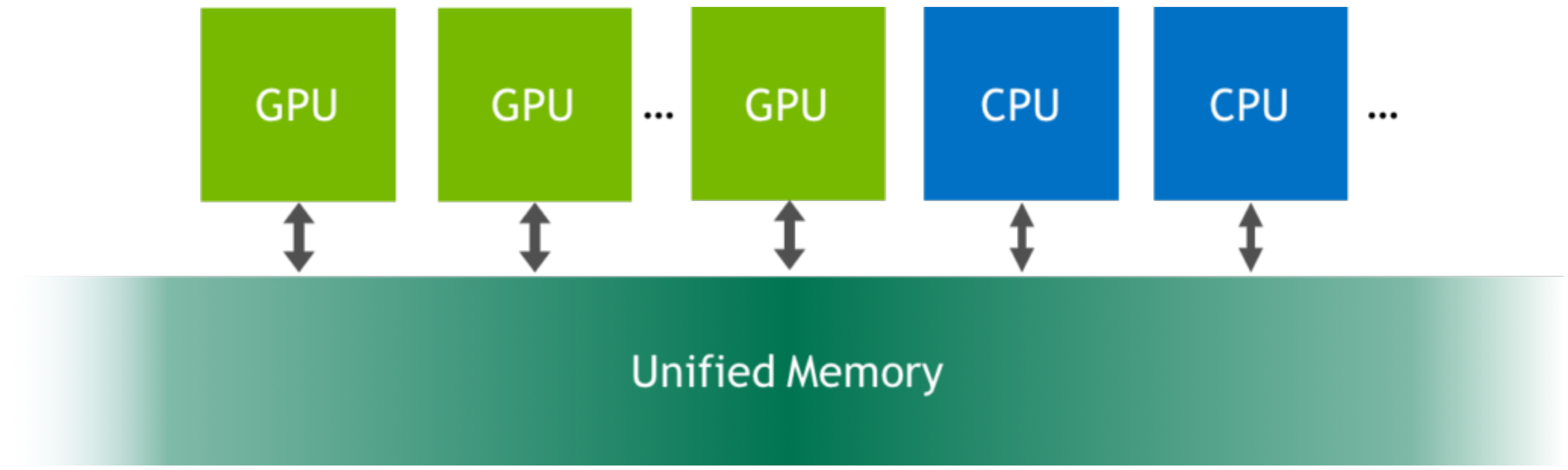
- Hopper H100 Confidential mode is primarily targeting compute acceleration

CUDA API Details

UVM: Optimized Memory Allocation Since CUDA 6.0

- CUDA handles memory in a flow called “Unified Virtual Memory” (UVM)
- Memory allocation done via cudaMallocManaged() will create a unified pointer that both GPU and CPU can access
- The driver handles all paging requirements, faults, code migration, etc.
- Creates incredibly easy developer experience

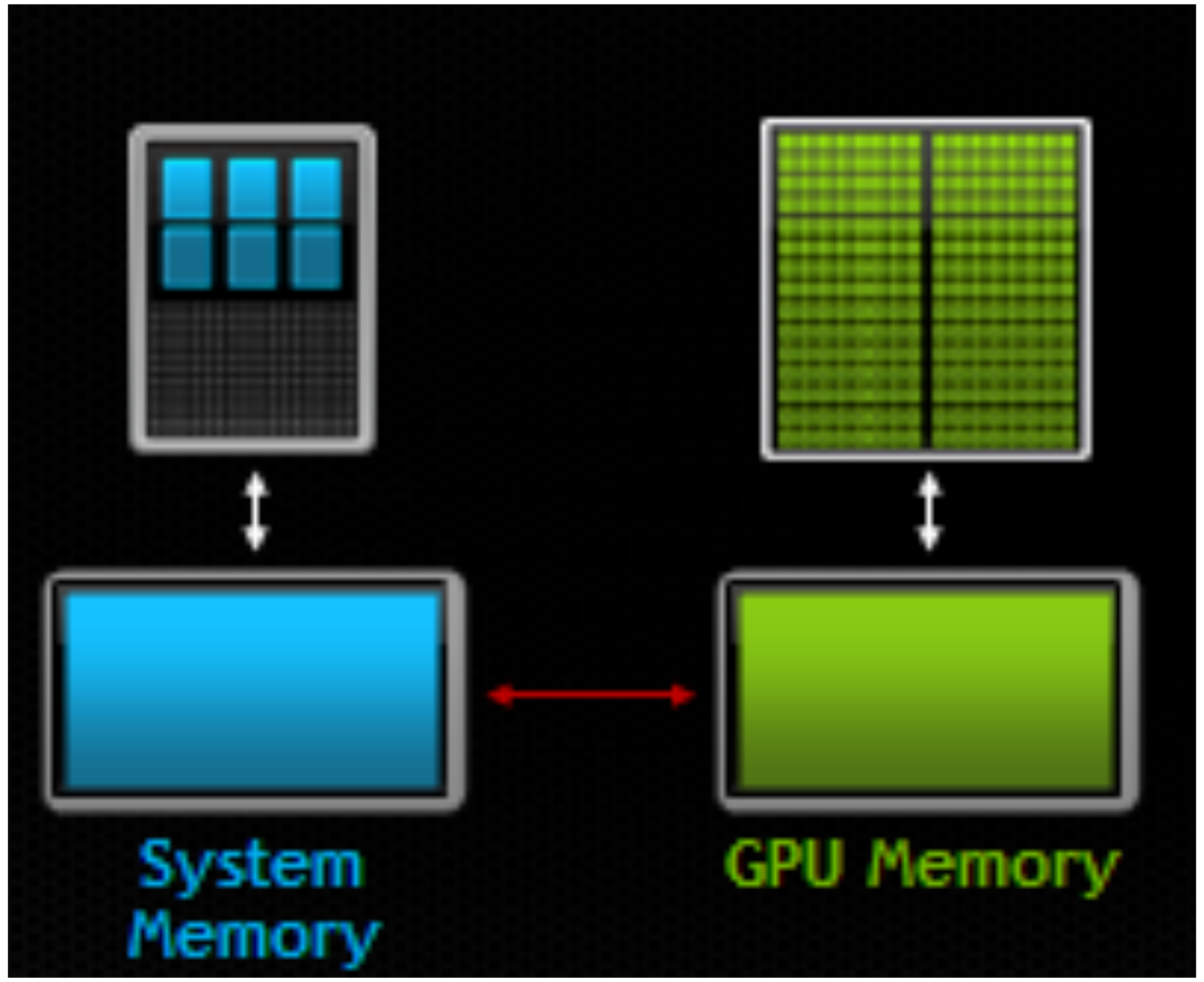




https://developer.nvidia.com/blog/unified-memory-in-cuda-6/
https://developer.nvidia.com/blog/unified-memory-cuda-beginners/

CUDA API Details

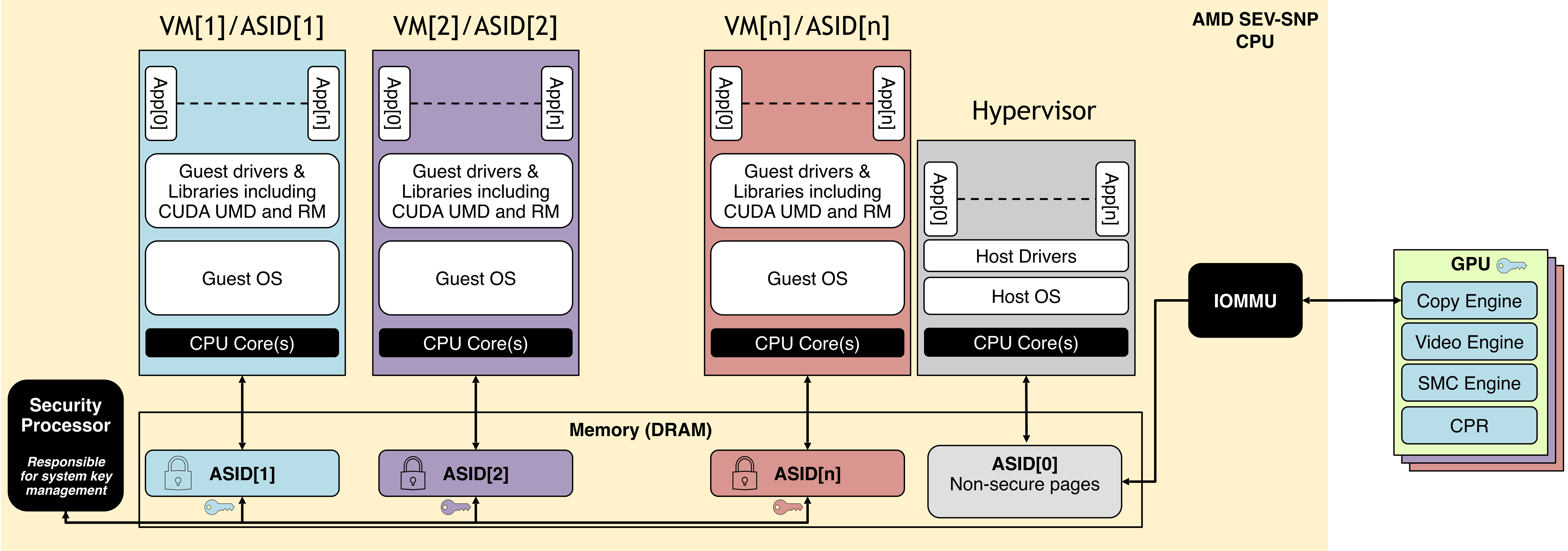
- cudaMallocManaged() is fully supported in Confidential Computing Mode
- CUDA APIs for allocating pinned system memory work with Confidential Computing:
  - cudaHostAlloc() # runtime;
  - cuMemHostAlloc() #driver
  - cudaMallocHost() # runtime;
  - cuMemMallocHost () #driver
- CUDA APIs for accessing CPU allocated memory (e.g., with malloc or new) are not possible with Confidential Computing:
  - cudaHostRegister() # runtime;
  - cuMemHostRegister() #driver
  - These APIs are rarely used, and where used can easily be substituted with the CUDA allocation APIs



CUDA API Details

Why Host Allocated Memory Needs Special Care

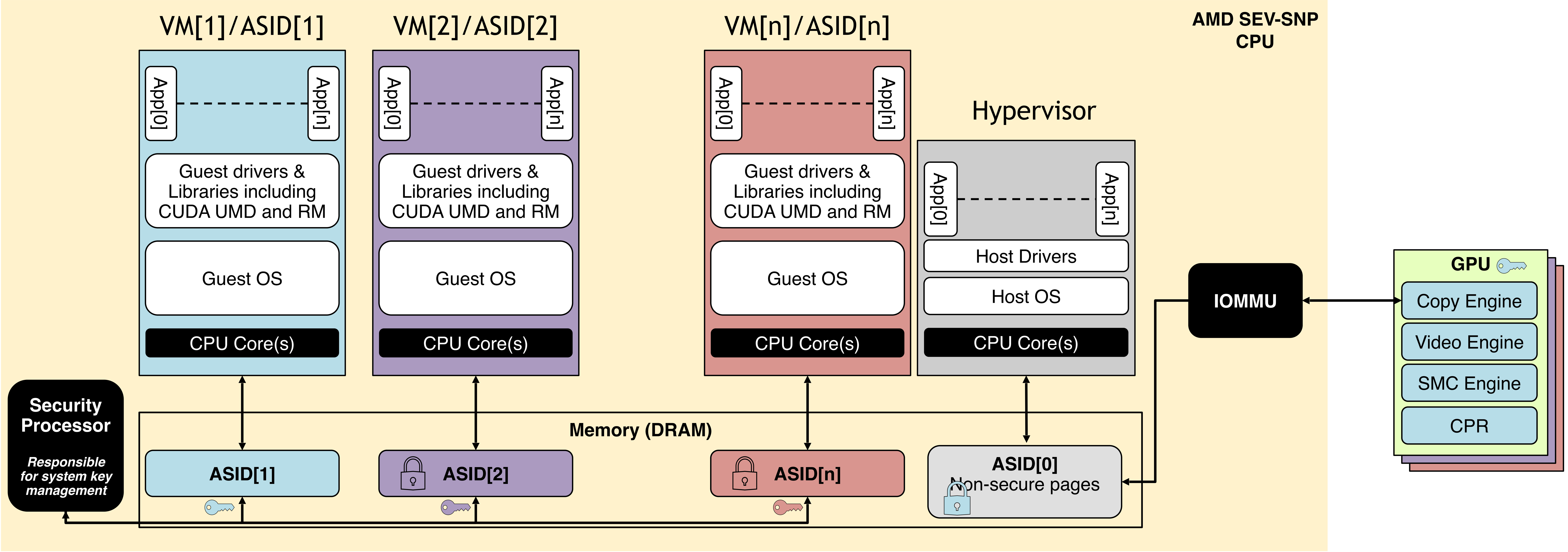
- In secure CPU systems, the IOMMU blocks accelerator access to secured VM memory
- In order for VMs to communicate with PCIe cards, they must send their data into non-secure memory
- Simply telling the GPU a VM's preallocated pointer will result in blocked accesses by the IOMMU
- UVM APIs will automatically handle Confidential copies to/from the "bounce buffer", transparent to the developer



CUDA API Details

Why Host Allocated Memory Needs Special Care

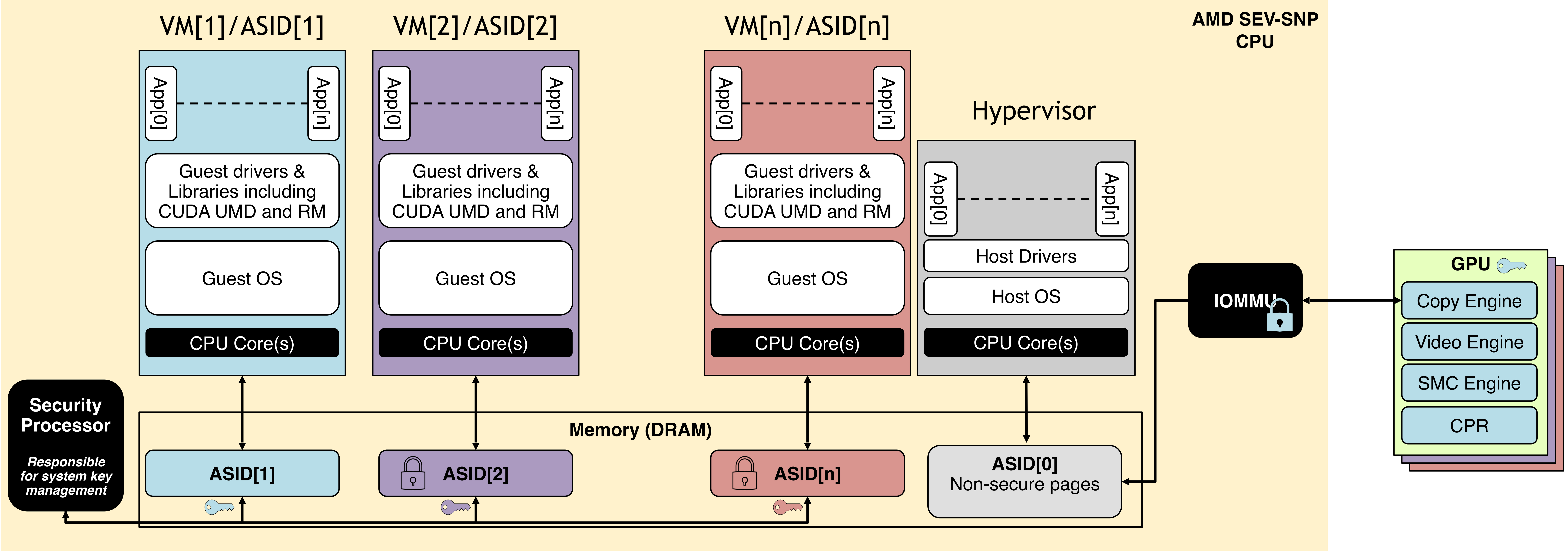
- In secure CPU systems, the IOMMU blocks accelerator access to secured VM memory
- In order for VMs to communicate with PCIe cards, they must send their data into non-secure memory
- Simply telling the GPU a VM's preallocated pointer will result in blocked accesses by the IOMMU
- UVM APIs will automatically handle Confidential copies to/from the "bounce buffer", transparent to the developer



CUDA API Details

Why Host Allocated Memory Needs Special Care

- In secure CPU systems, the IOMMU blocks accelerator access to secured VM memory
- In order for VMs to communicate with PCIe cards, they must send their data into non-secure memory
- Simply telling the GPU a VM's preallocated pointer will result in blocked accesses by the IOMMU
- UVM APIs will automatically handle Confidential copies to/from the "bounce buffer", transparent to the developer

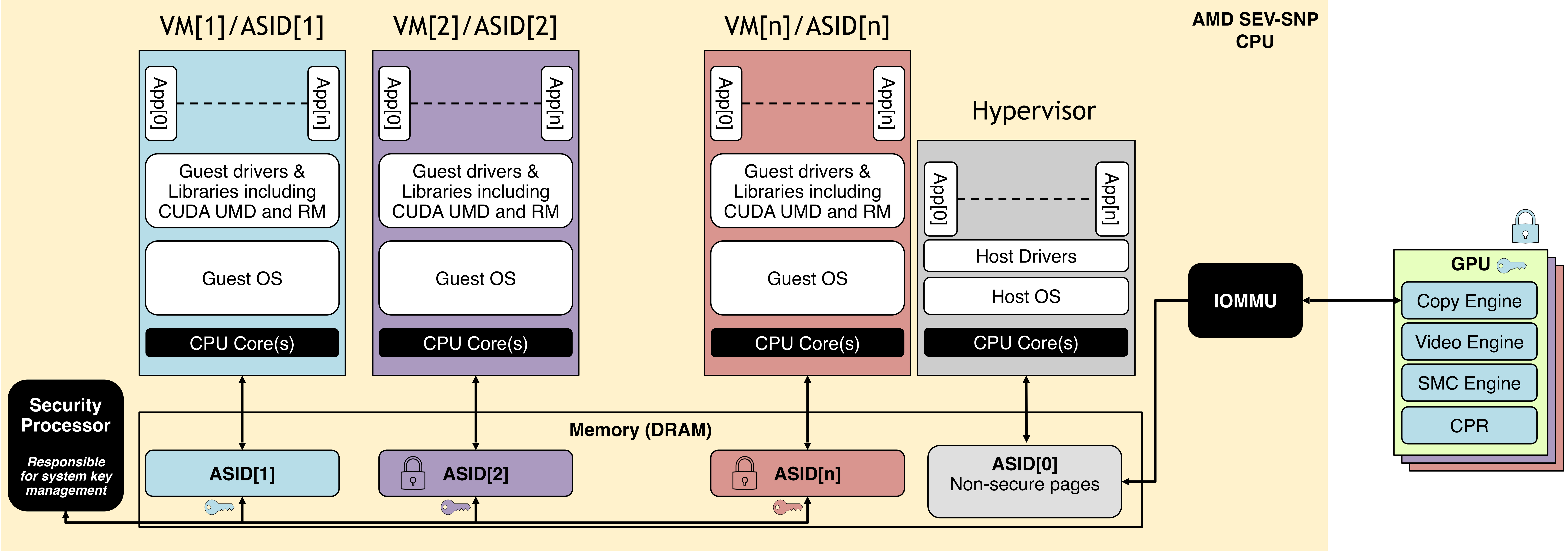


NVIDIA

CUDA API Details

Why Host Allocated Memory Needs Special Care

- In secure CPU systems, the IOMMU blocks accelerator access to secured VM memory
- In order for VMs to communicate with PCIe cards, they must send their data into non-secure memory
- Simply telling the GPU a VM's preallocated pointer will result in blocked accesses by the IOMMU
- UVM APIs will automatically handle Confidential copies to/from the "bounce buffer", transparent to the developer



Details on Other API considerations

Unsupported APIs with CC=on

- Confidential Computing on Hopper is Compute only
  - Graphics not supported
  - Therefore, compute-graphics interop APIs aren’t supported

- These APIs cannot work due to the security protections of the CPU TEE:
  - cuMemHostRegister/cudaHostRegister
  - cuStreamBatchMemOp[3]
  - cuStreamWaitValue[3]
  - cuStreamWriteValue[3]

Changed, but supported APIs

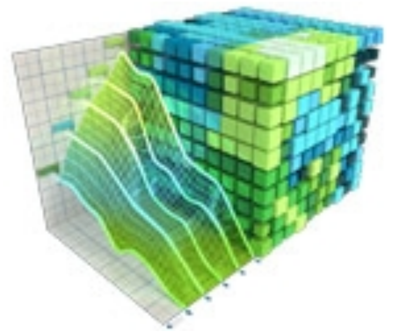
These APIs will Fall Back to UVM Based Calls

- cudaMallocHost -> cudaMallocManaged[1,2]
- cudaHostAlloc -> cudaMallocManaged[1,2]
- cudaFreeHost -> cudaFree

1 – Will prefer host memory

2 – Oversubscription of host memory vs. GPU allowed

Developer Tools Support

		Tool	Support Under CC
Modes		CC-Off – Standard H100 operation: No encryption, no bounce buffer	✓
		CC-On – All data/code is encrypted and authenticated. Firewalls prevent outside access	✗
		CC-DevTools – All data/code is encrypted and authenticated. Firewalls are dropped to enable tool access to performance counters	✓
Debugging		cuda-gdb - a seamless debugging experience that allows you to debug both the CPU and GPU	✓
		Nsight Visual Studio Code - application development environment for heterogeneous platforms	✓
Profiling		Nsight Systems - System-wide performance analysis tool	✓
		Nsight Compute* - Interactive profiler for CUDA and NVIDIA OptiX	✓
		CUPTI* - CUDA Profiling Tools Interface	✓
		NVTX - NVIDIA Tools Extension	✓
Sanitizer		Memcheck–- The memory access error and leak detection tool	✓
		Racecheck - The shared memory data access hazard detection tool	✗
		Initcheck -The uninitialized device global memory access detection tool.	✓
		Synccheck - The thread synchronization hazard detection tool.	✓

* Checkpoint/replay will not restore pinned memory under HCC

NVIDIA Libraries Working in Confidential Computing Mode

Library	Working in Hopper Confidential Compute Mode?
cuFFT	Yes
cuSPARSE	Yes
cuSPARSELt	Yes
cuBLAS	Yes
cuBLASLt	Yes
nvBLAS	Yes
Math API	Yes
NPP	Yes
nvJPEG	Yes
nvJPEG2000	Yes
nvTIFF	Yes
cuRAND	Yes
cuTENSOR	Yes
cuTensorNet	Yes
cuStateVec	Yes
cuSOLVER	In Progress

NVIDIA