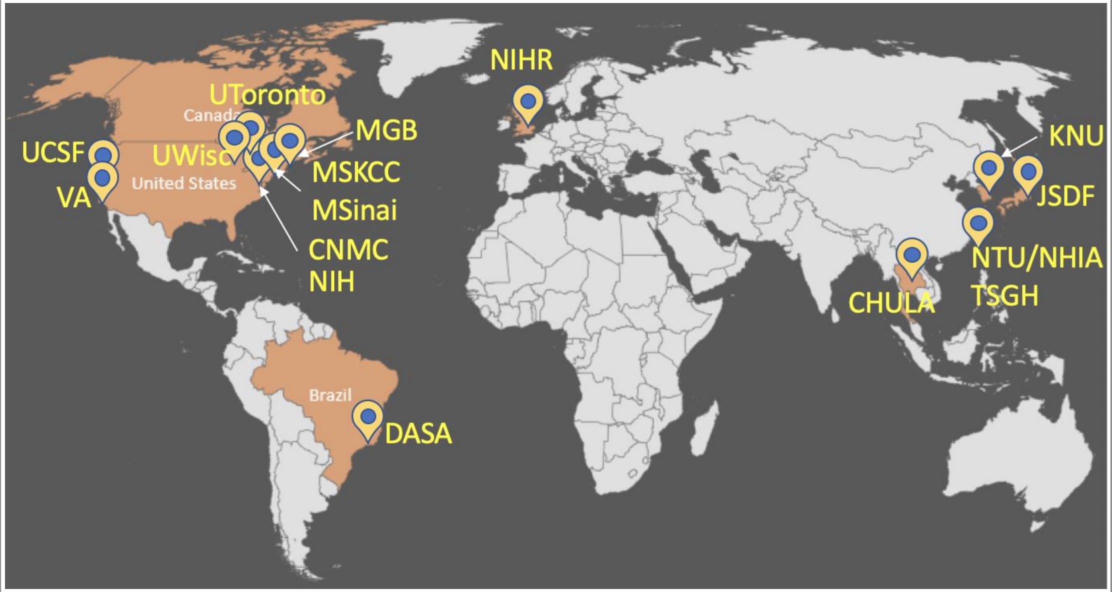
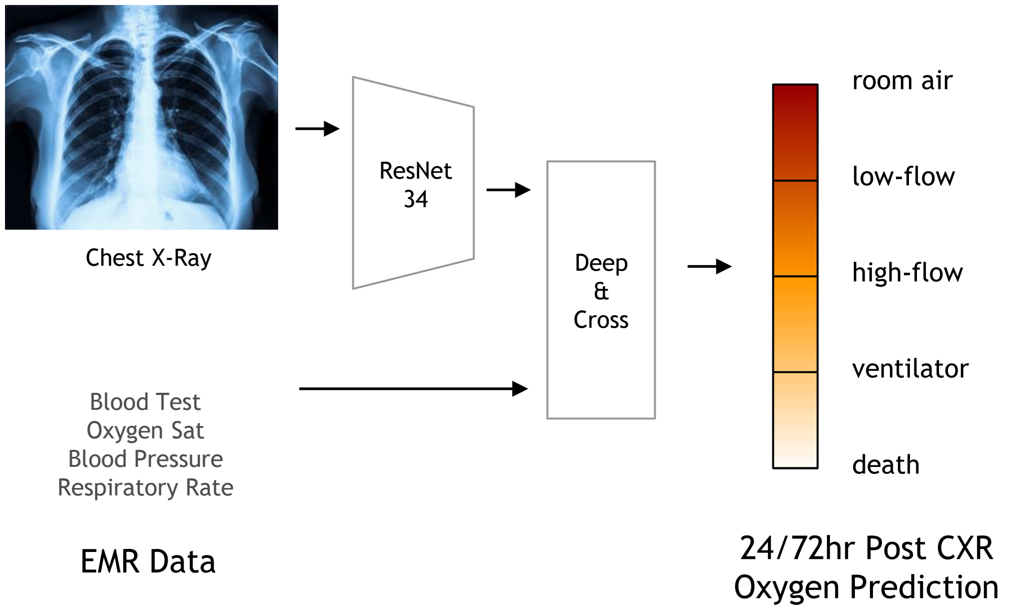
NVIDIA Flare – Federated Learning SDK

Warren Tseng, Solution Architect, NVIDIA Taiwan

CLARA FEDERATED LEARNING FOR COVID-19 PATIENT CARE
"EXAM" AI MODEL



Clara Federated Learning
20 Sites | 8 Countries
COVID-19 Oxygen Prediction



Global Model Achieved .93AUC
>25% Relative Improvement
Every Site Benefited Regardless of Dataset Size

Hospitals Build AI Model that Predicts Oxygen Needs of COVID-19 Patients | NVIDIA Blog

NVIDIA FEDERATED LEARNING

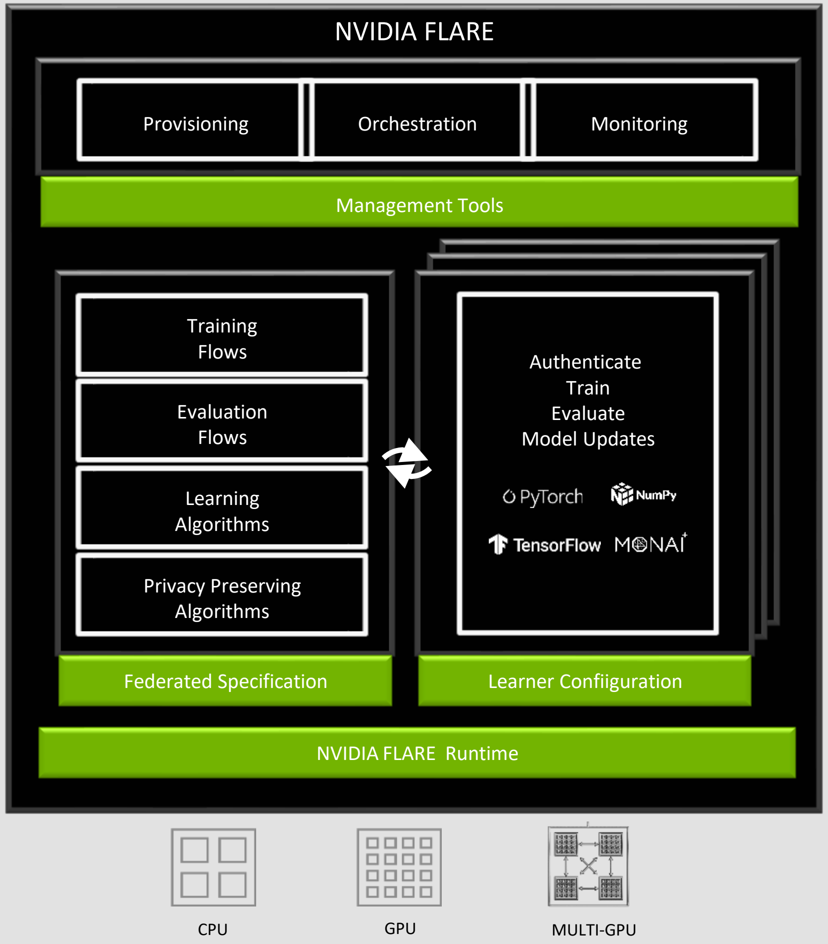
Applications across industries

NVIDIA FLARE

NVIDIA **F**ederated **L**earning **A**pplication **R**untime **E**nvironment
- An **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



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

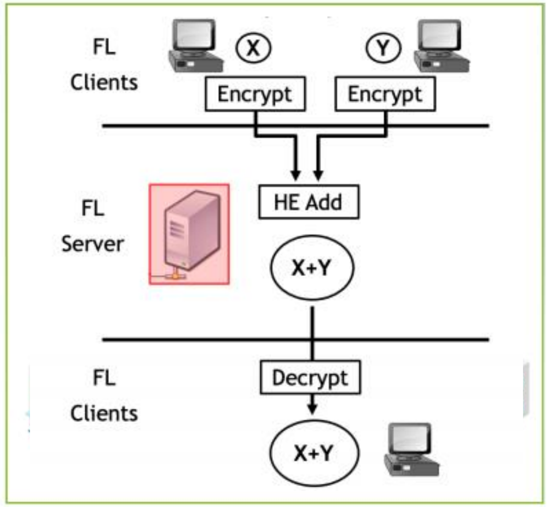
Homomorphic Encryption & Differential Privacy

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



Differential Privacy for BraTS18 Segmentation

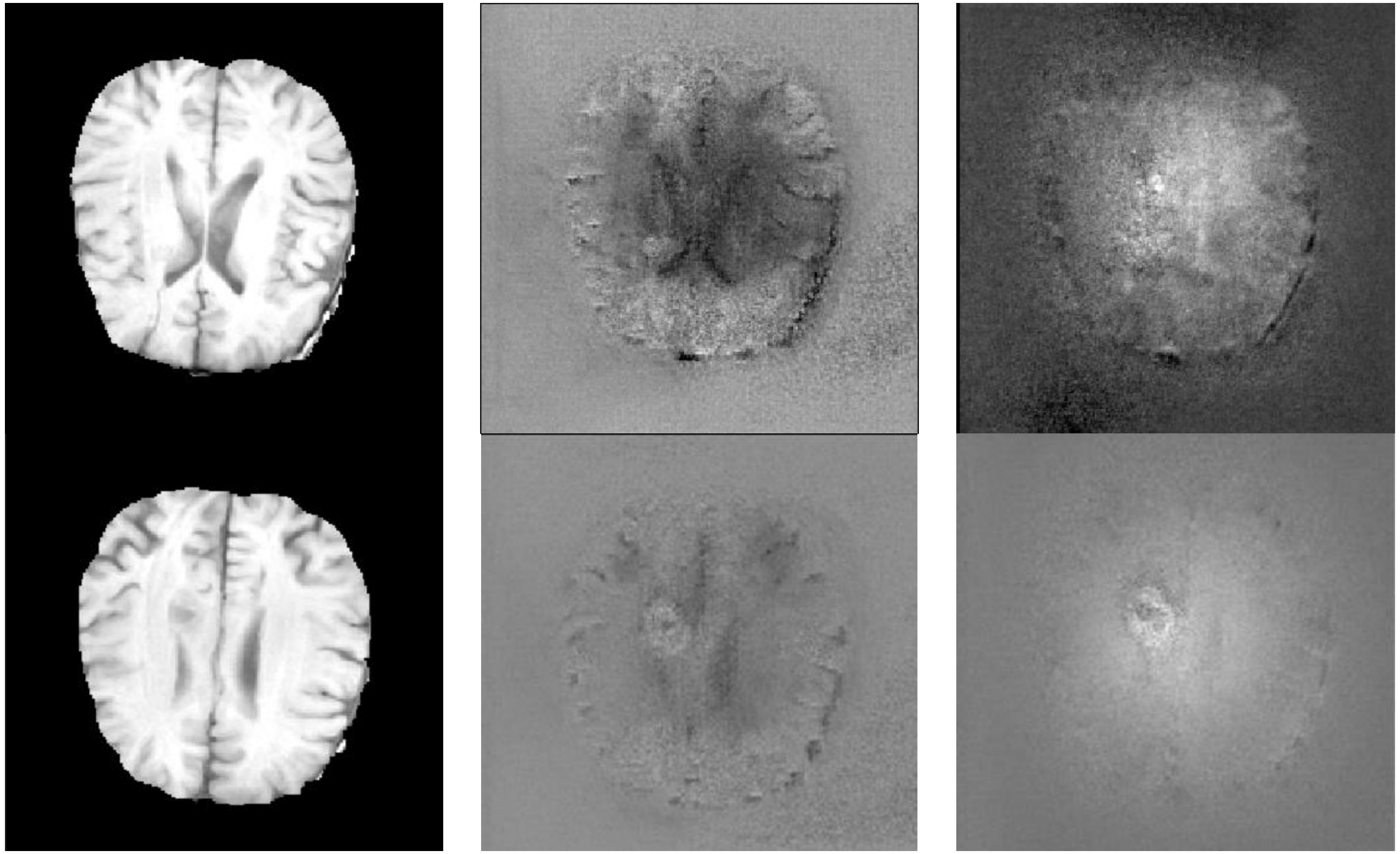
validation Dice scores of the global model for 600 training epochs:

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

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

MODEL INVERSION CASE STUDY

Training volumes	Reconstructions from FL model after training	Reconstructions from FL model trained with our privacy-preserving module



Understanding Deep Image Representations by Inverting Them
https://arxiv.org/abs/1412.0035

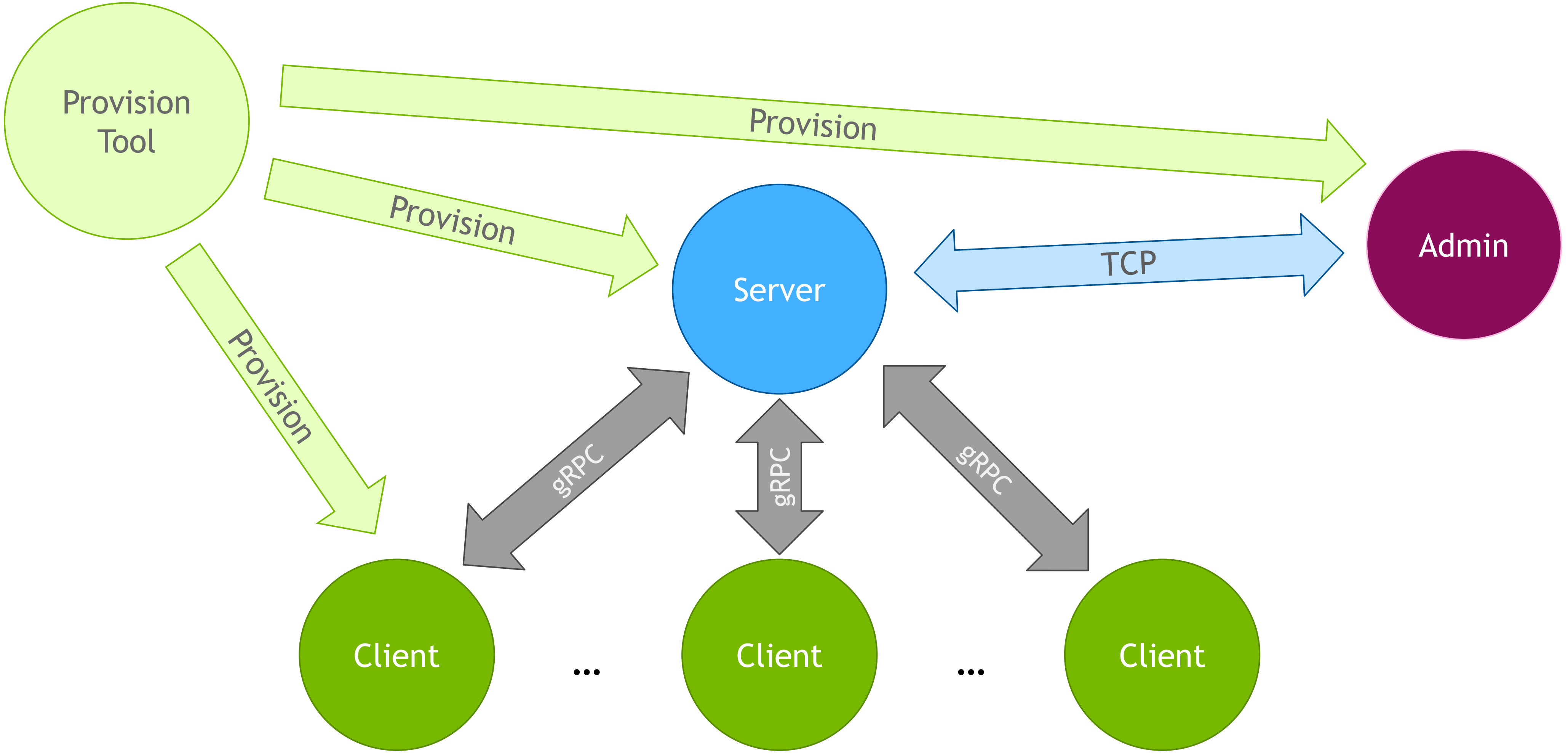
Li et al. 2019 https://arxiv.org/abs/1910.00962

NVFLARE 2.3 NEW FEATURES

- Cloud Deployment Support – Azure & AWS
- Job Signing - The submitter's private key is used to sign each file's digest to ensure that custom code is signed.
- Client-Side Model Initialization – Prevent running custom model initialization code on server. It could be a security risk.
- New Examples for Traditional ML – Linear/logistic regression, SVM, K-Means and Random Forest
- Vertical Learning Support

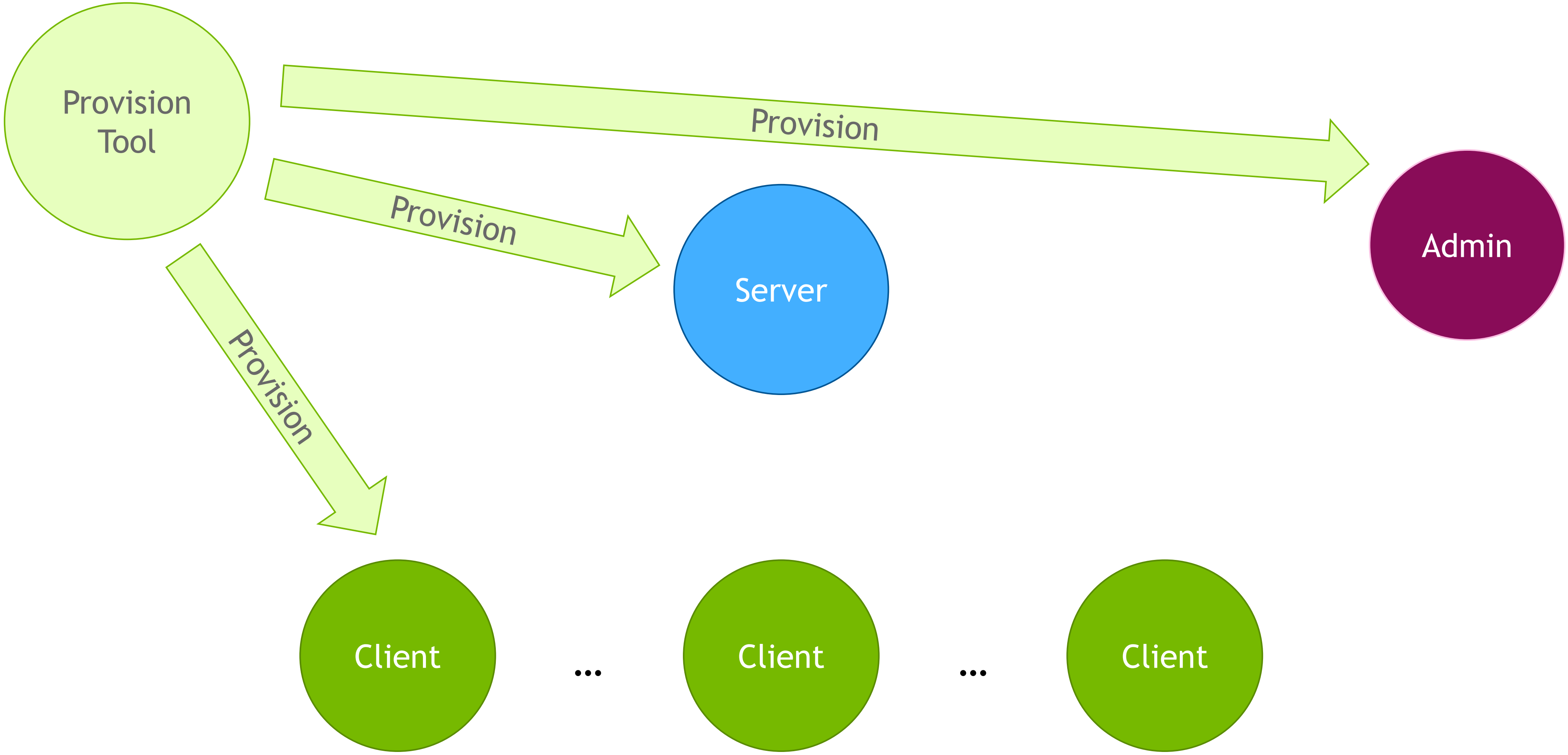
NVIDIA FLARE

High-level Architecture



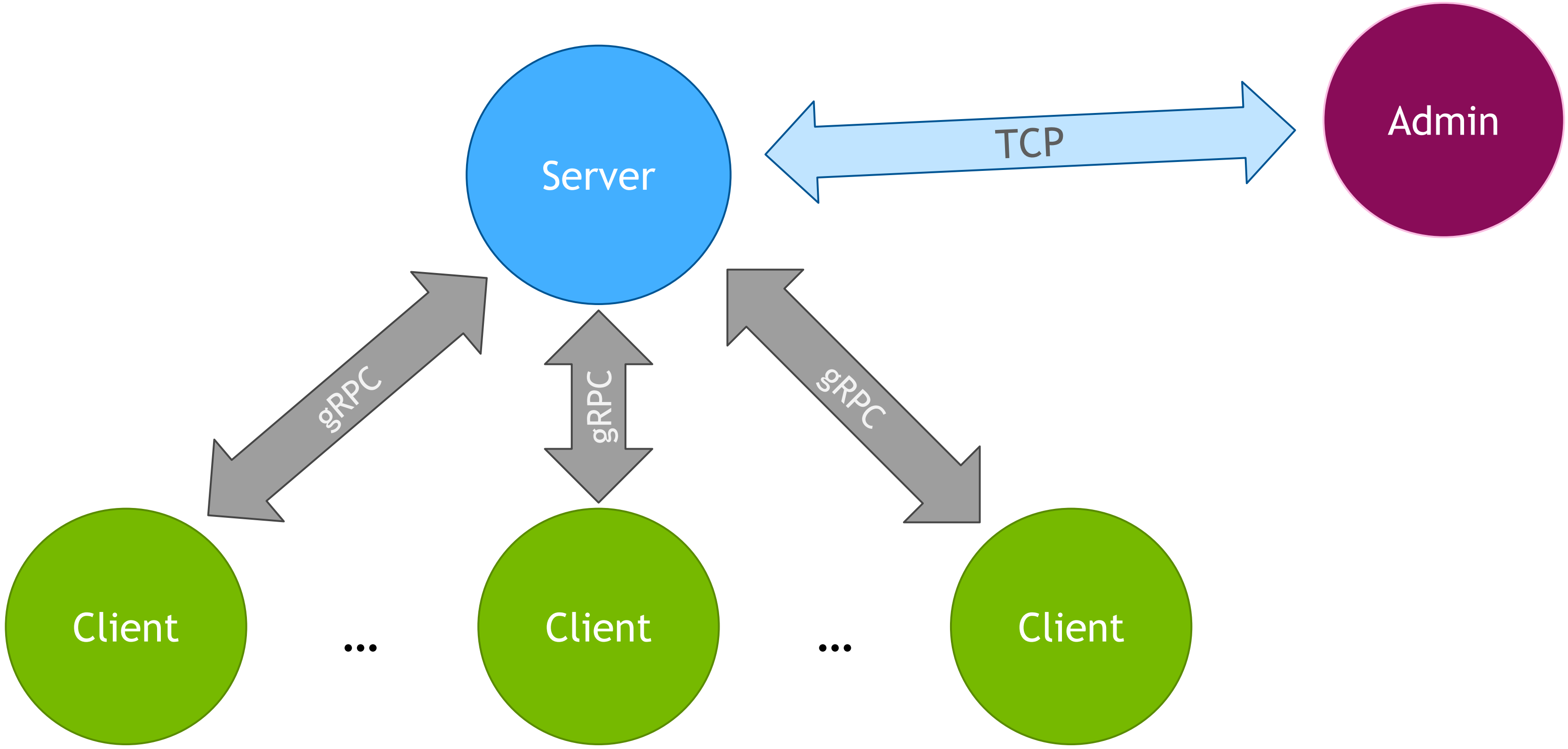
NVIDIA FLARE

High-level Architecture



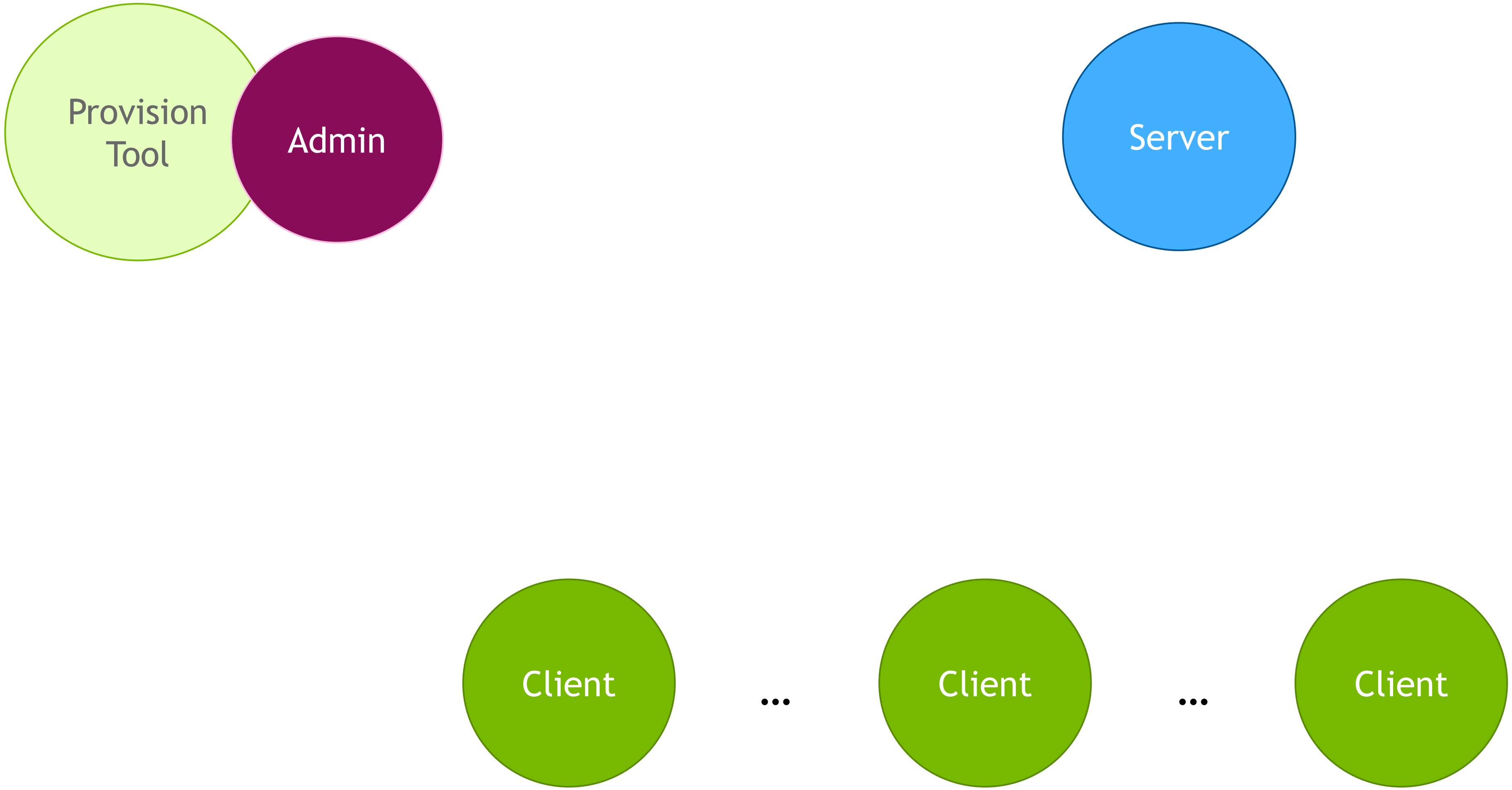
NVIDIA FLARE

High-level Architecture



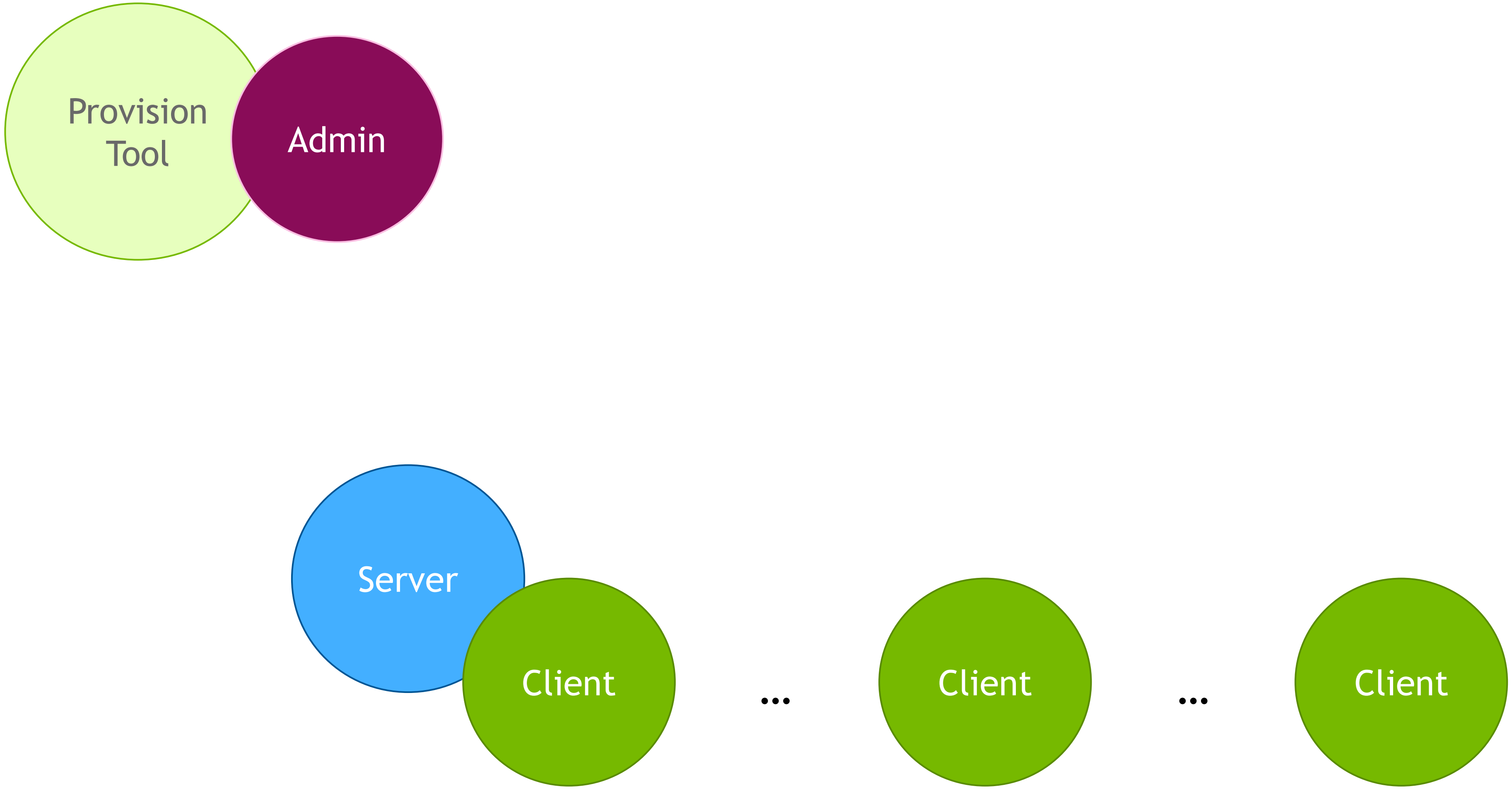
NVIDIA FLARE

Flexibility



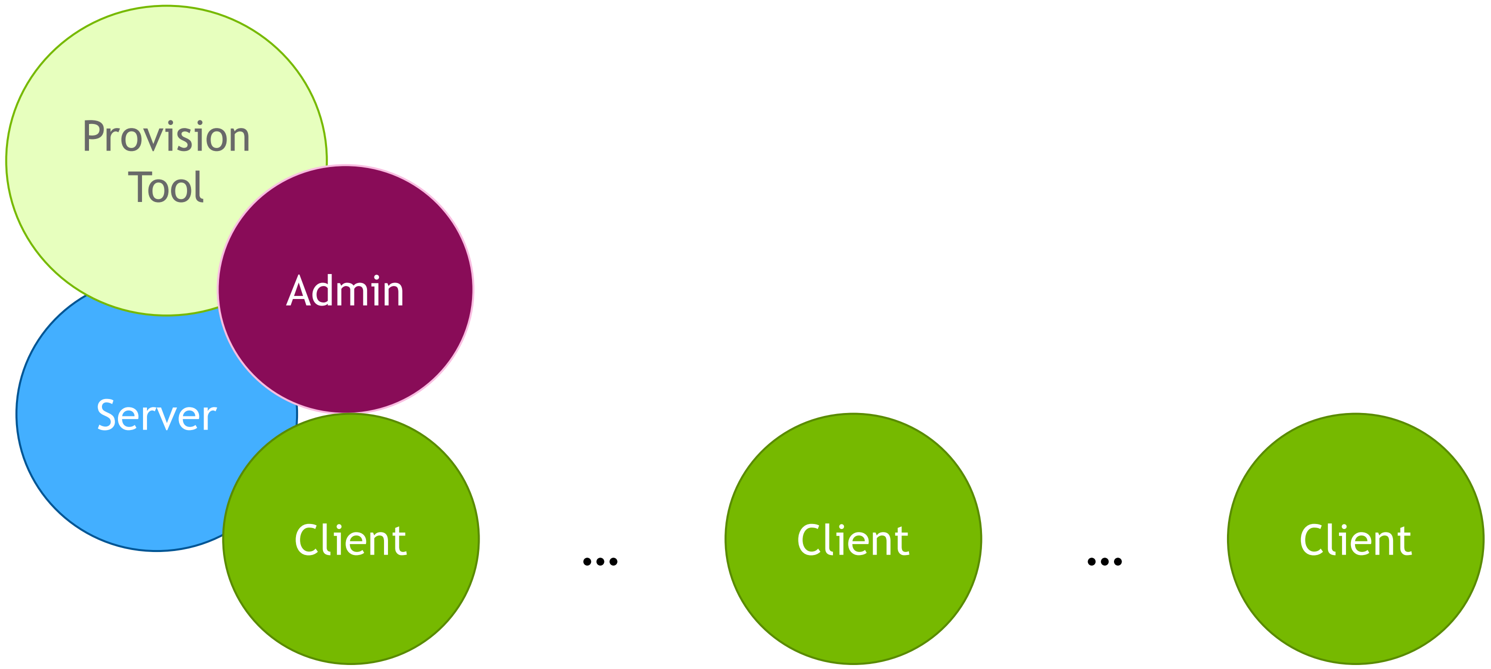
NVIDIA FLARE

Flexibility



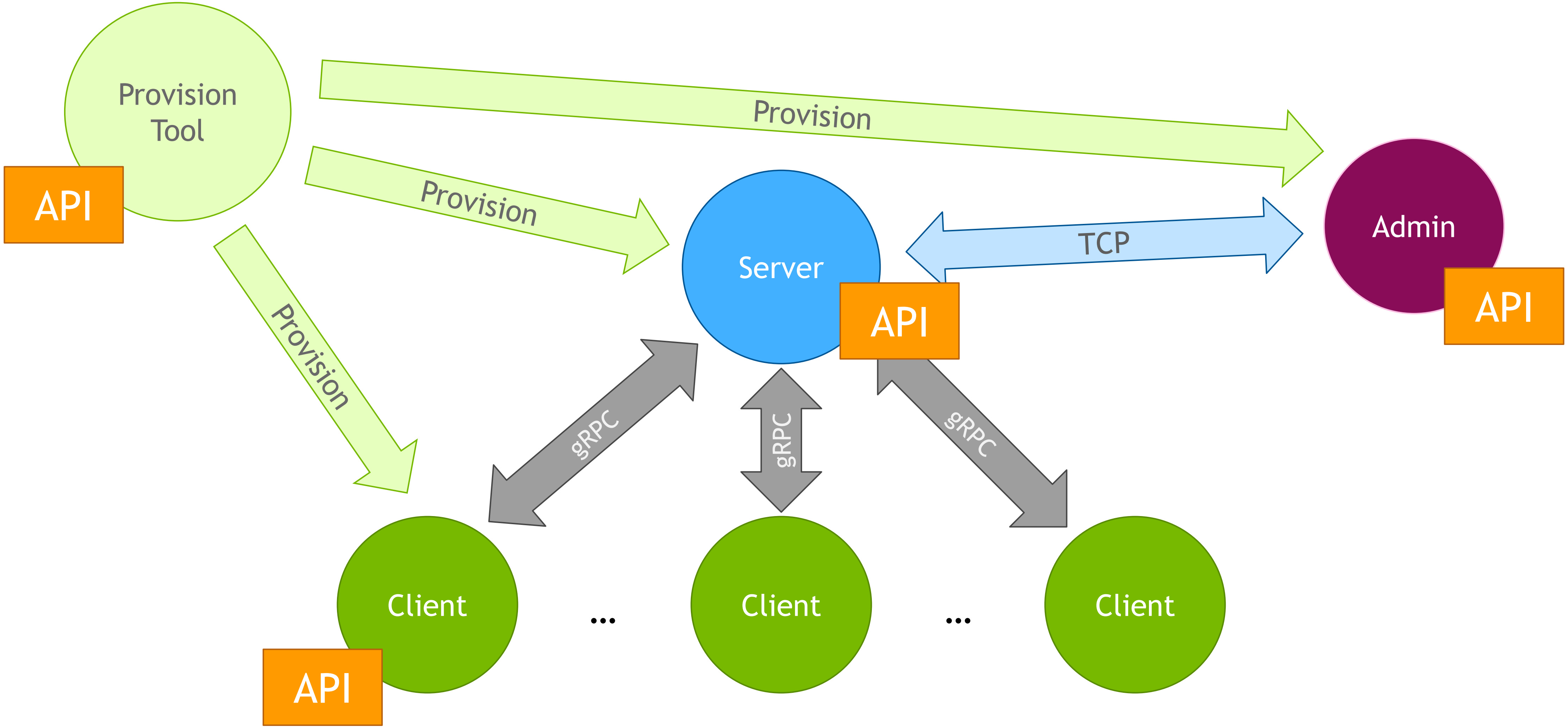
NVIDIA FLARE

Flexibility



NVIDIA FLARE

High-level Architecture



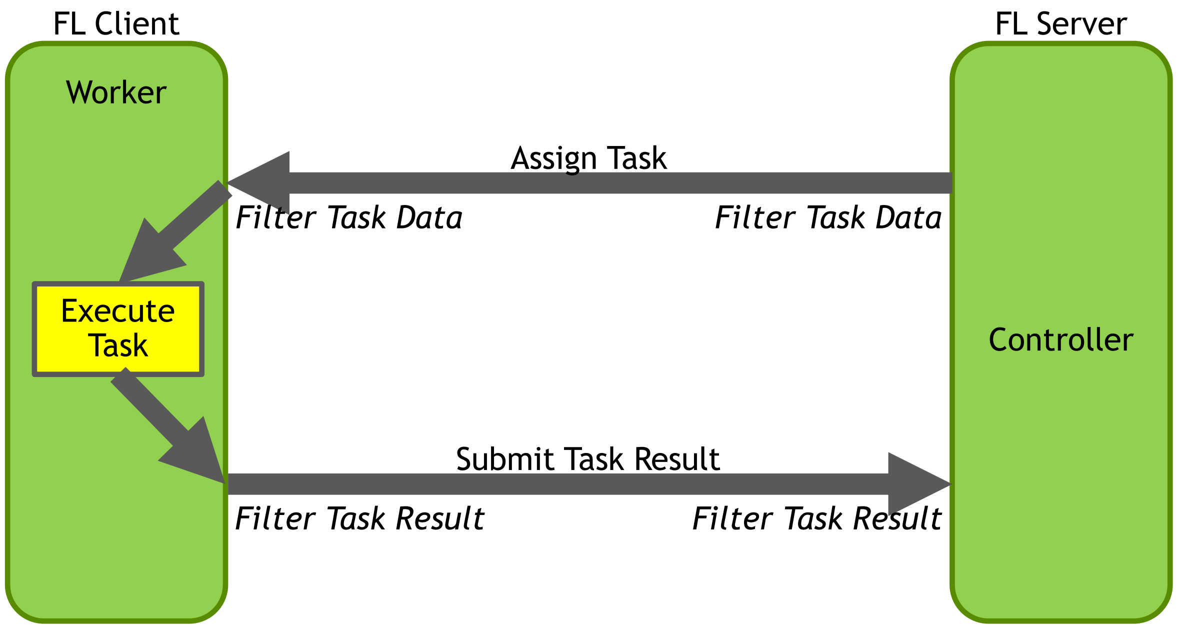
Controller and Worker API

Federated Learning Workflow

The Controller and Worker APIs define the overall control flow via Events, Tasks, and Executors.

- The Controller defines the series of Tasks to be executed by Workers and determines how these Tasks are distributed (broadcast, cyclic, send).
- The Worker implements Executors that execute specific named Tasks as defined and distributed by the Controller.
- The Controller aggregates the Workers' Task Result as defined in the Controller workflow.

Filters can be used in both the Controller and Executor and applied to both Task Data and Task Results.



NVFlare Adoptions

- NCKU – Pathology
- Sinica – FL Algorithm Developing
- SNAC (Australia) – Brain MRI
- NTUH, CGMH, …

RESOURCES

- Documentation: https://nvflare.readthedocs.io/en/main/index.html
- Getting Started: https://nvflare.readthedocs.io/en/main/getting_started.html
- Examples: https://github.com/NVIDIA/NVFlare/tree/dev/examples
  - Scikit-learn SVM: https://github.com/NVIDIA/NVFlare/tree/dev/examples/advanced/sklearn-svm
  - Federated Statistics: https://github.com/NVIDIA/NVFlare/tree/dev/examples/advanced/federated-statistics
- Demo:
  - https://www.youtube.com/watch?v=RnnMTjPm_PE&list=PL5uCDOVJqgeuaB0i1MbVS0k2mW83Pmxf5&index=22
  - https://www.youtube.com/watch?v=odB58L_HfnE&list=PL5uCDOVJqgeuaB0i1MbVS0k2mW83Pmxf5&index=25
  - https://www.youtube.com/watch?v=ahHH12dz9FM&list=PL5uCDOVJqgeuaB0i1MbVS0k2mW83Pmxf5&index=29
  - https://www.youtube.com/watch?v=P0_amvxqnuo&list=PL5uCDOVJqgeuaB0i1MbVS0k2mW83Pmxf5&index=30

NVIDIA

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.

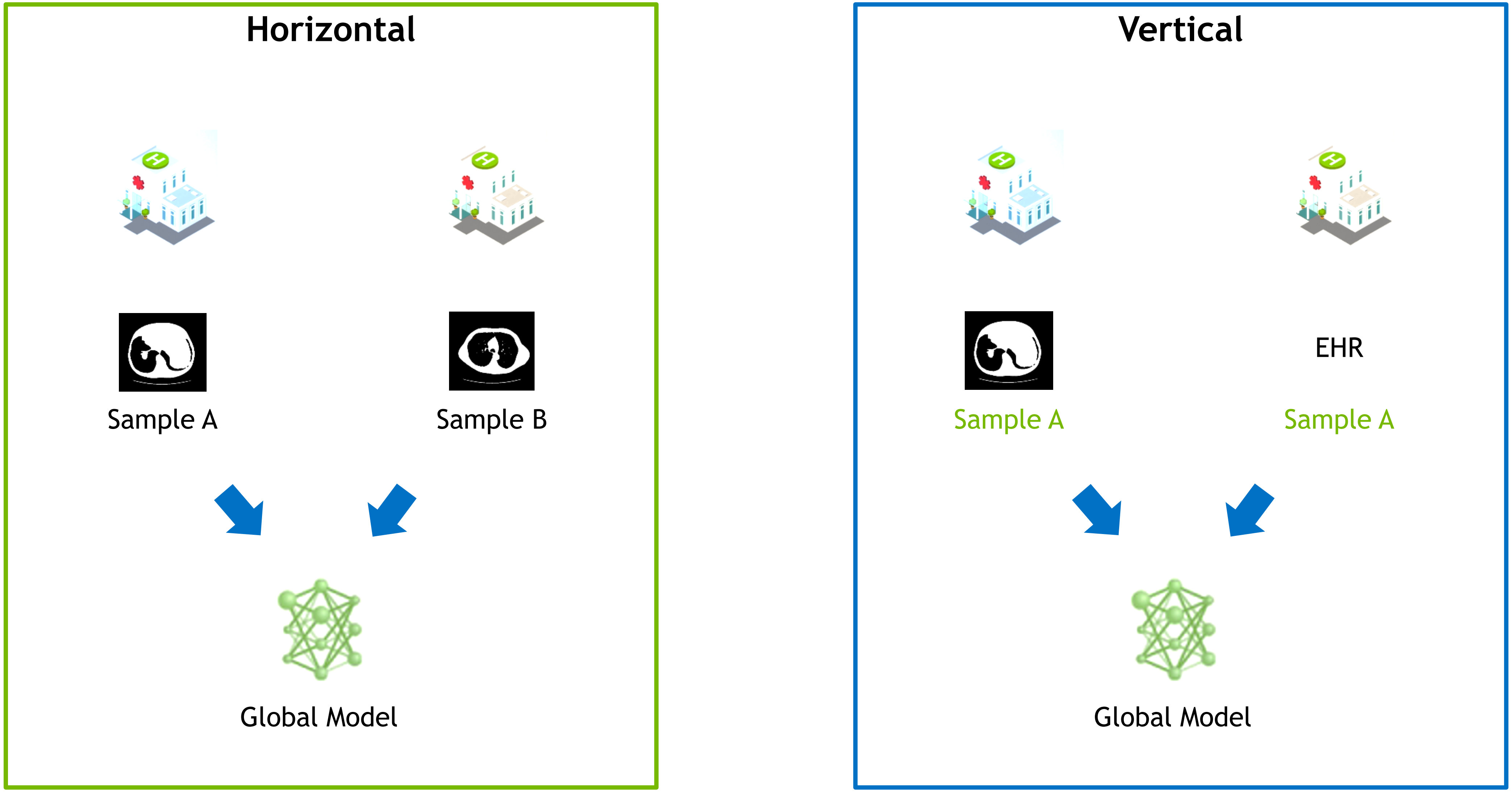
Ready to use FL specification and management tools enabling seamless execution.

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.

HORIZONTAL & VERTICAL LEARNING



NVFLARE 2.2 NEW FEATURES

From Research Simulation to Real World Deployment

FL SIMULATOR

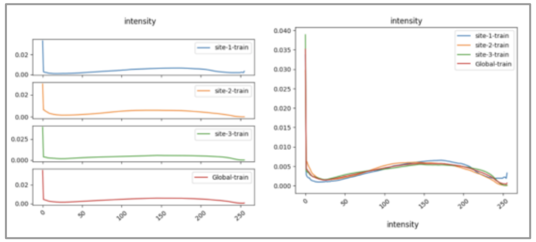
Rapid Development and Debugging

```python
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

BUILDING AI FOR REAL-WORLD CLINICAL PERFORMANCE

Taking Algorithms Beyond Proof-of-Concept

REAL-WORLD AI DESIGN

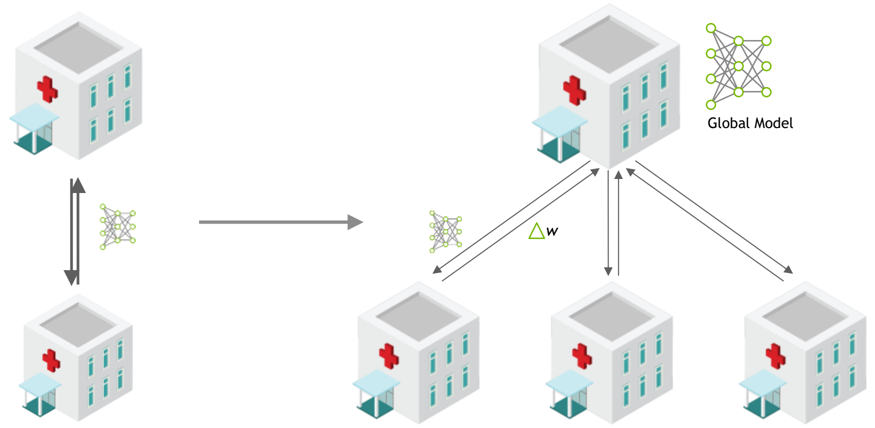
External Validation, Multiple Institutions, Prospective Data

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”

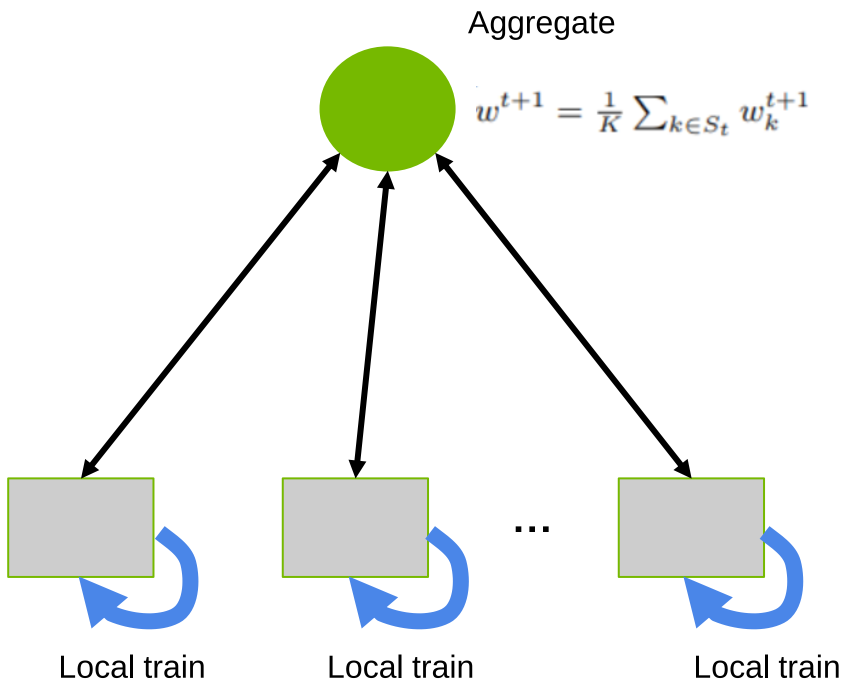
Federated Learning
“Generalize”

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.

FedAvg

Communication-Efficient Learning of Deep Networks from Decentralized Data

https://arxiv.org/pdf/1602.05629.pdf



Algorithm 1 Federated Averaging (`FedAvg`)

Input: $K, T, \eta, E, w^0, N, p_k, k = 1, \cdots, N$

for $t = 0, \cdots, T-1$ **do**

  Server selects a subset S_t of K devices at random (each device k is chosen with probability p_k)

  Server sends w^t to all chosen devices

  Each device $k \in S_t$ updates w^t for E epochs of SGD on F_k with step-size η to obtain w_k^{t+1}

  Each device $k \in S_t$ sends w_k^{t+1} back to the server

  Server aggregates the w's as $w^{t+1} = \frac{1}{K}\sum_{k \in S_t} w_k^{t+1}$

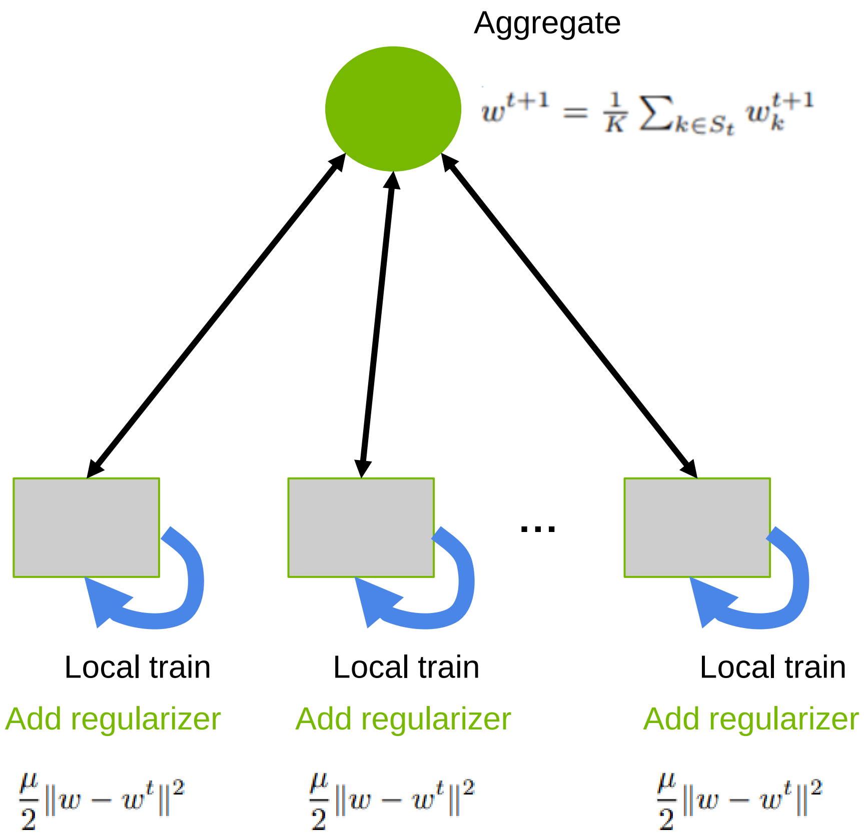
end for

→ FullModelShareableGenerator + InTimeAccumulateWeightedAggregator

FedProx

FEDERATED OPTIMIZATION IN HETEROGENEOUS NETWORKS

https://arxiv.org/pdf/1812.06127.pdf



Algorithm 2 `FedProx` (Proposed Framework)

Input: $K, T, \mu, \gamma, w^0, N, p_k, k = 1, \cdots, N$

for $t = 0, \cdots, T-1$ **do**

Server selects a subset S_t of K devices at random (each device k is chosen with probability p_k)

Server sends w^t to all chosen devices

Each chosen device $k \in S_t$ finds a w_k^{t+1} which is a γ_k^t-inexact minimizer of: $w_k^{t+1} \approx \arg\min_w h_k(w; w^t) = F_k(w) + \frac{\mu}{2}\|w - w^t\|^2$

Each device $k \in S_t$ sends w_k^{t+1} back to the server

Server aggregates the w's as $w^{t+1} = \frac{1}{K}\sum_{k \in S_t} w_k^{t+1}$

end for

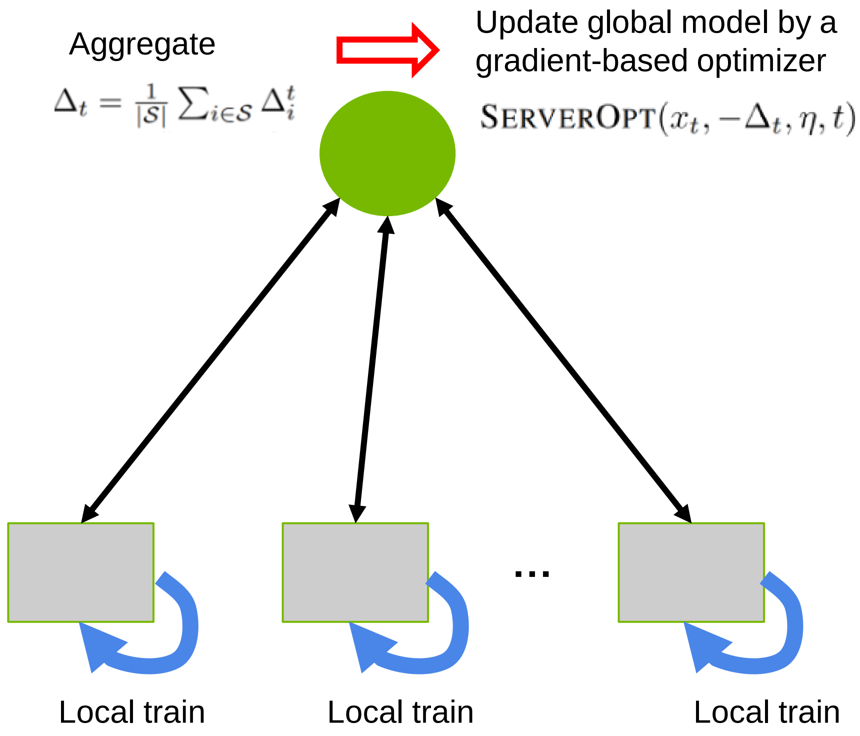
→ Learner

NVIDIA

FedOPT

Adaptive Federated Optimization

https://arxiv.org/pdf/2003.00295.pdf



Algorithm 1 FedOpt

Input: x_0, ClientOpt, ServerOpt

for $t = 0, \cdots, T-1$ **do**

    Sample a subset $\mathcal{S}$ of clients

    $x_{i,0}^t = x_t$

    for each client $i \in \mathcal{S}$ **in parallel do**

        for $k = 0, \cdots, K-1$ **do**

            Compute an unbiased estimate $g_{i,k}^t$ of $\nabla F_i(x_{i,k}^t)$

            $x_{i,k+1}^t = \textsc{ClientOpt}(x_{i,k}^t, g_{i,k}^t, \eta_l, t)$

        $\Delta_i^t = x_{i,K}^t - x_t$

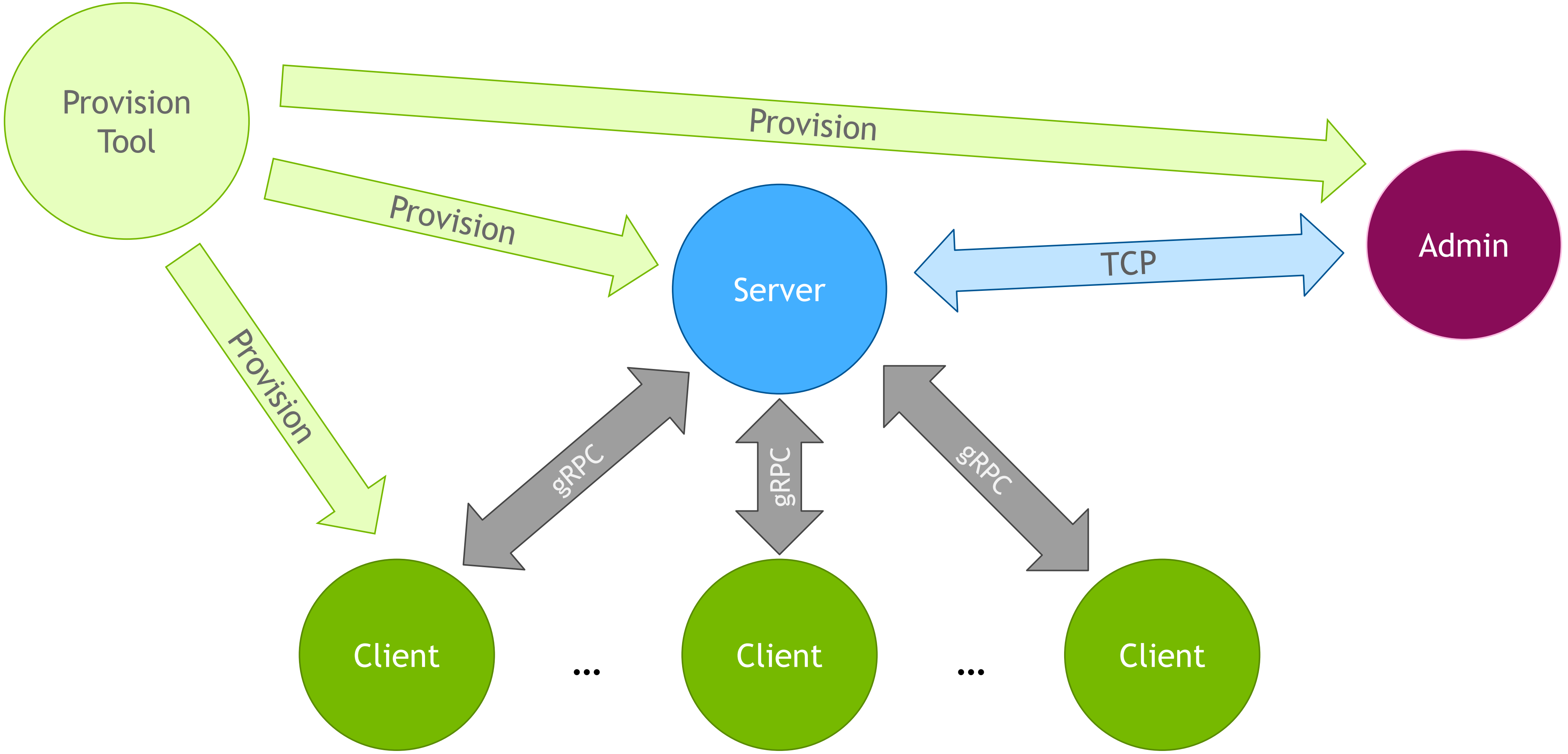
    $\Delta_t = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\Delta_i^t$

    $x_{t+1} = \textsc{ServerOpt}(x_t, -\Delta_t, \eta, t)$

$\longrightarrow$ PTFedOptModelShareableGenerator + InTimeAccumulateWeightedAggregator

NVIDIA FLARE v2.0

High-level Architecture



“ *Using the MONAI imaging framework integrated into Flywheel, with training done on NVIDIA DGX BasePOD, we can apply our state-of-the-art research tools to every single abdominal CT we've ever performed at UW Madison since 2004. 10,000 cases alone used to take **six to eight months** just to get through, and we can now easily process them in **a day**.* “

— John Garrett, Assistant Professor and Director of Informatics, Department of Radiology, University of Wisconsin-Madison

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

AI ACCELERATES THE ENTIRE RADIOLOGICAL WORKFLOW

Challenge

Researchers at the University of Wisconsin-Madison wanted to determine if AI could speed up tedious tasks in the radiologic interpretation process.

They also wanted to use AI to improve patient outcomes via opportunistic screening, but limited data and disparate data sources were hindrances.

Needed infrastructure to handle large, complex data but also tools to make AI training easy, portable, and reproducible.

Solution

They leveraged MONAI from the NVIDIA Clara application framework integrated in Flywheel data management platform to pre-process data from multiple systems and hospitals.

Using NVIDIA Federated Learning Application Runtime Environment, or FLARE, in collaboration with other hospitals, to securely train AI models on DGX BasePOD for medical imaging, annotation and classification.

Containerized software from NVIDIA AI Enterprise enabled the university to easily replicate their workflows to other clinics and institutions.

NVIDIA DGX BasePOD for Healthcare and Life Sciences
DGX A100 for training

NVIDIA Base Command
DGX system software

NVIDIA AI Enterprise
AI Software Suite

NVIDIA Clara Train SDK
MONAI for pre-processing,
FLARE for Federated Learning

1M+ Images processed in less than a day

10K Cases processed in a day vs 6 to 8 months previously

NVIDIA