Decentralized AI: Confidential Federated Learning With NVIDIA FLARE

Chester Chen

Senior Product & Engineering Manager

NVIDIA Federated Learning

Decentralization & AI Summit

August 06, 2024

Berkeley
UNIVERSITY OF CALIFORNIA

Powered by Zoom

Agenda

- NVIDIA FLARE Overview
- Confidential Federated Learning: Use Cases and Processes
- Confidential Computing: Tech Stack
- Enabling Confidential Federated Learning with NVIDIA FLARE

Berkeley

Powered by Zoom

NVIDIA

Agenda

- NVIDIA FLARE Overview
- Confidential Federated Learning: Use Cases and Processes
- Confidential Computing: Tech Stack
- Enabling Confidential Federated Learning with NVIDIA FLARE

Berkeley

Powered by Zoom

NVIDIA

NVIDIA FLARE Overview

Agenda

- NVIDIA FLARE Overview
- Confidential Federated Learning: Use Cases and Processes
- Confidential Computing: Tech Stack
- Enabling Confidential Federated Learning with NVIDIA FLARE

Berkeley

Powered by Zoom

NVIDIA FLARE Overview

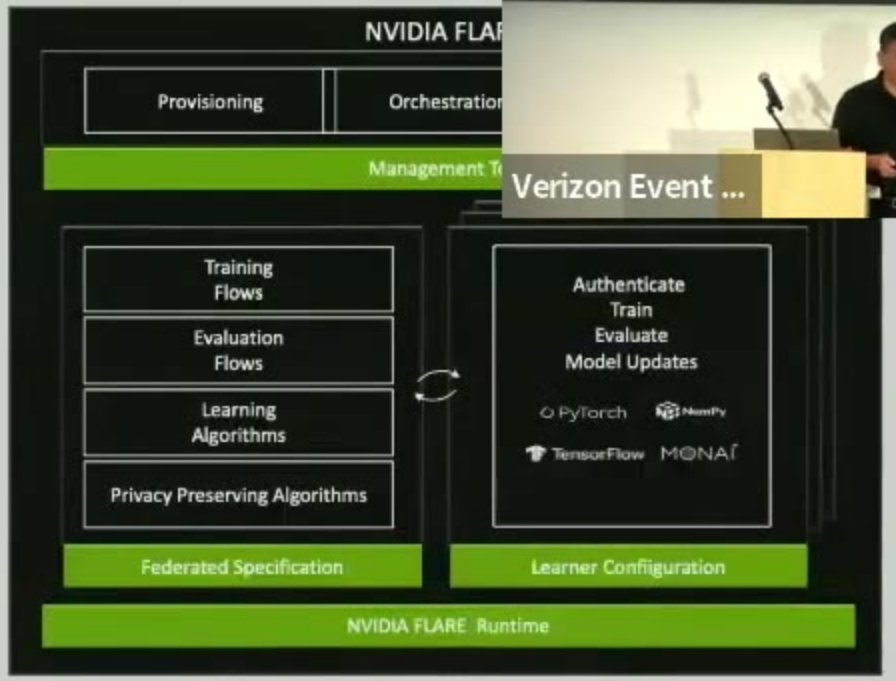
NVIDIA FLARE

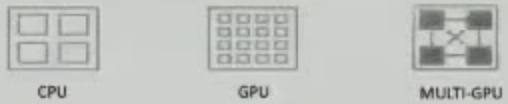
Open-Source, Enterprise Federated Learning & Compute Framework

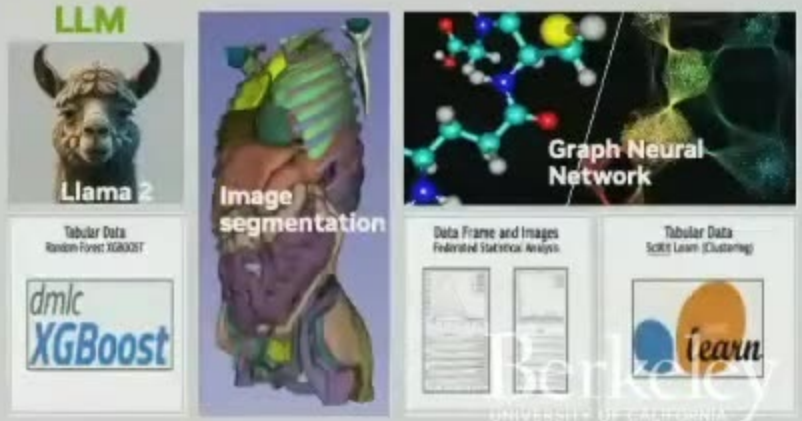
- **Apache License 2.0** to catalyze FL research & development
- **Designed for production**, not just for research
- **Enables cross-internet**, distributed, multi-party collaborative Learning
- **Production scalability** with high availability and **multi-task** execution
- **Easy to convert** existing ML/DL workflows to a Federated paradigm with few lines of code changes
- **LLM streaming, LLM fine tuning**
- **Framework, model, domain and task agnostic**
- **Privacy Preserving** Technologies
  - Homomorphic Encryption (HE), Differential Privacy (DP)
  - Multi-party computing (Private Set intersection, PSI)
  - Confidential Computing (CC)
- **Confidential FL**: end-to-end Federated Learning with Confidential Computing
- **Layered, pluggable, customizable** federated compute architecture
- Secure Provisioning, Orchestration & Monitoring

GitHub: https://github.com/nvidia/nvFlare

Web: https://nvidia.github.io/NVFlare/







Framework agnostic | Model agnostic | Domain agnostic | Task agnostic

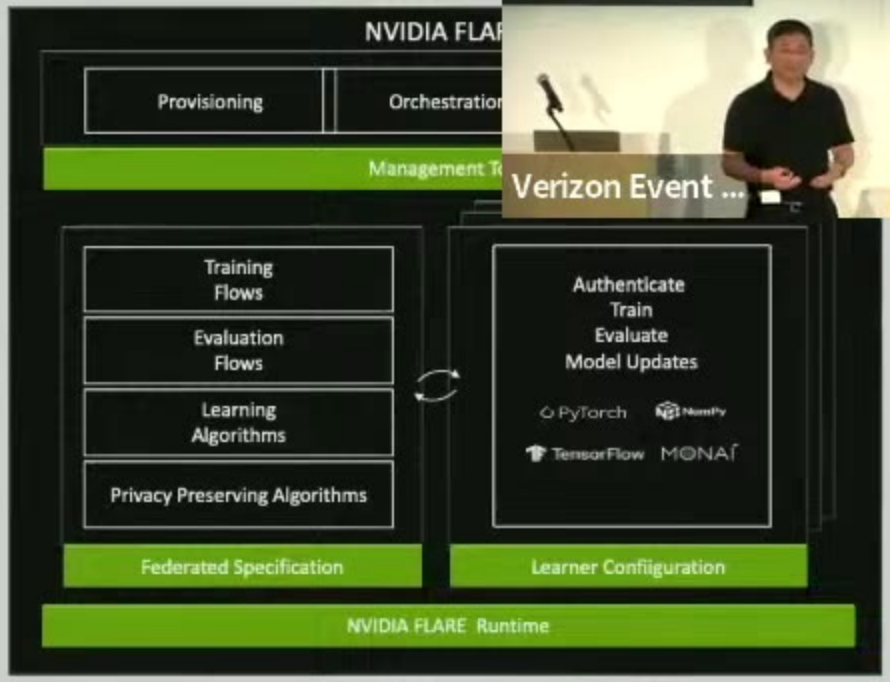
NVIDIA FLARE

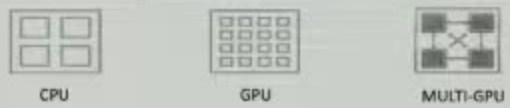
Open-Source, Enterprise Federated Learning & Compute Framework

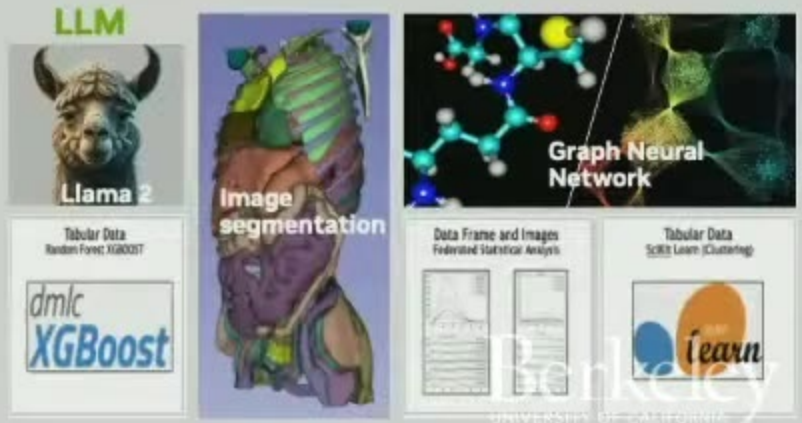
- **Apache License 2.0** to catalyze FL research & development
- **Designed for production**, not just for research
- **Enables cross-internet**, distributed, multi-party collaborative Learning
- **Production scalability** with high availability and **multi-task** execution
- **Easy to convert** existing ML/DL workflows to a Federated paradigm with few lines of code changes
- **LLM streaming, LLM fine tuning**
- **Framework, model, domain and task agnostic**
- **Privacy Preserving** Technologies
  - Homomorphic Encryption (HE), Differential Privacy (DP)
  - Multi-party computing (Private Set intersection, PSI)
  - Confidential Computing (CC)
- **Confidential FL**: end-to-end Federated Learning with Confidential Computing
- **Layered, pluggable, customizable** federated compute architecture
- Secure Provisioning, Orchestration & Monitoring

GitHub: https://github.com/nvidia/nvFlare

Web: https://nvidia.github.io/NVFlare/







Framework agnostic | Model agnostic | Domain agnostic | Task agnostic

NVIDIA FLARE Architecture

Federated Computing Engine

- **Layered, Pluggable Open Architecture**
  - Each layer's component are composable and pluggable
- **Network: Communication & Messaging layer**
  - Drivers ➔ gRPC, http + websocket, TCP, any plugin driver
  - CellNet: logical end point-to-point (cell to cell) network
  - Message: reliable streaming message
- **Federated Computing Layer**
  - Resource-based job scheduling, job monitoring, concurrent job lifecycle management, High-availability management
  - Plugin component management
  - Configuration management
  - Local event and federated event handling
- **Federated Workflow**
  - SAG, Cyclic, Cross-site Evaluation, Swarm Learning, Federated Analytics
- **Federated Learning Algorithms**
  - FedAvg, FedOpt, FedProx, Scalffold, Ditto, XGBoost, GNN, PSI, LLM (p-tuning, SFT, PEFT), KM, Scikit-Learn
- **Pythonic Programming APIs**
  - Client API, Controller API, Job Construction API, Job Monitoring API
- **Productivity & Deployment Tools:**
  - Simulator, provision, POC, Cloud deployment, preflight check, more

NVIDIA FLARE Architecture

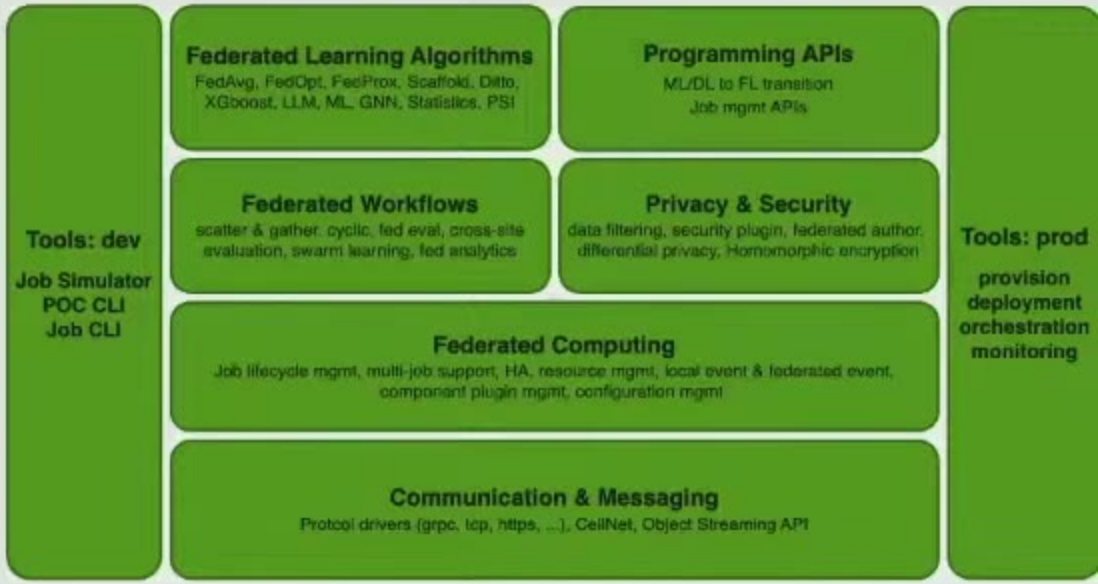
Federated Computing Engine

- **Layered, Pluggable Open Architecture**
  - Each layer's component are composable and pluggable
- **Network: Communication & Messaging layer**
  - Drivers ➔ gRPC, http + websocket, TCP, any plugin driver
  - CellNet: logical end point-to-point (cell to cell) network
  - Message: reliable streaming message
- **Federated Computing Layer**
  - Resource-based job scheduling, job monitoring, concurrent job lifecycle management, High-availability management
  - Plugin component management
  - Configuration management
  - Local event and federated event handling
- **Federated Workflow**
  - SAG, Cyclic, Cross-site Evaluation, Swarm Learning, Federated Analytics
- **Federated Learning Algorithms**
  - FedAvg, FedOpt, FedProx, Scalffold, Ditto, XGBoost, GNN, PSI, LLM (p-tuning, SFT, PEFT), KM, Scikit-Learn
- **Pythonic Programming APIs**
  - Client API, Controller API, Job Construction API, Job Monitoring API
- **Productivity & Deployment Tools:**
  - Simulator, provision, POC, Cloud deployment, preflight check, more

NVIDIA FLARE Architecture

Federated Computing Engine

- **Layered, Pluggable Open Architecture**
  - Each layer's component are composable and pluggable
- **Network: Communication & Messaging layer**
  - Drivers ➔ gRPC, http + websocket, TCP, any plugin driver
  - CellNet: logical end point-to-point (cell to cell) network
  - Message: reliable streaming message
- **Federated Computing Layer**
  - Resource-based job scheduling, job monitoring, concurrent job lifecycle management, High-availability management
  - Plugin component management
  - Configuration management
  - Local event and federated event handling
- **Federated Workflow**
  - SAG, Cyclic, Cross-site Evaluation, Swarm Learning, Federated Analytics
- **Federated Learning Algorithms**
  - FedAvg, FedOpt, FedProx, Scalffold, Ditto, XGBoost, GNN, PSI, LLM (p-tuning, SFT, PEFT), KM, Scikit-Learn
- **Pythonic Programming APIs**
  - Client API, Controller API, Job Construction API, Job Monitoring API
- **Productivity & Deployment Tools:**
  - Simulator, provision, POC, Cloud deployment, preflight check, more



NVIDIA FLARE: Summary

A domain-agnostic, open-source, extensible FL framework

- **Federated Computing** -- a federated computing framework at core
- **Built for productivity** -- designed for maximum productivity, providing a range of tools to enhance user experience
- **Built for security & privacy** -- prioritizes robust security and privacy preservation
- **Built for concurrency & scalability** -- designed for concurrency, supporting resource-based multi-job execution
- **Built for customization** -- structured in layers, with each layer composed of customizable components
- **Built for integration** -- multiple integration options with third-party system
- **Built for production** -- robust, production-scale deployment in real-world federated learning and computing scenarios
- **Rich examples repository** -- wealth of built-in implementations, tutorials and examples
- **Growing application categories --** medical imaging, medical devices, edge device application, financial services, HPC and autonomous driving vehicles

GitHub : https://github.com/NVIDIA/NVFlare

Web: https://nvidia.github.io/NVFlare/

NVIDIA FLARE: Summary

A domain-agnostic, open-source, extensible FL framework

- **Federated Computing** -- a federated computing framework at core
- **Built for productivity** -- designed for maximum productivity, providing a range of tools to enhance user experience
- **Built for security & privacy** -- prioritizes robust security and privacy preservation
- **Built for concurrency & scalability** -- designed for concurrency, supporting resource-based multi-job execution
- **Built for customization** -- structured in layers, with each layer composed of customizable components
- **Built for integration** -- multiple integration options with third-party system
- **Built for production** -- robust, production-scale deployment in real-world federated learning and computing scenarios
- **Rich examples repository** -- wealth of built-in implementations, tutorials and examples
- **Growing application categories --** medical imaging, medical devices, edge device application, financial services, HPC and autonomous driving vehicles

GitHub : https://github.com/NVIDIA/NVFlare

Web: https://nvidia.github.io/NVFlare/

NVIDIA FLARE: Summary

A domain-agnostic, open-source, extensible FL framework

- **Federated Computing** -- a federated computing framework at core
- **Built for productivity** -- designed for maximum productivity, providing a range of tools to enhance user experience
- **Built for security & privacy** -- prioritizes robust security and privacy preservation
- **Built for concurrency & scalability** -- designed for concurrency, supporting resource-based multi-job execution
- **Built for customization** -- structured in layers, with each layer composed of customizable components
- **Built for integration** -- multiple integration options with third-party system
- **Built for production** -- robust, production-scale deployment in real-world federated learning and computing scenarios
- **Rich examples repository** -- wealth of built-in implementations, tutorials and examples
- **Growing application categories --** medical imaging, medical devices, edge device application, financial services, HPC and autonomous driving vehicles

GitHub : https://github.com/NVIDIA/NVFlare

Web: https://nvidia.github.io/NVFlare/

Confidential Federated Learning: Use Cases and Processes

Confidential Computing: Use Cases

CC Use Cases across industries

Government	Healthcare	Financial Services	Manufacturing	Enterprise	Any Industry
Security Risk	Patient Data Privacy	Fraud Detection, AML	Supply Chain Analysis	Multinational HR Analysis	Data Clean Room
cross-agency, or cross-country multi-party collaboration	cross hospitals and corss-institutional research. Federated AI and multi-party collaboration is needed	Multi-Party PSI, Collaborate data sharing, Federated AI training and Inference	Enforce Quality Control Procedures requires federated AI and multi-party collaboration	Protect sensitive data while perform analytics. Require federated AI	securely share data for data cleaning, training and analytics

Federated Learning Use Cases

Requirements

- Prevent Personal Information Leak
- Prevent data Leak
- Protect Model IP

FL Use Cases that requires CC

- Build **Explicit Trust**
- **Prevent** code, model, data **tampering**
- **Secure Aggregation** at Server Node
  - Secure aggregation node
  - Aggregation code protection
- **Secure Training** at Client Node
  - Training node protection with TEE
  - Model IP protection with TEE
  - Prevent data leak
- **Federated Inference Protection**
  - Input data protection
  - Model protection

Federated Learning Use Cases

Requirements

- Prevent Personal Information Leak
- Prevent data Leak
- Protect Model IP

FL Use Cases that requires CC

- Build **Explicit Trust**
- **Prevent** code, model, data **tampering**
- **Secure Aggregation** at Server Node
  - Secure aggregation node
  - Aggregation code protection
- **Secure Training** at Client Node
  - Training node protection with TEE
  - Model IP protection with TEE
  - Prevent data leak
- **Federated Inference Protection**
  - Input data protection
  - Model protection

Federated Learning Use Cases

Requirements

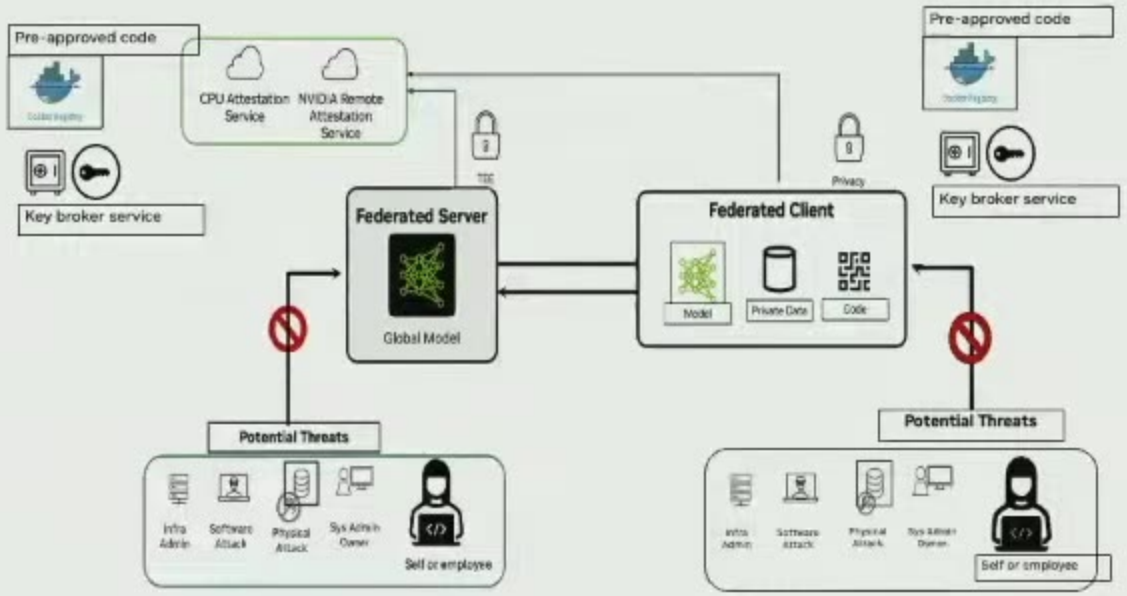
- Prevent Personal Information Leak
- Prevent data Leak
- Protect Model IP

FL Use Cases that requires CC

- Build **Explicit Trust**
- **Prevent** code, model, data **tampering**
- **Secure Aggregation** at Server Node
  - Secure aggregation node
  - Aggregation code protection
- **Secure Training** at Client Node
  - Training node protection with TEE
  - Model IP protection with TEE
  - Prevent data leak
- **Federated Inference Protection**
  - Input data protection
  - Model protection

Confidential Federated Learning: Processes

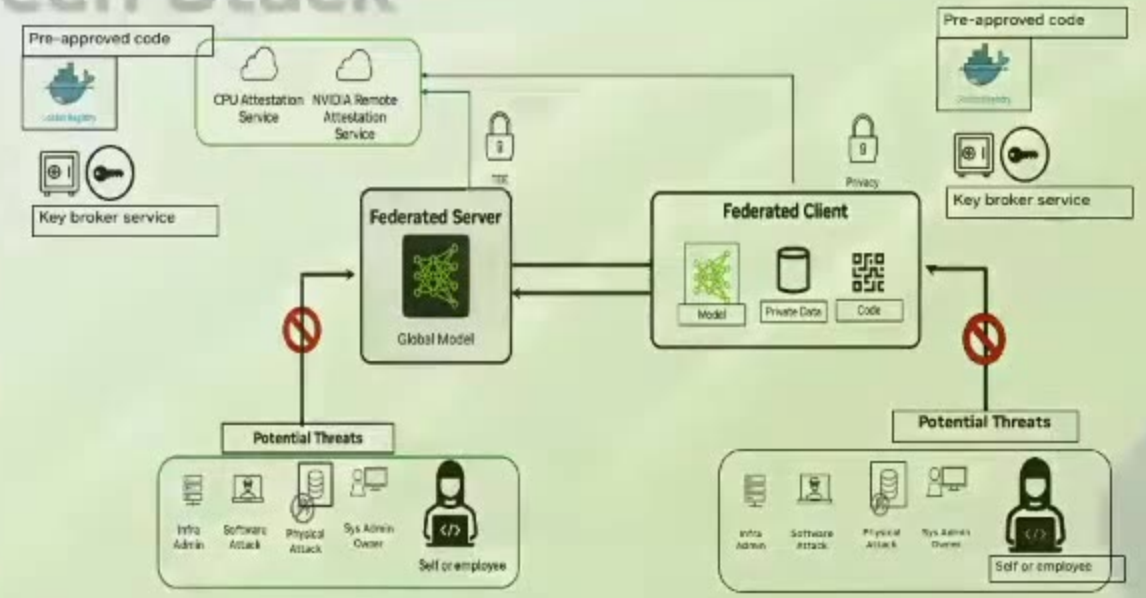
- **Attestations**
  - Federated Learning Requests multi-SDK attestation
  - FL Servers needs to verify all client's trustworthiness
  - Attestation at different points, self and cross verifications via attestation service
- **CC Policies:**
  - Bootup policy – provided by hardware vendor
  - Self-verification CC policy – user defined
  - Cross-verification CC policy -- user defined
- **Protected Assets**
  - **Code:** Training code (client), aggregation code (server)
  - **Data:** input data
  - **Model**: initial model, intermediate model, output model
  - **Workspace**: check points, logs, temp data, output directory
- **Key Broker Service**
  - Key management depends on user case, global model ownership, key release management process
- **Bootstrap**:
  - Need a process to generate the keys, policies and input them into key-vault to avoid tempering



Confidential Federated Learning: Processes

- **Attestations**
  - Federated Learning Requests multi-SDK attestation
  - FL Servers needs to verify all client's trustworthiness
  - Attestation at different points, self and cross verifications via attestation service
- **CC Policies:**
  - Bootup policy – provided by hardware vendor
  - Self-verification CC policy – user defined
  - Cross-verification CC policy -- user defined
- **Protected Assets**
  - **Code:** Training code (client), aggregation code (server)
  - **Data:** input data
  - **Model**: initial model, intermediate model, output model
  - **Workspace**: check points, logs, temp data, output directory
- **Key Broker Service**
  - Key management depends on user case, global model ownership, key release management process
- **Bootstrap**:
  - Need a process to generate the keys, policies and input them into key-vault to avoid tempering

Confidential Computing Tech Stack



Powered by Zoom

Confidential Computing Tech Stack

NVIDIA Confidential Computing Introduction

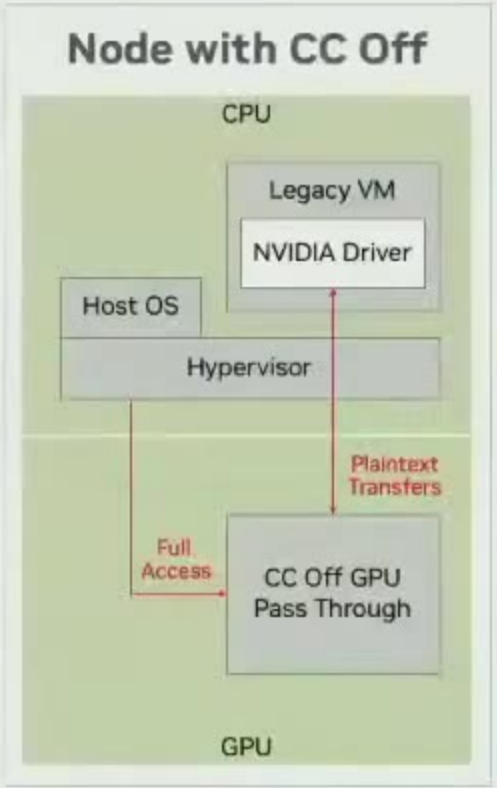
Protecting Data and Code from Hypervisor and Physical Attacks

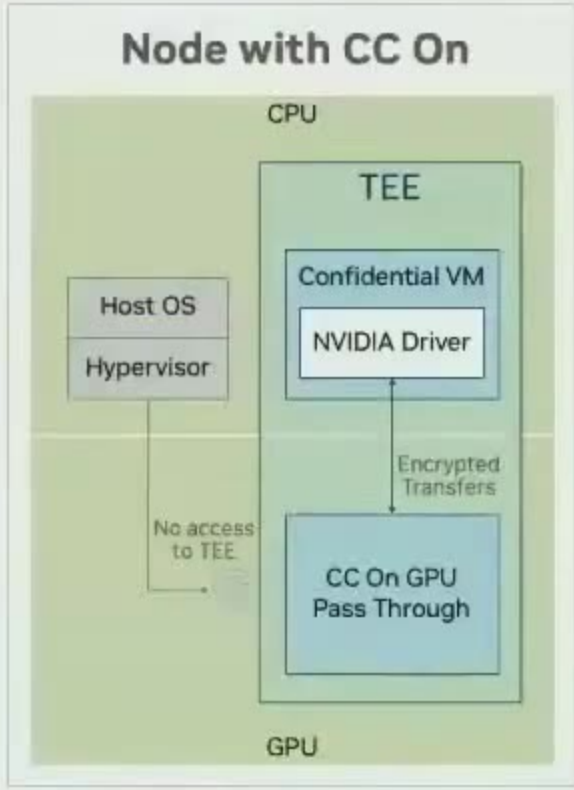
Prerequisites:

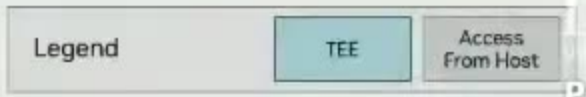
- CPU with support for a Virtualized-based TEE ("Confidential VM")
- Supported variants are AMD Milan or later, or Intel SPR and later.

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







Confidential Computing Tech Stack

Platform	Pipeline Mgmt	Security Mgmt	Project Mgmt	User Mgmt	Data Mgmt
Orchestration	Cluster Mgmt.	Storage Mgmt.	Monitoring Service		
AI/ML Runtime	NVFLARE				
Key Broker Service	Vendor Key Broker Services	CoCo Trustee	... more ...		
Attestation SDK	CPU Attestation SDKs	GPU Attestation Services	Cloud CSP or Vendor Attestation Services		
Virtualization	Confidential VM	Confidential Container (CoCo) / K8s	Confidential Container	Other Approaches	
Operating System	Ubuntu 22.04				
Hardware	CPU Confidential Computing Technology	NVIDIA Hopper, Blackwell			

Confidential Computing Tech Stack

Platform	Pipeline Mgmt	Security Mgmt	Project Mgmt	User Mgmt	Data Mgmt
Orchestration	Cluster Mgmt.	Storage Mgmt.	Monitoring Service		
AI/ML Runtime	NVFLARE				
Key Broker Service	Vendor Key Broker Services	CoCo Trustee	... more ...		
Attestation SDK	CPU Attestation SDKs	GPU Attestation Services	Cloud CSP or Vendor Attestation Services		
Virtualization	Confidential VM	Confidential Container (CoCo) / K8s	Confidential Container	Other Approaches	
Operating System	Ubuntu 22.04				
Hardware	CPU Confidential Computing Technology	NVIDIA Hopper, Blackwell			

Confidential Computing Tech Stack

Enabling Confidential Federated Learning with NVIDIA FLARE

Platform	Pipeline Mgmt	Security Mgmt	Project Mgmt	User Mgmt	Data Mgmt
Orchestration	Cluster Mgmt.	Storage Mgmt.	Monitoring Service		
AI/ML Runtime	NVFLARE				
Key Broker Service	Vendor Key Broker Services	CoCo Trustee	... more ...		
Attestation SDK	CPU Attestation SDKs	GPU Attestation Services	Cloud CSP or Vendor Attestation Services		
Virtualization	Confidential VM	Confidential Container (CoCo) / K8s	Confidential Container	Other Approaches	
Operating System	Ubuntu 22.04				
Hardware	CPU Confidential Computing Technology	NVIDIA Hopper, Blackwell			

Berkeley UNIVERSITY OF CALIFORNIA

Powered by Zoom

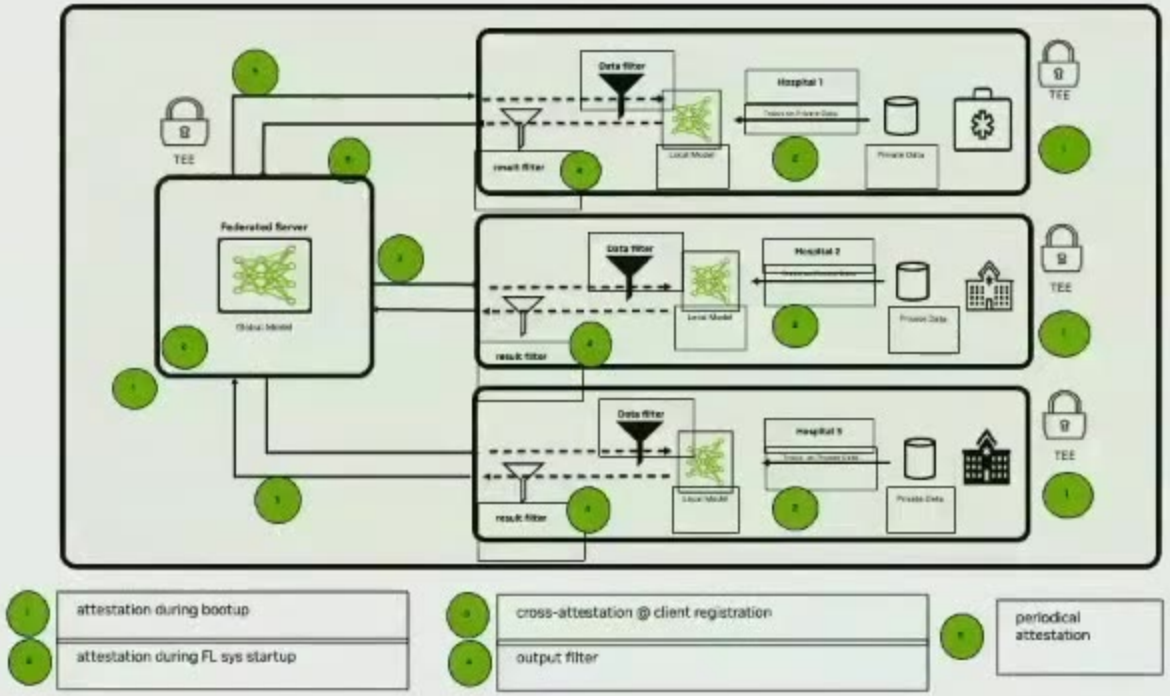
NVIDIA

Enabling Confidential Federated Learning with NVIDIA FLARE

Berkeley
UNIVERSITY OF CALIFORNIA

NVIDIA FLARE + Confidential Computing Integration

- **CC Enables the "Lift & Shift" Capability**
  - Existing application don't need to be modified to shift from non-TEE to TEE env with new hardware-based protection
- **Build Explicit Trust**
  - Attestation Service Integration
    - Different CPU/GPU attestations SDKs
  - Design to verify the trust worthiness with CC attestation service
    - Self-Test at start
    - Cross-verification at client registration
    - Repeat attestation tests periodically
  - Confidential VM
    - Bare Metal CVM, CSP CVM
  - Confidential Containers (**CoCo**) on K8s
    - SSH lockdown
    - Require additional Trustee services features



Secure Aggregation: NVIDIA FLARE with CC

NVIDIA FLARE + Azure Confidential Computing

Infrastructure setup: Hybrid deployment

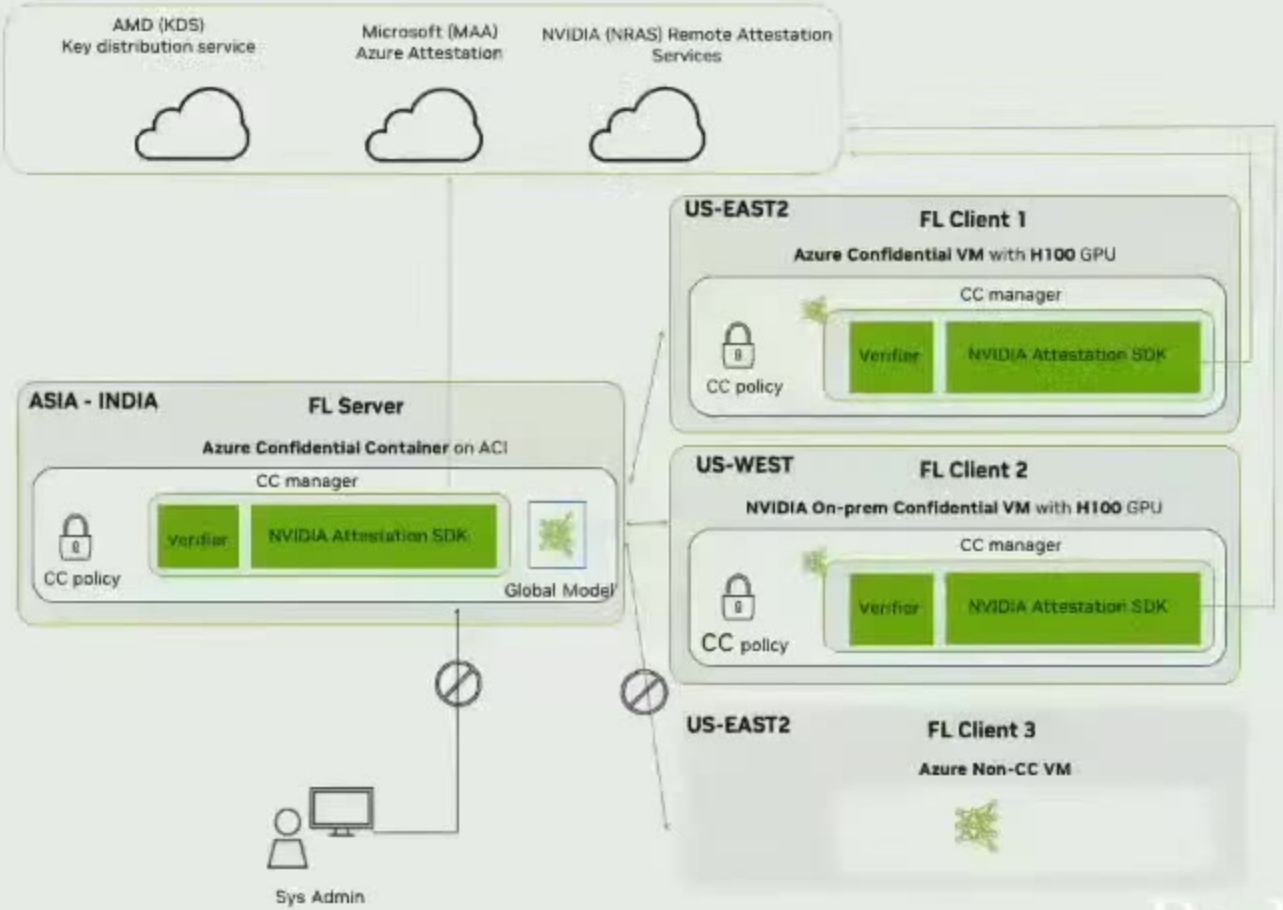
- **FL Server:** Azure **Confidential Container** on **ACI**
  - AMD CPU
- **FL Client 1:** Azure **Confidential VM**
  - AMD CPU + NVIDIA H100
- **FL Client 2:** On-Prem **Confidential VM**
  - AMD CPU + NVIDIA H100
- **FL Client 3:** Azure Non-CC instance (CPU)

CC Summit SF

- Medical Imaging: Spleen 3D CT segmentation with MONAI bundle

User Benefits:

- Available on Azure Cloud
- Any FL application written in FLARE could executed without any change
- End-to-end security enforcement and secure aggregation for all federated learning applications
- No one can SSH into FL server
- Failure of attestation verification will cause the job to fail or system shutdown



Note:
This is part of the join presentation with Azure Team at CC Summit at SF, 2024.
We also had a similar presentation at using XGBoost model at GTC 2024

Get Involved: Upcoming Events

Just us at

NVIDIA FLARE DAY 2024:

Exploring Real-world Examples of Federated Learning

Online Event

September 18, 2024

NVIDIA FLARE DAY 2024 : https://nvidiaflareday.splashthat.com/

NVLARE Github : https://github.com/NVIDIA/NVFlare

Web Page : https://nvidia.github.io/NVFlare/

RFP: NVIDIA Academic Grant Program Federated Learning with FLARE

Topic available in Fall

Berkeley Center for Decentraliz

Keynote Speaker

Building an Open, Responsible AI Economy

Dawn Song
Professor
UC Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA

Powered by Zoom