
NVIDIA FEDERATED LEARNING

BUILDING AI FOR REAL-WORLD CLINICAL PERFORMANCE

Taking Algorithms Beyond Proof-of-Concept

REAL-WORLD AI DESIGN

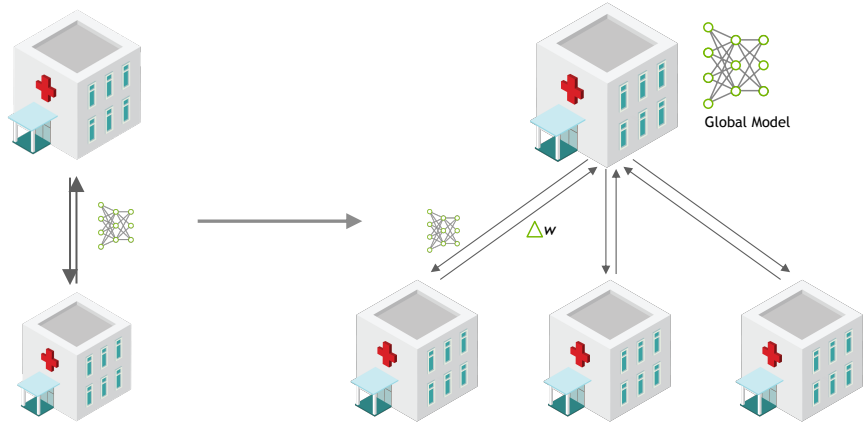
External Validation, Multiple Institutions, Prospective Data

Design Characteristic	All Articles (n = 516)	Articles Published in Medical Journals (n = 437)
External validation		
Used	31 (6.0)	27 (6.2)
Not used	485 (94.0)	410 (93.8)
In studies that used external validation		
Diagnostic cohort design	5 (1.0)	5 (1.1)
Data from multiple institutions	15 (2.9)	12 (2.7)
Prospective data collection	4 (0.8)	4 (0.9)
Fulfillment of all of above three criteria	0 (0)	0 (0)
Fulfillment of at least two criteria	3 (0.6)	3 (0.7)
Fulfillment of at least one criterion	21 (4.1)	18 (4.1)

Only 6% of published AI studies have external validation
Few included multiple institutions

FEDERATED LEARNING PARADIGM

Model to Data | Generalize Model



Transfer Learning
“Adapt”

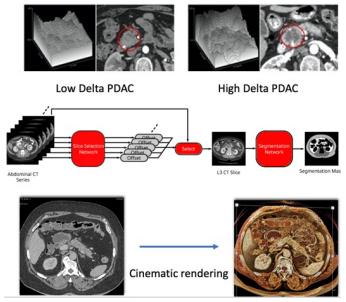
Federated Learning
“Generalize”

Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol. 2019 Mar;20(3):405-410. doi: 10.3348/kjr.2019.0025. PMID: 30799571; PMCID: PMC6389801.

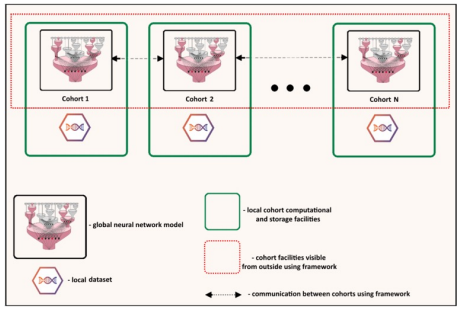
FEDERATED LEARNING MOMENTUM

Breaking Healthcare Data Siloes

EDRN
Early Detection of Pancreatic Cancer



ERASMUS GENNET
Genome Wide Association Study



EXAM
COVID-19 Oxygen Requirement Prediction

nature medicine
ARTICLES
https://doi.org/10.1038/s41591-021-01506-3

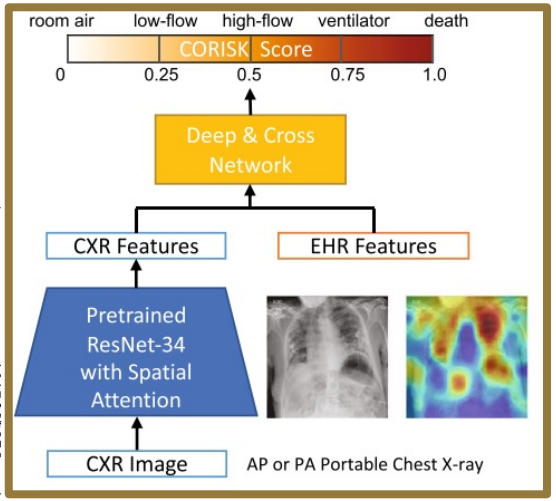
Federated learning for predicting clinical outcomes in patients with COVID-19

Ittai Dayan[1,56], Holger R. Roth[2,56], Aoxiao Zhong[3,4,56], Ahmed Harouni[2], Amilcare Gentili[5], Anas Z. Abidin[2], Andrew Liu[2], Anthony Beardsworth Costa[6], Bradford J. Wood[7,8], Chien-Sung Tsai[9], Chih-Hung Wang[10,11], Chun-Nan Hsu[12], C. K. Lee[2], Peiying Ruan[2], Daguang Xu[2], Dufan Wu[3], Eddie Huang[2], Felipe Campos Kitamura[13], Griffin Lacey[2], Gustavo César de Antônio Corradi[13], Gustavo Nino[14], Hao-Hsin Shin[15], Hirofumi Obinata[16], Hui Ren[3], Jason C. Crane[17], Jesse Tetreault[2], Jiahui Guan[2], John W. Garrett[18], Joshua D. Kaggie[19], Jung Gil Park[20], Keith Dreyer[1,21], Krishna Juluru[15], Kristopher Kersten[2], Marcio Aloisio Bezerra Cavalcanti Rockenbach[21], Marius George Linguraru[22,23], Masoom A. Haider[24,25], Meena AbdelMaseeh[25], Nicola Rieke[2], Pablo F. Damasceno[17], Pedro Mario Cruz e Silva[2], Pochuan Wang[26,27], Sheng Xu[7,8], Shuichi Kawano[16], Sira Sriswasdi[28,29], Soo Young Park[30], Thomas M. Grist[31], Varun Buch[21], Watsamon Jantarabenjakul[32,33], Weichung Wang[26,27], Won Young Tak[30], Xiang Li[3], Xihong Lin[3,34], Young Joon Kwon[6], Abood Quraini[2], Andrew Feng[2], Andrew N. Priest[35], Baris Turkbey[4,36], Benjamin Glicksberg[37], Bernardo Bizzo[21], Byung Seok Kim[38], Carlos Tor-Díez[22], Chia-Cheng Lee[39], Chia-Jung Hsu[39], Chin Lin[40,41,42], Chiu-Ling Lai[43], Christopher P. Hess[17], Colin Compas[2], Deepeksha Bhatia[2], Eric K. Oermann[44], Evan Leibovitz[21], Hisashi Sasaki[16], Hitoshi Mori[16], Isaac Yang[2], Jae Ho Sohn[17], Krishna Nand Keshava Murthy[15], Li-Chen Fu[45], Matheus Ribeiro Furtado de Mendonça[13], Mike Fralick[46], Min Kyu Kang[20], Mohammad Adil[2], Natalie Gangai[15], Peerapon Vateekul[47], Pierre Elnajjar[15], Sarah Hickman[19], Sharmila Majumdar[17], Shelley L. McLeod[48,49], Sheridan Reed[7,8], Stefan Gräf[50], Stephanie Harmon[4,51], Tatsuya Kodama[16], Thanyawee Puthanakit[32,33], Tony Mazzulli[52,53,54], Vitor Lima de Lavor[13], Yothin Rakvongthai[55], Yu Rim Lee[30], Yuhong Wen[2], Fiona J. Gilbert[19,56], Mona G. Flores[2,56]✉ and Quanzheng Li[3,56]

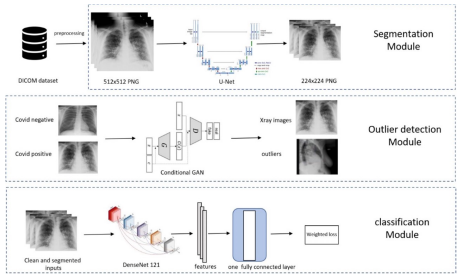
Federated learning (FL) is a method used for training artificial intelligence models with data from multiple sources while maintaining data anonymity, thus removing many barriers to data sharing. Here we used data from 20 institutes across the globe to train a FL model, called EXAM (electronic medical record (EMR) chest X-ray AI model), that predicts the future oxygen requirements of symptomatic patients with COVID-19 using inputs of vital signs, laboratory data and chest X-rays. EXAM achieved an average area under the curve (AUC) >0.92 for predicting outcomes at 24 and 72 h from the time of initial presentation to the emergency room, and it provided 16% improvement in average AUC measured across all participating sites and an average increase in generalizability of 38% when compared with models trained at a single site using that site's data. For prediction of mechanical ventilation treatment or death at 24 h at the largest independent test site, EXAM achieved a sensitivity of 0.950 and specificity of 0.882. In this study, FL facilitated rapid data science collaboration without data exchange and generated a model that generalized across heterogeneous, unharmonized datasets for prediction of clinical outcomes in patients with COVID-19, setting the stage for the broader use of FL in healthcare.

The scientific, academic, medical and data science communities have come together in the face of the COVID-19 pandemic crisis to rapidly assess novel paradigms in artificial intelligence (AI) that are rapid and secure, and potentially incentivize data sharing and model training and testing without the usual privacy and data ownership hurdles of conventional collaborations[1,2]. Healthcare providers, researchers and industry have pivoted their focus to address unmet and critical clinical needs created by the crisis, with remarkable results[3–11]. Clinical trial recruitment has been expedited and facilitated by national regulatory bodies and an international cooperative spirit[12–14]. The data analytics and AI disciplines have always fostered open

A full list of affiliations appears at the end of the paper.



U MINNESOTA, FAIRVIEW
X-RAY Covid-19 Classification



MELLODDY
Multi-task Learning Chemical Assays

PHARMA PARTNERS

AMGEN · astellas · AstraZeneca · BAYER · Boehringer Ingelheim · gsk · janssen · MERCK · NOVARTIS · SERVIER

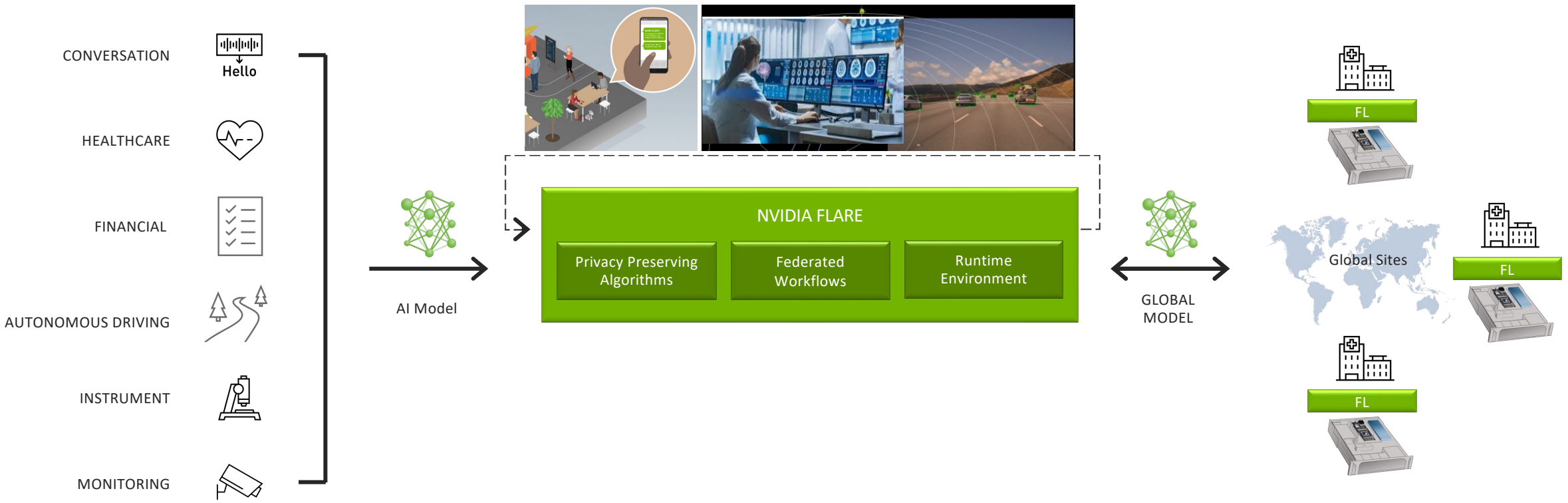
PUBLIC PARTNERS

MOLGENSYSTEM 1782 · IKTOS · KU LEUVEN · loodse · NVIDIA · OWKIN · SUBSTRA

NVIDIA

NVIDIA FEDERATED LEARNING

Applications across industries

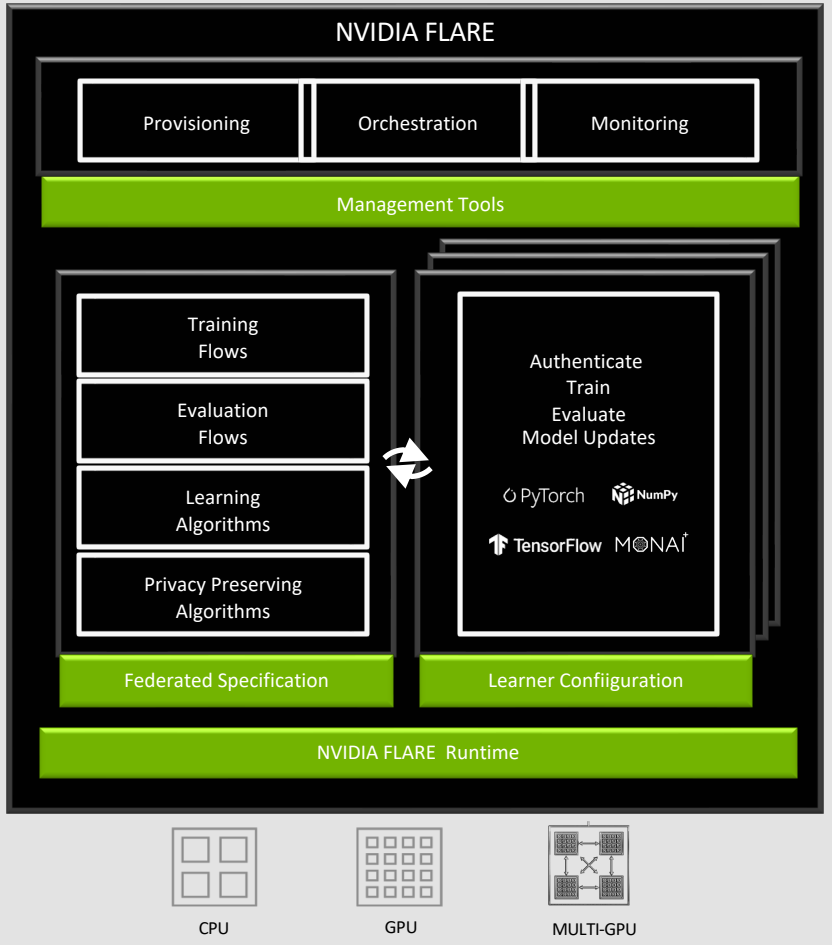


NVIDIA FLARE

Open-Source SDK for Federated Learning

- Apache License 2.0 to catalyze FL research & development
- Enables Distributed, Multi-Party Collaborative Learning
- Production Scalability with high availability and multi-task execution
- Adapt existing ML/DL workflows to a Federated paradigm
- Privacy Preserving Algorithms
  - Homomorphic Encryption & Differential Privacy
- Secure Provisioning, Orchestration & Monitoring
- Programmable APIs for Extensibility

Available on Github: https://github.com/nvidia/nvFlare



PERSONAS (WHO & VALUE PROP FOR EACH)

FL RESEARCHERS

Enables ease of getting started with FL experiments execution & evaluation in real world.

Extensible APIs for ease of creating custom implementations for new federated workflows, learning & privacy preserving algorithms.

DATA SCIENTISTS

Extend existing DL/ML workflows with a Federated paradigm and explore potential of Federated learning.

Ready to use FL specification and management tools enabling seamless execution.

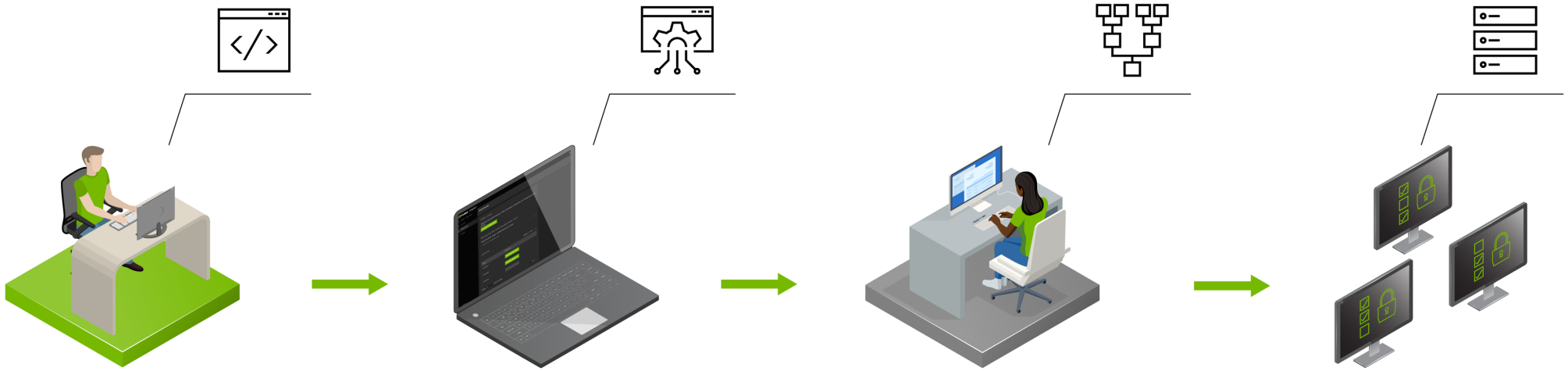
PLATFORM DEVELOPERS

A robust, extensible foundation to customize a platform offering for end users.

Built-in implementations of Federated learning spec & Aux APIs to build custom offerings.

NVIDIA FLARE Workflow

From rapid research prototyping to streamlined real world deployment



Rapid Development

Streamlined Deployment

Simplified Operations

Real-World Site Security

FEDERATED 2.2 NEW FEATURES

From Research Simulation to Real World Deployment

FL SIMULATOR

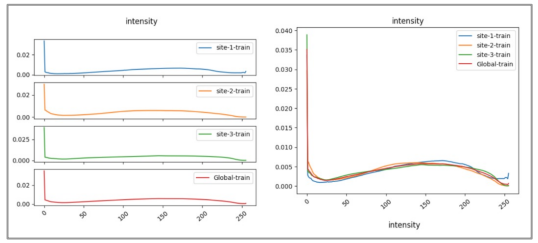
Rapid Development and Debugging

```
def run_simulator(simulator_args):
    simulator = SimulatorRunner(
        job_folder=simulator_args.job_folder,
        workspace=simulator_args.workspace,
        clients=simulator_args.clients,
        n_clients=simulator_args.n_clients,
        threads=simulator_args.threads,
        gpu=simulator_args.gpu,
        max_clients=simulator_args.max_clients,
    )
    run_status = simulator.run()

    return run_status
```

FEDERATED STATS

Analyze data distributions



FRAMEWORKS INTEGRATION

MONAI & XGBOOST

dmlc XGBoost

MONAI+

FLARE DASHBOARD

Streamlined operation & deployment

UNIFIED CLI

Multi-task Learning Chemical Assays

CLIENT CONTROLLED PP

Genome Wide Association Study

NVIDIA FLARE KEY CAPABILITIES

Runtime-ready and extensible suite of features

Privacy-Preserving Algorithms

NVIDIA FLARE provides privacy-preserving algorithms that ensure each change to the global model stays hidden and prevent the server from reverse-engineering the submitted weights and discovering any training data.

Training and Evaluation Workflows

Built-in workflow paradigms use local and decentralized data to keep models relevant at the edge, including learning algorithms for FedAvg, FedOpt, and FedProx.

Extensible Management Tools

Management tools help secure provisioning using SSL certifications, orchestration through an admin console, and monitoring of federated learning experiments using TensorBoard for visualization.

Supports Popular ML/DL Frameworks

Flexible in design, the SDK can be used with PyTorch, Tensorflow, and even Numpy, which allows for integrating federated learning into your current workflow.

Extensive API

Its extensive and open-source API enables researchers to develop new federated workflow strategies, innovative learning, and privacy-preserving algorithms.

Reusable Building Blocks

NVIDIA FLARE provides an easy way to perform federated learning experiments by utilizing the reusable building blocks and example walkthroughs.

https://developer.nvidia.com/flare

SECURITY & PRIVACY

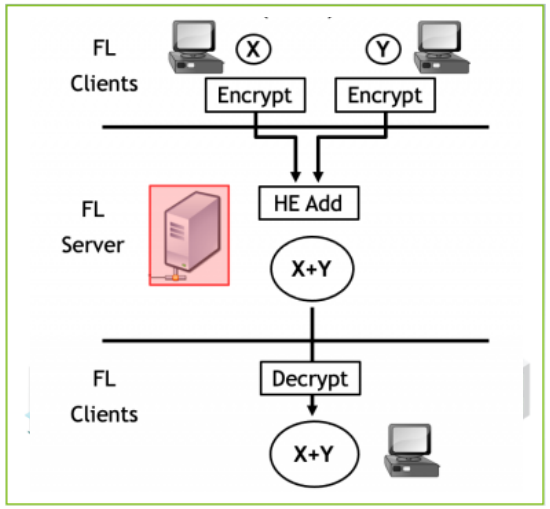
Homomorphic Encryption & Differential Privacy

Federated Learning with Homomorphic Encryption

What if I don't trust the server?

Homomorphic encryption (HE)

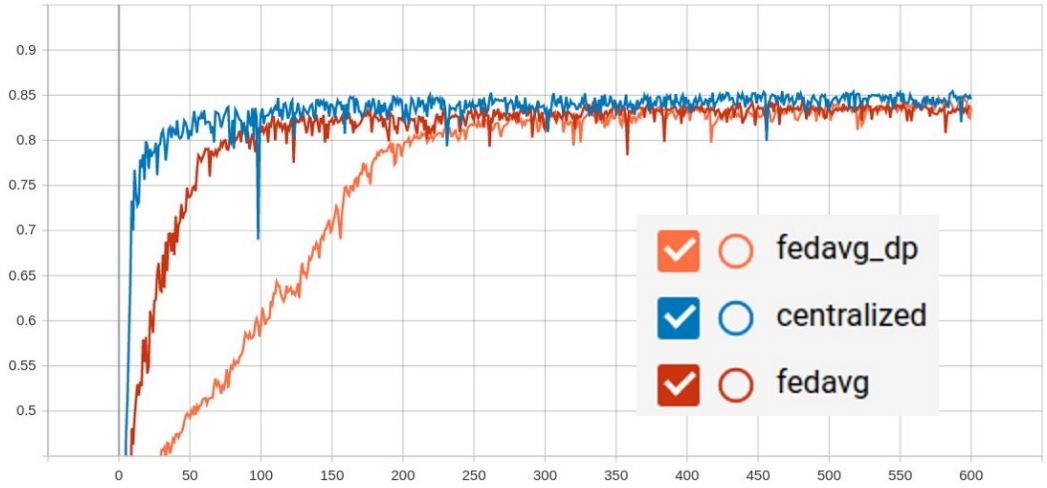
A form of encryption that permits users to perform computations on encrypted data



Blog: https://developer.nvidia.com/blog/federated-learning-with-homomorphic-encryption/
Example: https://github.com/NVIDIA/NVFlare/tree/main/examples/cifar10

Differential Privacy for BraTS18 Segmentation

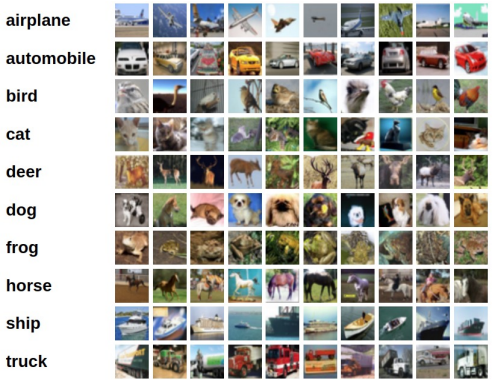
validation Dice scores of the global model for 600 training epochs:

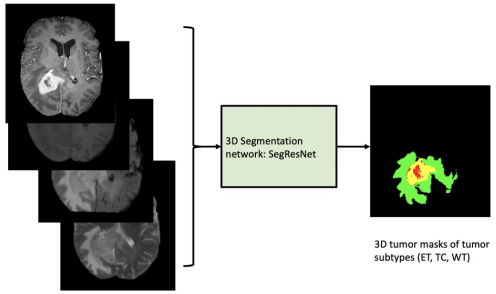


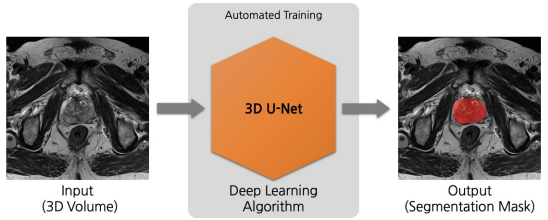
Example: https://github.com/NVIDIA/NVFlare/tree/main/examples/brats18

END-TO-END EXAMPLES (CIFAR10, BRATS18, PROSTATE)

- Comprehensive example for researchers to compare algorithms
  1. Set up a virtual environment
  2. Create your FL workspace
  3. Run automated experiments
     1. Varying data heterogeneity of data splits
     2. Centralized training
     3. FedAvg on different data splits
     4. Advanced FL algorithms (FedProx and FedOpt)
     5. Secure aggregation using homomorphic encryption
     6. Differential privacy
  4. Results



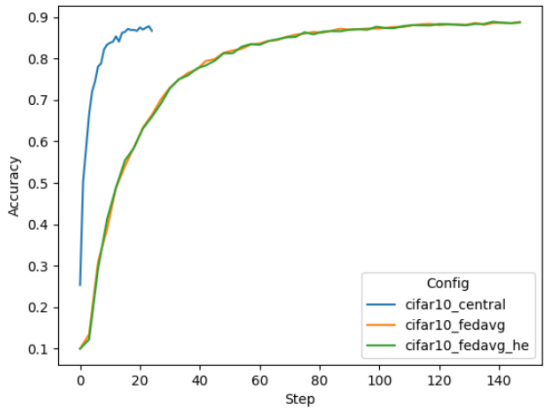




4.1 Central vs. FedAvg

With a data split using `alpha=1.0`, i.e. a non-heterogeneous split, we achieve the follc achieve a similar performance to central training and that HE does not impact the per aggregation step.

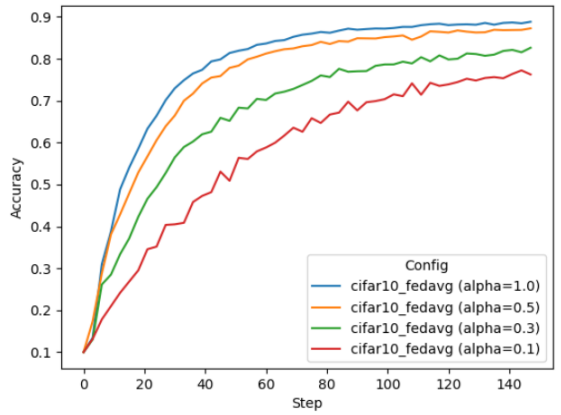
Config	Alpha	Val score
cifar10_central	1.0	0.8668
cifar10_fedavg	1.0	0.8840
cifar10_fedavg_he	1.0	0.8868



4.2 Impact of client data heterogeneity

We also tried different `alpha` values, where lower values cause higher heterogeneity. Thi FedAvg algorithms.

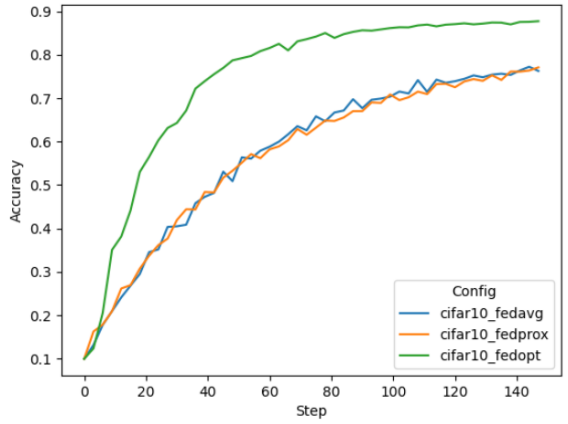
Config	Alpha	Val score
cifar10_fedavg	1.0	0.8840
cifar10_fedavg	0.5	0.8727
cifar10_fedavg	0.3	0.8264
cifar10_fedavg	0.1	0.7626



4.3 FedProx vs. FedOpt

Finally, we are comparing an `alpha` setting of 0.1, causing a high client data heteroc namely FedProx and FedOpt. Both achieve a better performance compared to FedAv convergence rate by utilizing SGD with momentum to update the global model on tl amount of training steps.

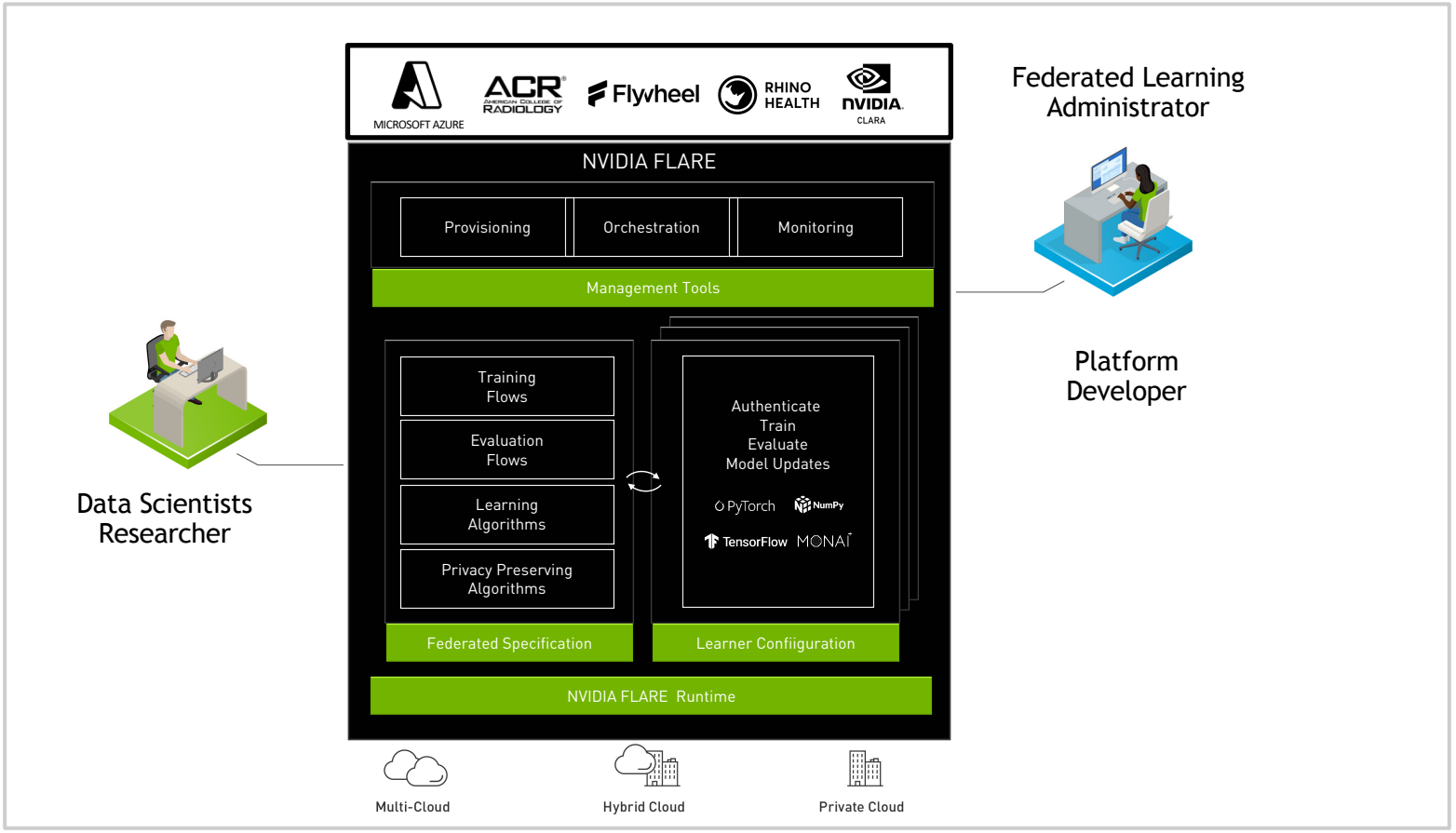
Config	Alpha	Val score
cifar10_fedavg	0.1	0.7626
cifar10_fedprox	0.1	0.7709
cifar10_fedopt	0.1	0.7963

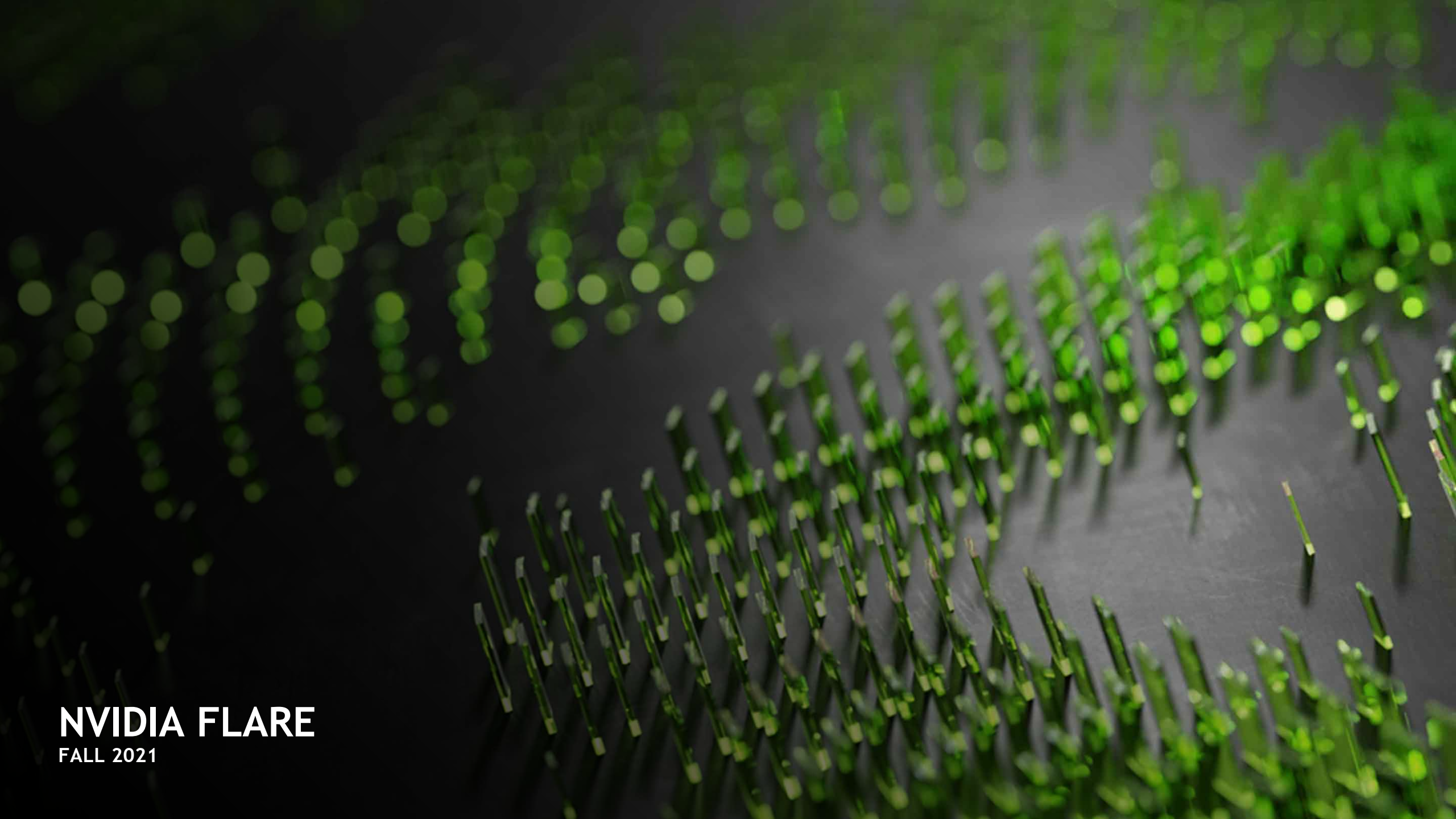


NVIDIA

NVIDIA FLARE 2.0

Accelerating the Federated Learning Paradigm




NVIDIA FLARE
FALL 2021

LEARNING ALGORITHMS

Federated Averaging (FedAvg)

- Weighted average to update global model

McMahan et al.

Federated Proxy (FedProx)

- Clients add a loss to stay close the global model
- Avoids models drifting away from global model in heterogenous datasets

Li et al.

Adaptive Federated Optimization (FedOpt)

- Global model is updated using an optimizer (SGD w. momentum, Adam, Yogi, Adagrad, etc.)

Reddi et al.

Cyclic Weight Transfer

- Models are continuously fine-tuned and circulated around institutions

Chang et al.

SCAFFOLD

- Adds correction terms during training to deal with non-IID

Karimireddy et al.

Ditto

- Fairness through personalization

Li et al.

Algorithms can be extended

- Differential privacy
- Homomorphic Encryption

More to come...

https://developer.nvidia.com/blog/creating-robust-and-generalizable-ai-models-with-nvidia-flare/

FLARE 2.1: BUILT FOR SCALABILITY

High-Availability & Multi-Task Execution

- High availability (HA) supports multiple FL servers and automatically activates a backup server when the currently active server becomes unavailable.
- This is managed by a new entity in the federation, the overseer, that's responsible for monitoring the state of all participants and orchestrating the cutover to a backup server when needed.
- Multi-job execution supports resource-based multi-job execution by allowing for concurrent runs, provided that the resources required by the jobs are satisfied

https://developer.nvidia.com/blog/experimenting-with-novel-distributed-applications-using-nvidia-flare-2-1/

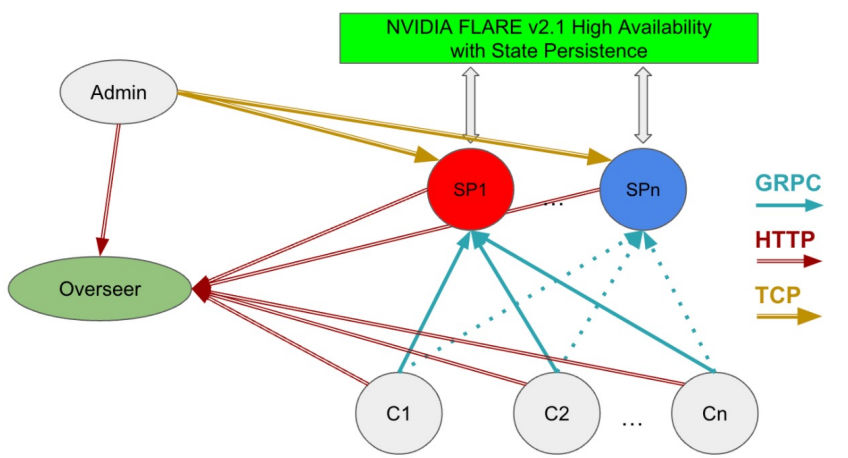


Figure 3. The NVIDIA FLARE deployment for high availability (HA)

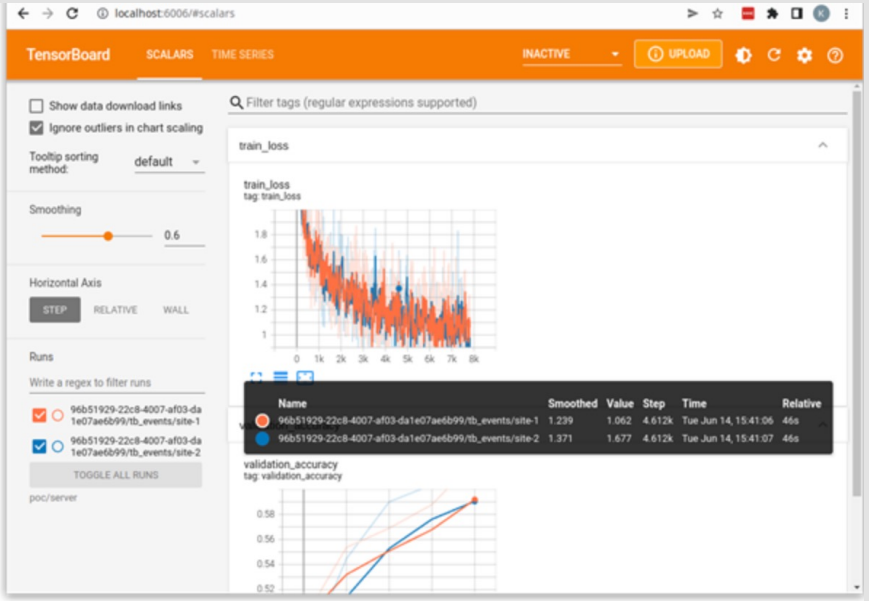


Figure 2. Example TensorBoard output from the hello-pt-tb *application*