第八章：新型分布式机器学习系统的工业应用

2021年9月

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

Heterogeneity

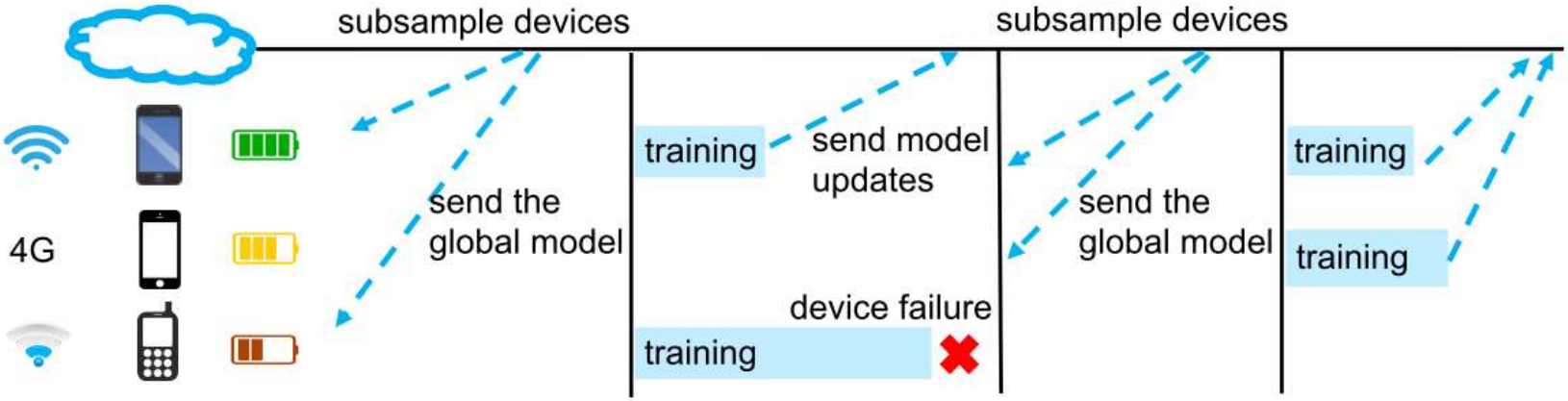
- Systems heterogeneity
- Statistical heterogeneity

Systems heterogeneity

- Asynchronous communication
- Active sampling
- Fault tolerance

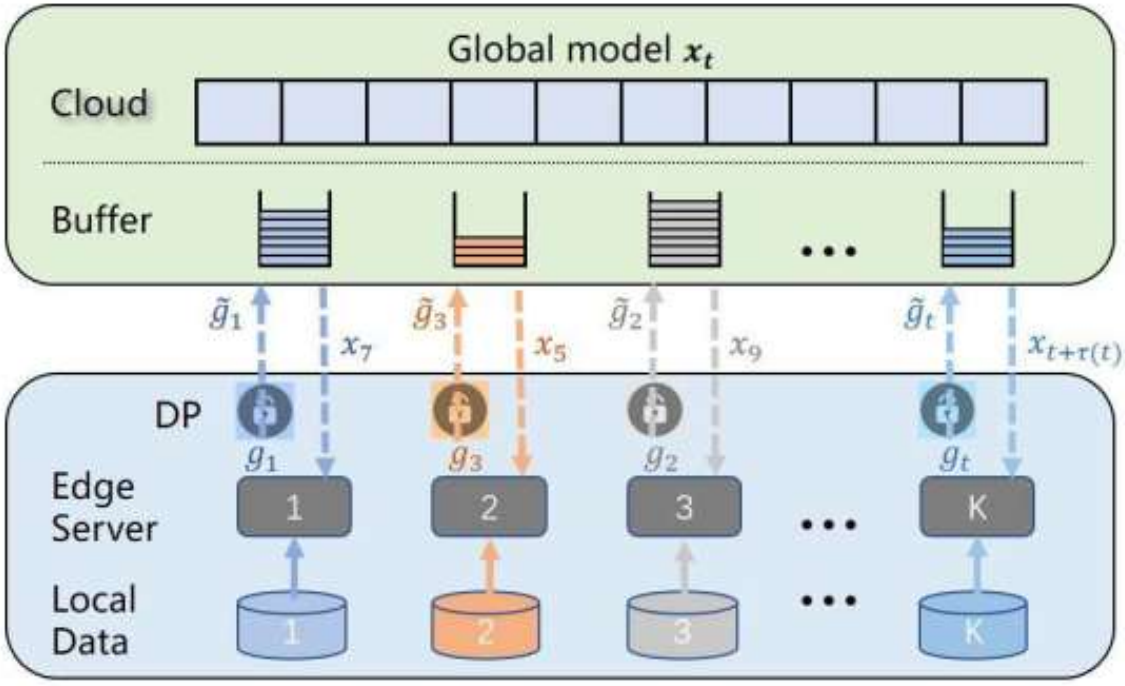
Systems heterogeneity

- Asynchronous communication
  - Synchronous schemes are simple and guarantee a serial-equivalent computational model, but they are also more susceptible to stragglers in the face of device variability.



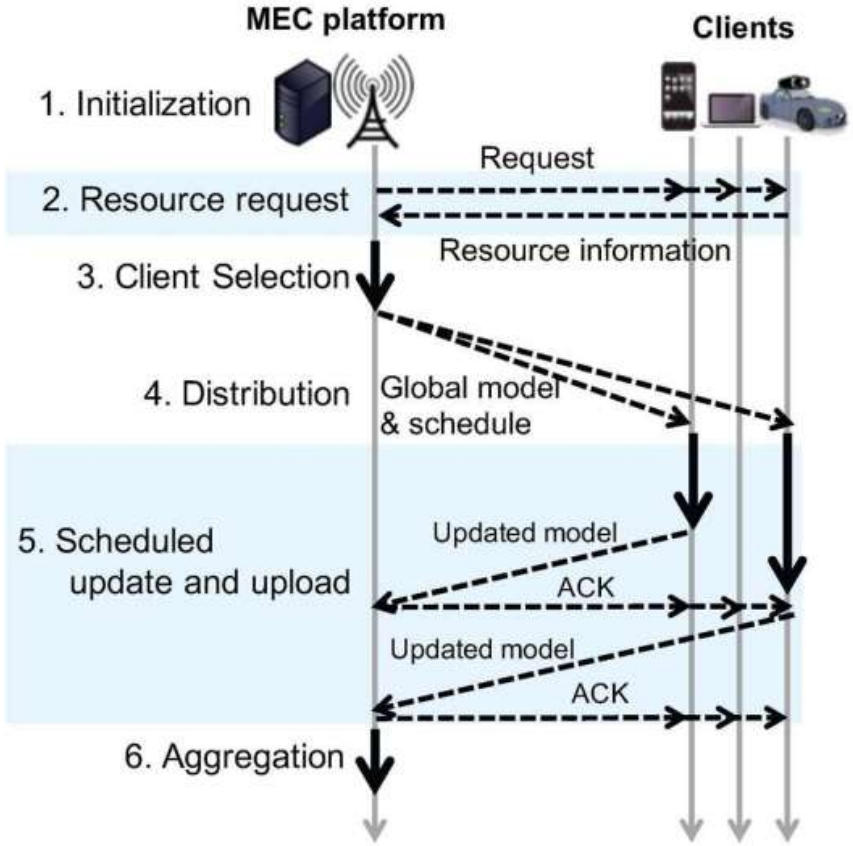
Systems heterogeneity

- Asynchronous communication
  - Asynchronous schemes are an attractive approach to mitigate stragglers in heterogeneous environments.



Systems heterogeneity

- Active sampling
  - Actively selecting participating devices at each round.

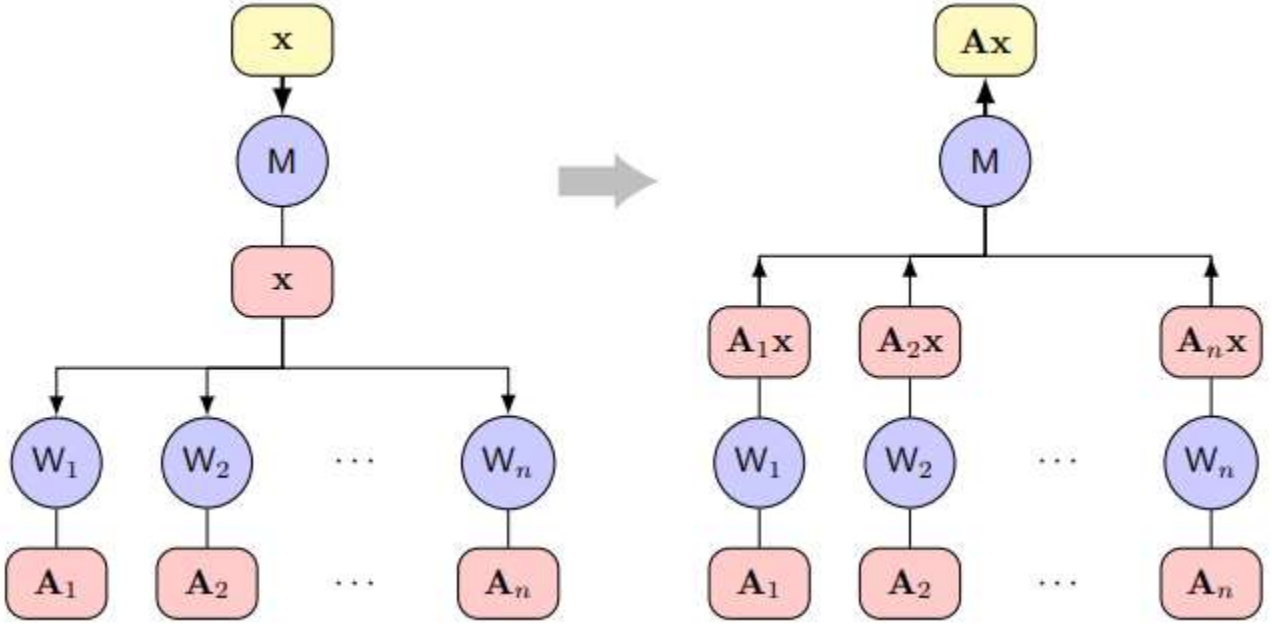


Systems heterogeneity

- Fault tolerance
  - Fault tolerance has been extensively studied in the systems community and is a fundamental consideration of classical distributed systems.
  - When learning over remote devices, however, fault tolerance becomes more critical.
  - One practical strategy is to simply ignore such device failure, which may introduce bias into the device sampling scheme if the failed devices have specific data characteristics.

Systems heterogeneity

- Fault tolerance
  - Coded computation is another option to tolerate device failures by introducing algorithmic redundancy.

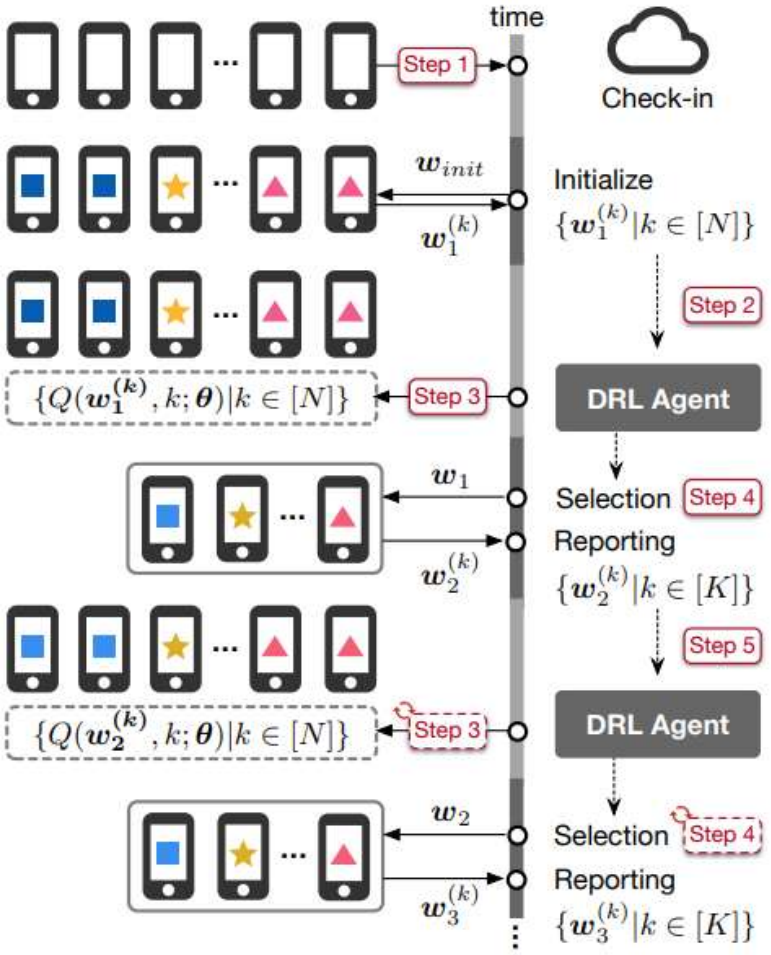


Statistical heterogeneity

- Overcome the non-IID issue
- Utilize the non-IID feature

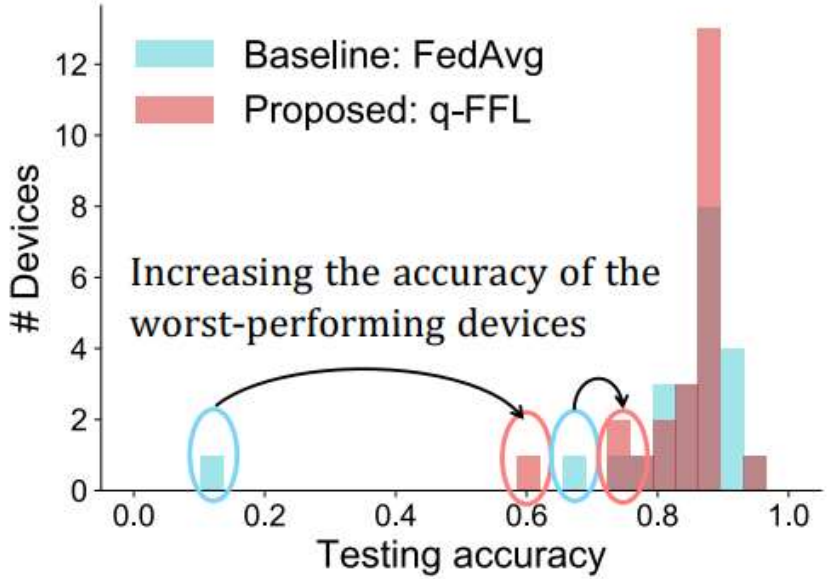
Statistical heterogeneity

- Overcome the non-IID issue
  - Although the data is not independent and identically distributed among all the clients, we can relieve this issue by client selection.
  - Client selection can be formulated as a deep reinforcement learning problem in federated learning.
  - It solely relies on model weight information to determine which device may improve the global model the most —thus preserving the same level of privacy as the original FL does.



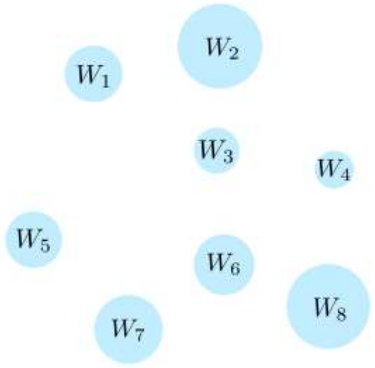
Statistical heterogeneity

- Overcome the non-IID issue
  - Devices with higher loss are given higher relative weight to encourage less variance in the final accuracy distribution.

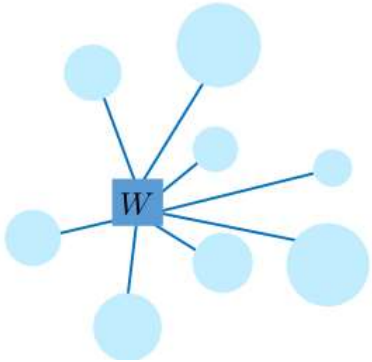


Statistical heterogeneity

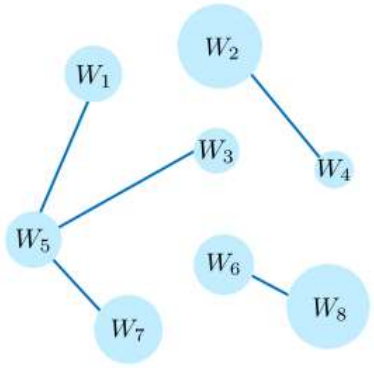
- Utilize the non-IID feature
  - Non-IID data is not just an issue for federated learning, but also a natural feature in this setting.
  - Personalized federated learning is welcomed.



Learn personalized models for each device; do not learn from peers.



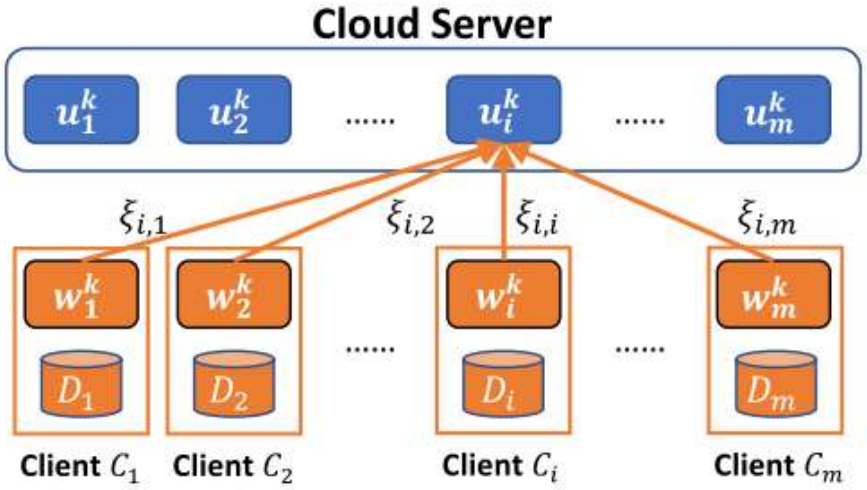
Learn a global model; learn from peers.



Learn personalized models for each device; learn from peers.

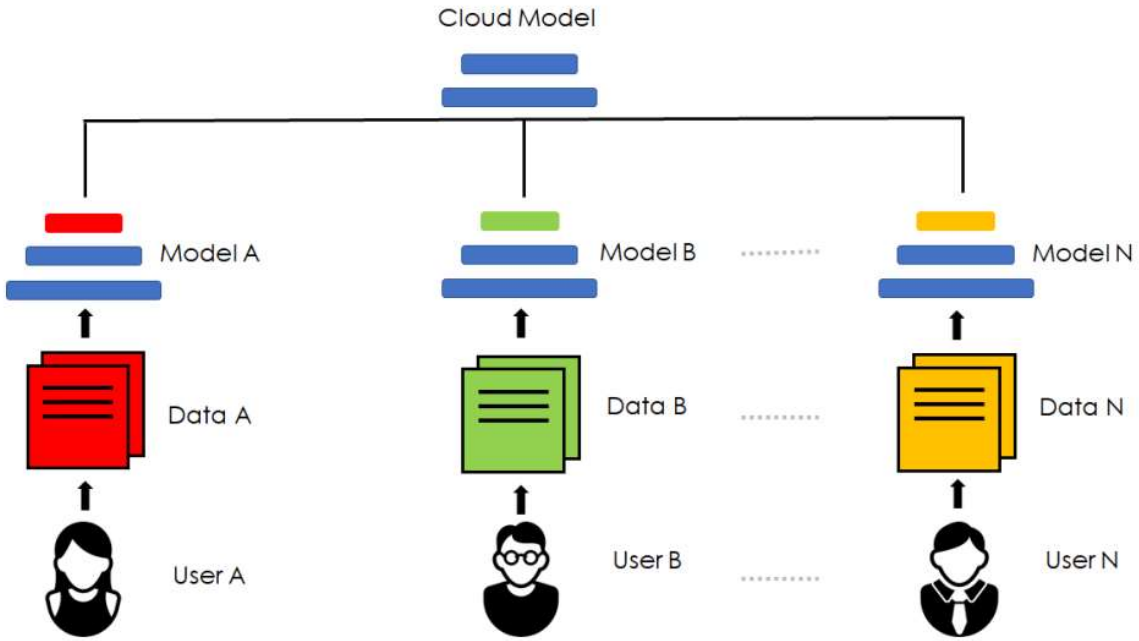
Statistical heterogeneity

- Utilize the non-IID feature
  - FedAMP allows each client to own a local personalized model, it maintains a personalized cloud model on the cloud server for each client.
  - FedAMP realizes the attentive message passing mechanism by attentively passing the personalized model of each client as a message to the personalized cloud models with similar model parameters.
  - FedAMP updates the personalized cloud model of each client by a weighted convex combination of all the messages it receives.



Statistical heterogeneity

- Utilize the non-IID feature
  - The base layers are shared with the parameter server while the personalization layers are kept private by each device.

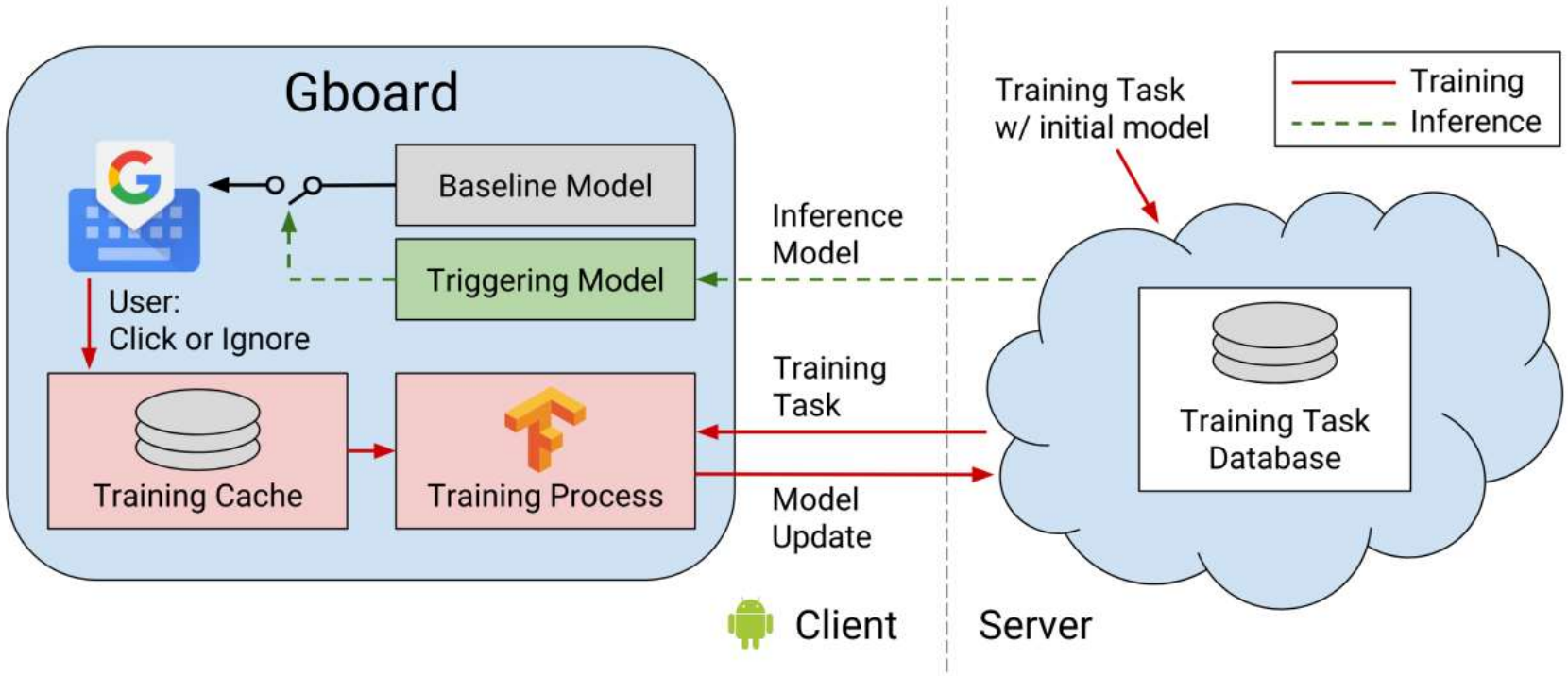


Industrial Study

- Gboard
- Recommender system
- Blockchain
- Autonomous driving
- Health
- IOT
- UAV

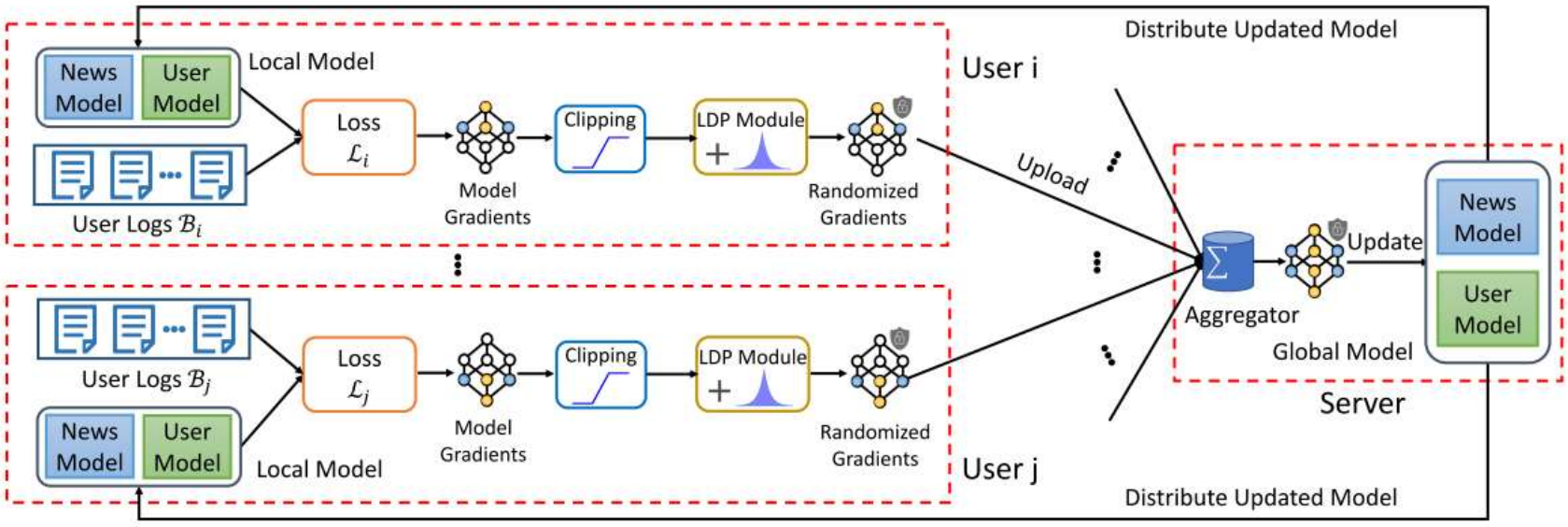
Industrial Study

- Gboard
  - Google's first implementation of federated learning.
  - Triggering model is trained federated to tune the results of the pre-trained baseline model for better performance.



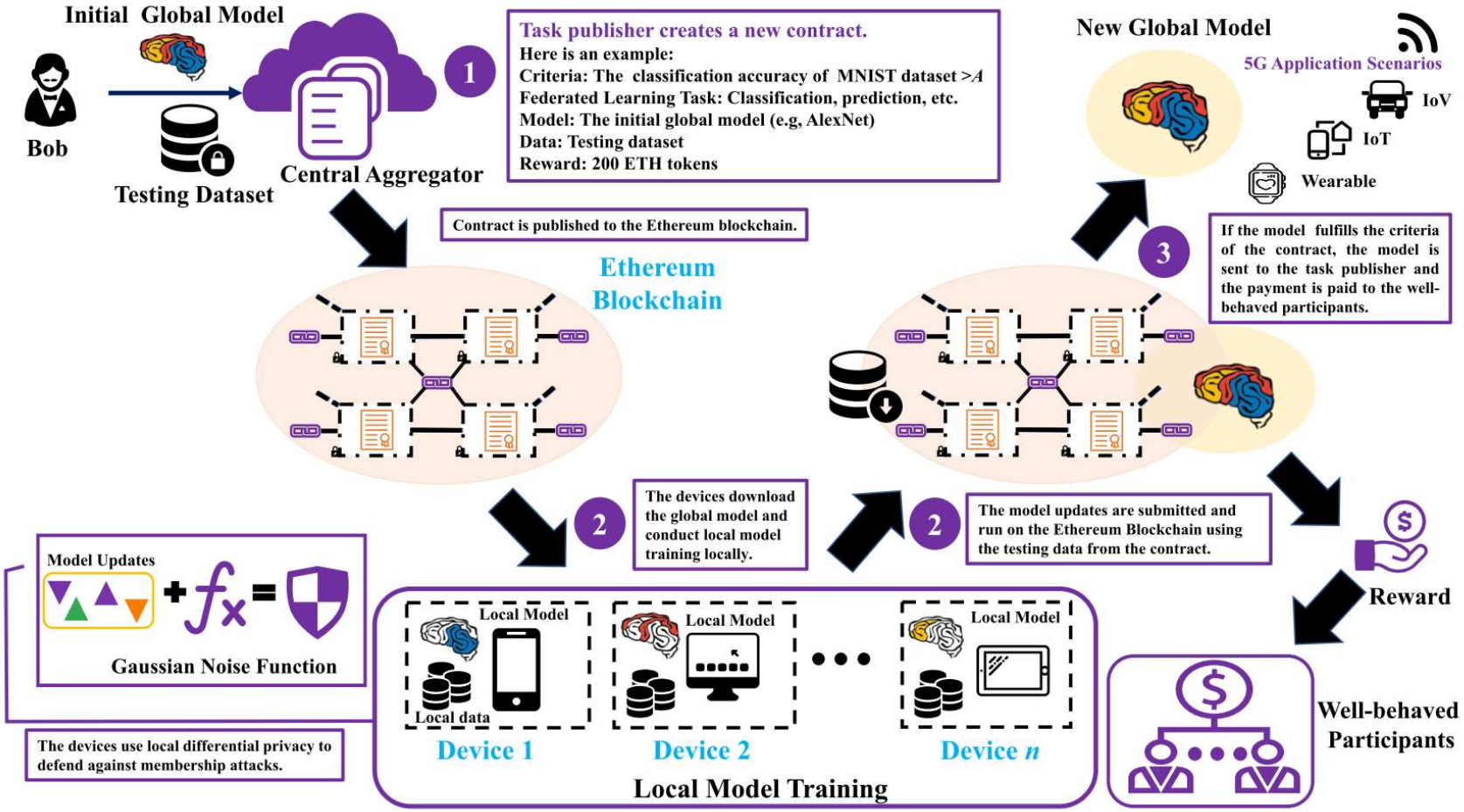
Industrial Study

- Recommender system
  - The news model aims to learn news representations to model news content.
  - The user model is used to learn user representations to model their personal interest.
  - LDP denotes the local differential privacy



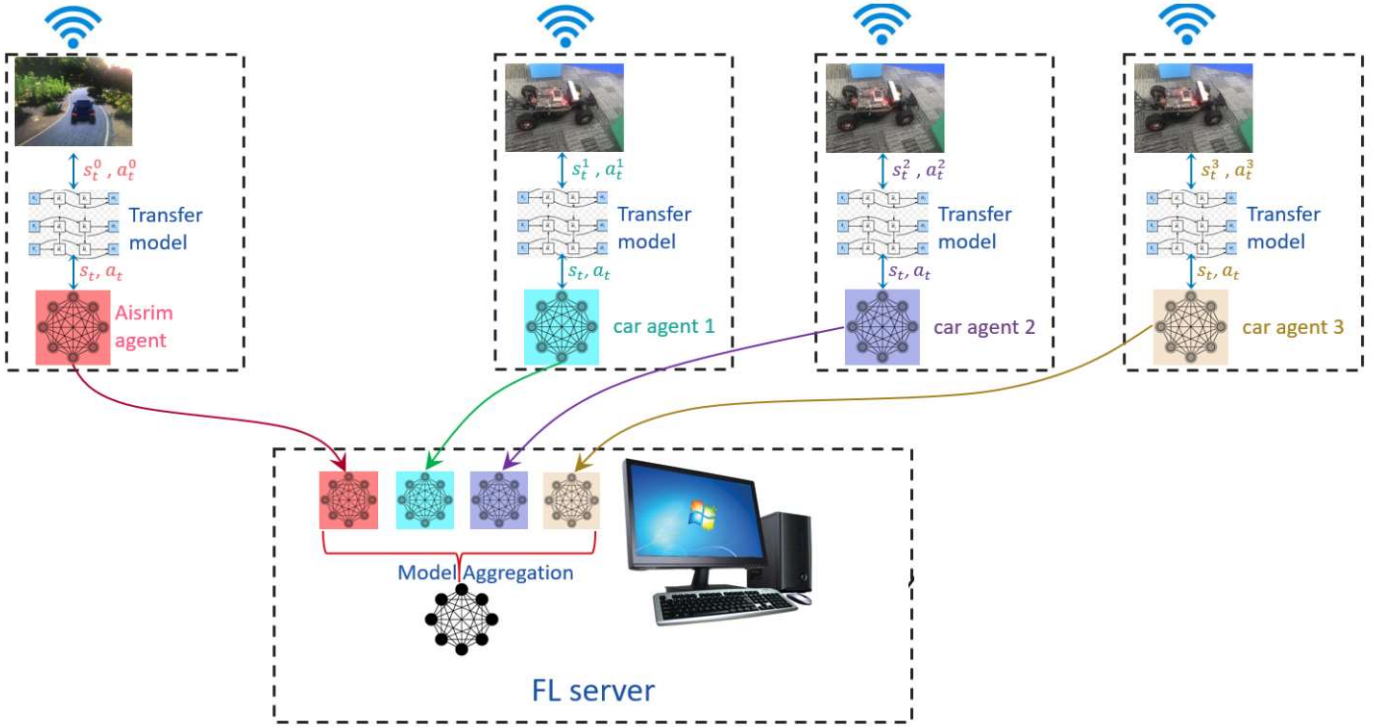
Industrial Study

- Blockchain



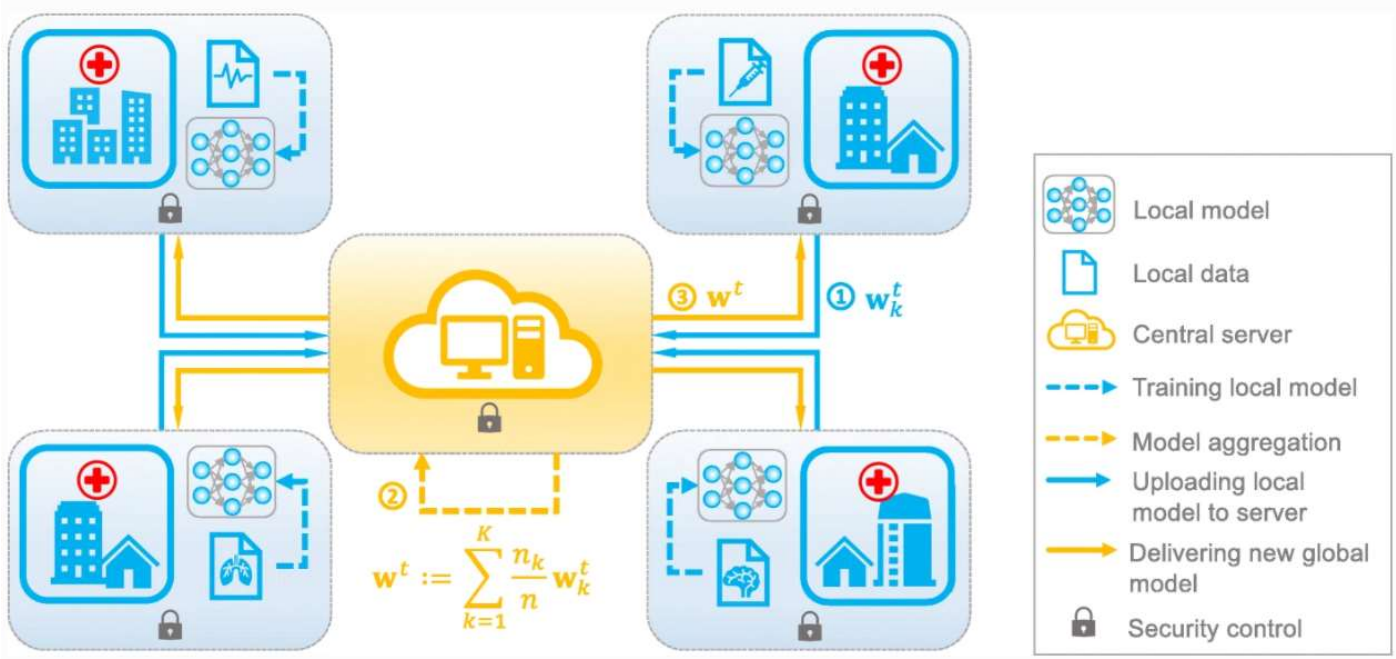
Industrial Study

- Autonomous driving
  - The FTRL framework for collision avoidance RL tasks of autonomous driving cars
  - Global model is asynchronously updated by different RL agents.
  - Transfer knowledge from virtual world (Airsim platform) to real world



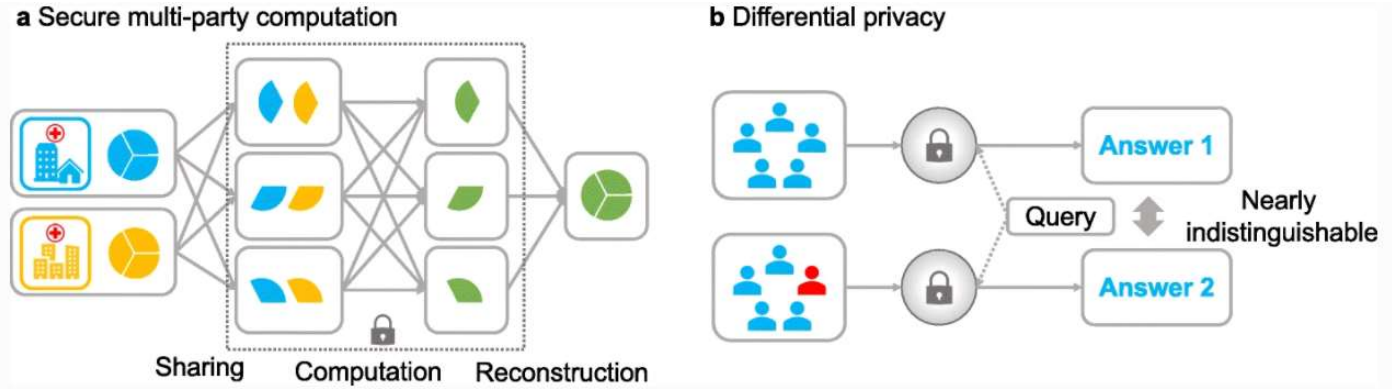
Industrial Study

- Health
  - The workflow



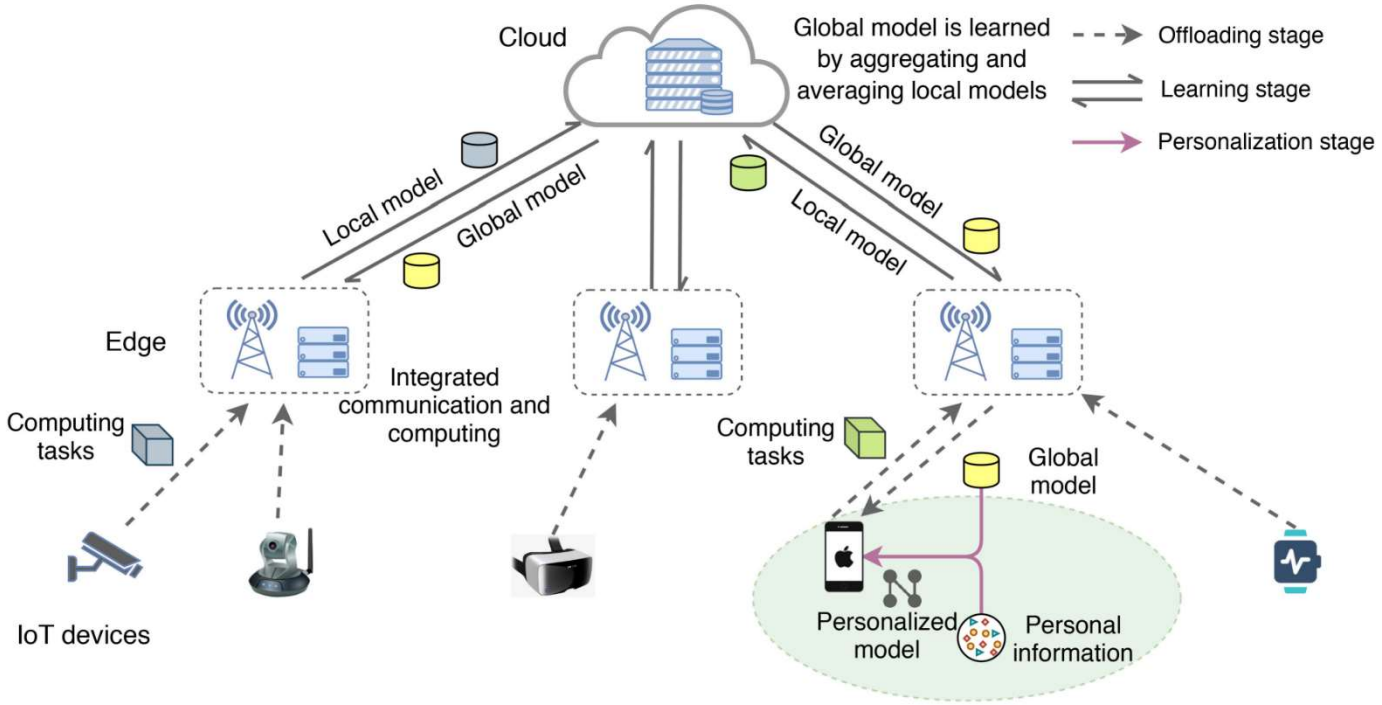
Industrial Study

- Health
  - Security is the most significant consideration.
  - Secure multi-party computation.
  - Differential privacy.



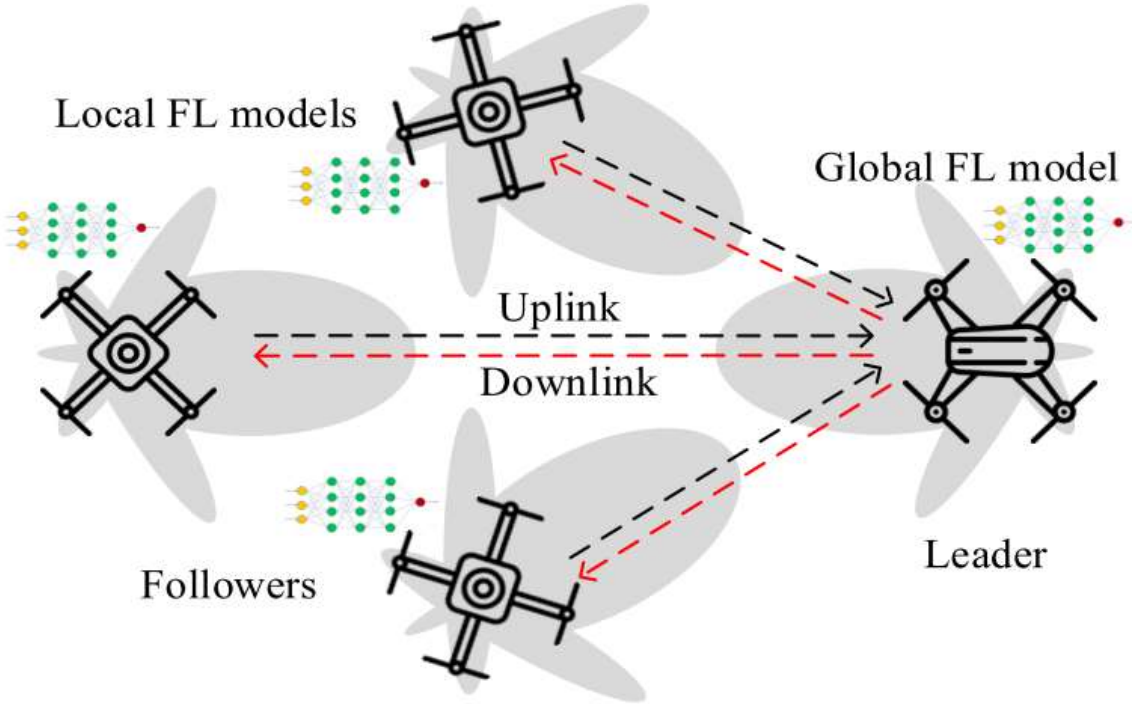
Industrial Study

- IOT
  - Personalized federated learning framework for intelligent IoT applications.
  - Supports flexible selection of personalized federated learning approaches.



Industrial Study

- UAV (Unmanned aerial vehicle)
  - Due to the high mobility of UAVs and their limited energy and stringent energy limitations, the analysis in previous federated learning work cannot be directly applied for UAV swarms.
  - Use a sample average approximation approach from stochastic programming along with a dual method from convex optimization.



Simulation platform

- We introduce the platform from https://github.com/TsingZ0/PFL-Non-IID
  - Environments
  - Datasets
  - Algorithms
  - How to start simulating
  - Practical setting
  - Easy to extend

Simulation platform

- Environments
  - Download conda from https://docs.conda.io/en/latest/miniconda.html and install it.
  - Install all the requested python packages according to *env.yml*
  - Cudatoolkit will be installed by the last step, which means CUDA is ready.

Simulation platform

- Datasets
  - CV datasets: MNIST, Cifar10, Cifar100, Fashion-MNIST
  - NLP datasets: AG_News, Sogou_News

  - In Non-IID setting, there are three situations exist. The first one is the extreme Non-IID setting, the second one is real-world Non-IID setting and the third one is feature skew Non-IID.
  - In the pathological Non-IID setting, for example, the data on each client only contains the specific number of labels (maybe only two labels), though the data on all clients contains 10 labels such as MNIST dataset.
  - In the real-world Non-IID setting, the number of labels for each client is randomly chosen.
  - In the feature skew Non-IID, specific Gaussian noise is added to each clients according to their IDs.

Simulation platform

- Algorithms

- FedAvg — Communication-Efficient Learning of Deep Networks from Decentralized Data *AISTATS 2017*
- Per-FadAvg — Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach *NeurIPS 2020*
- pFedMe — Personalized Federated Learning with Moreau Envelopes *NeurIPS 2020*
- FedProx — Federated Optimization for Heterogeneous Networks *ICLR 2020*
- FedFomo — Personalized Federated Learning with First Order Model Optimization *ICLR 2021*
- MOCHA — Federated multi-task learning *NIPS 2017*
- FedPlayer — Federatedlearning with personalization layers
- FedAMP & HeurFedAMP — Personalized Cross-Silo Federated Learning on Non-IID Data *AAAI 2021*

Simulation platform

- How to start simulating
  - Build dataset: Datasets
  - Train and evaluate the model:
    - cd ./system
    - python main.py -data mnist -m cnn -algo FedAvg -gr 2500 -did 0 -go cnn # for FedAvg and MNIST

    Or you can uncomment the lines you need in ./system/auto_train.sh and run:
    - cd ./system
    - sh auto_train.sh
  - Plot the result test accuracy and training loss curves and save to figures:
    - python plot.py
    - Then check the figures in ./figures.

  Note: All the hyper-parameters have been tuned for all the algorithms, which are recorded in ./system/auto_train.sh

Simulation platform

- Practical setting
  - If you need to simulate FL in a practical setting, which include client dropout, slow trainers, slow senders and network TTL, you can set the following parameters to realize it.
  - Train and evaluate the model:
    - -cdr: The dropout rate for total clients. The selected clients will randomly drop at each training round.
    - -tsr and -ssr: The rates for slow trainers and slow senders among all clients. Once a client was selected as "slow trainers", for example, it will always train slower than original one. So does "slow senders".
    - -tth: The threshold for network TTL (ms).

Simulation platform

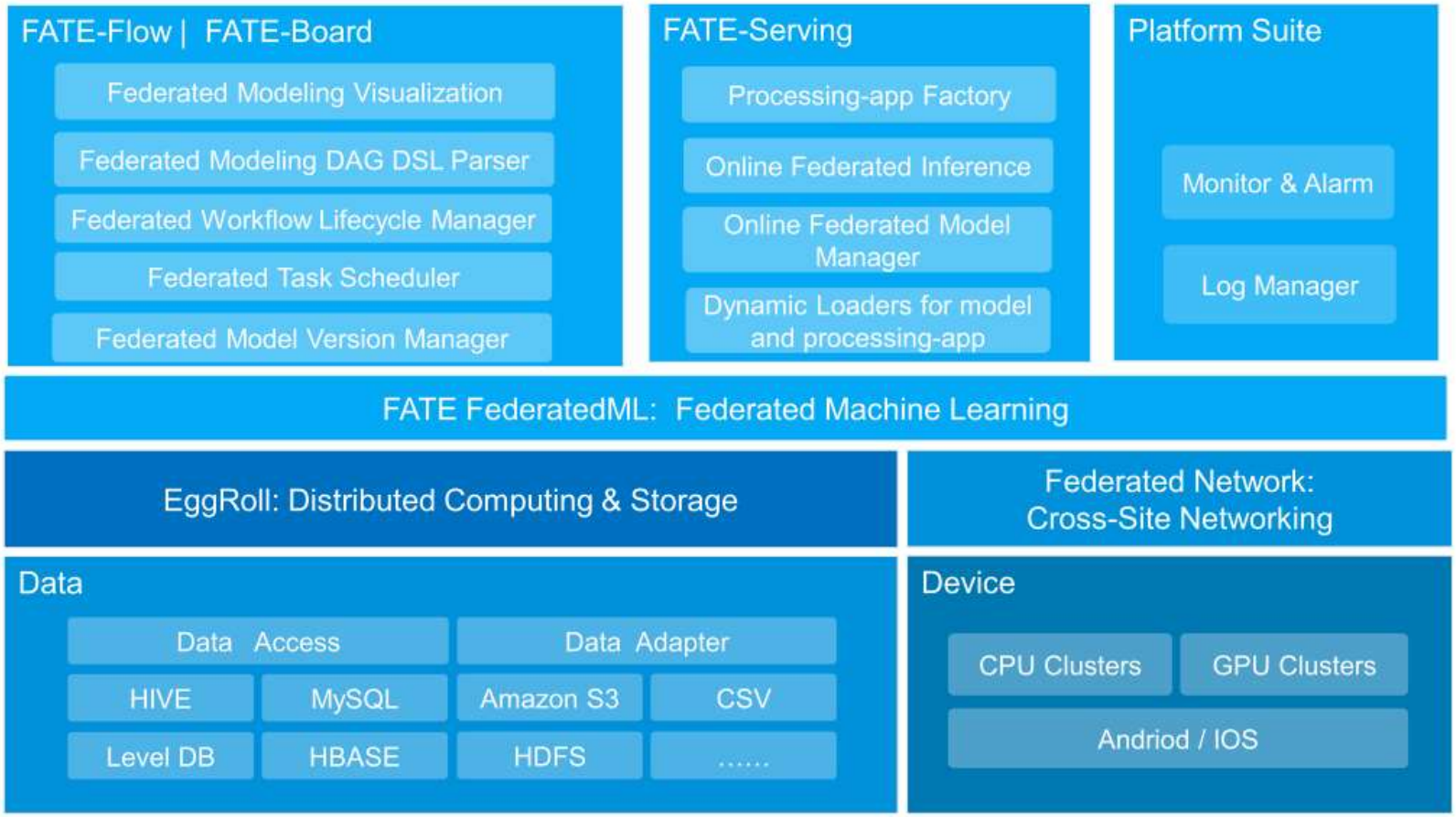
- Easy to extend
  - To add a new dataset into this platform, all you need to do is writing the download code and using the utils the same as ./dataset/generate_mnist.py (you can also consider it as the template).
  - To add a new algorithm, you can utilize the class server and class client, which are wrote in ./system/flcore/servers/serverbase.py and ./system/flcore/clients/clientbase.py, respectively.
  - To add a new model, just add it into ./system/flcore/trainmodel/models.py.
  - If you have your own optimizer while training, please add it into ./system/flcore/optimizers/fedoptimizer.py

Industrial platform (FATE)

- We introduce the platform from https://fate.fedai.org, which is the first published industrial platform that supports standalone-deploy and cluster-deploy using docker or Kubernetes.
  - Technical architecture overview
  - Core function
  - Online inference service

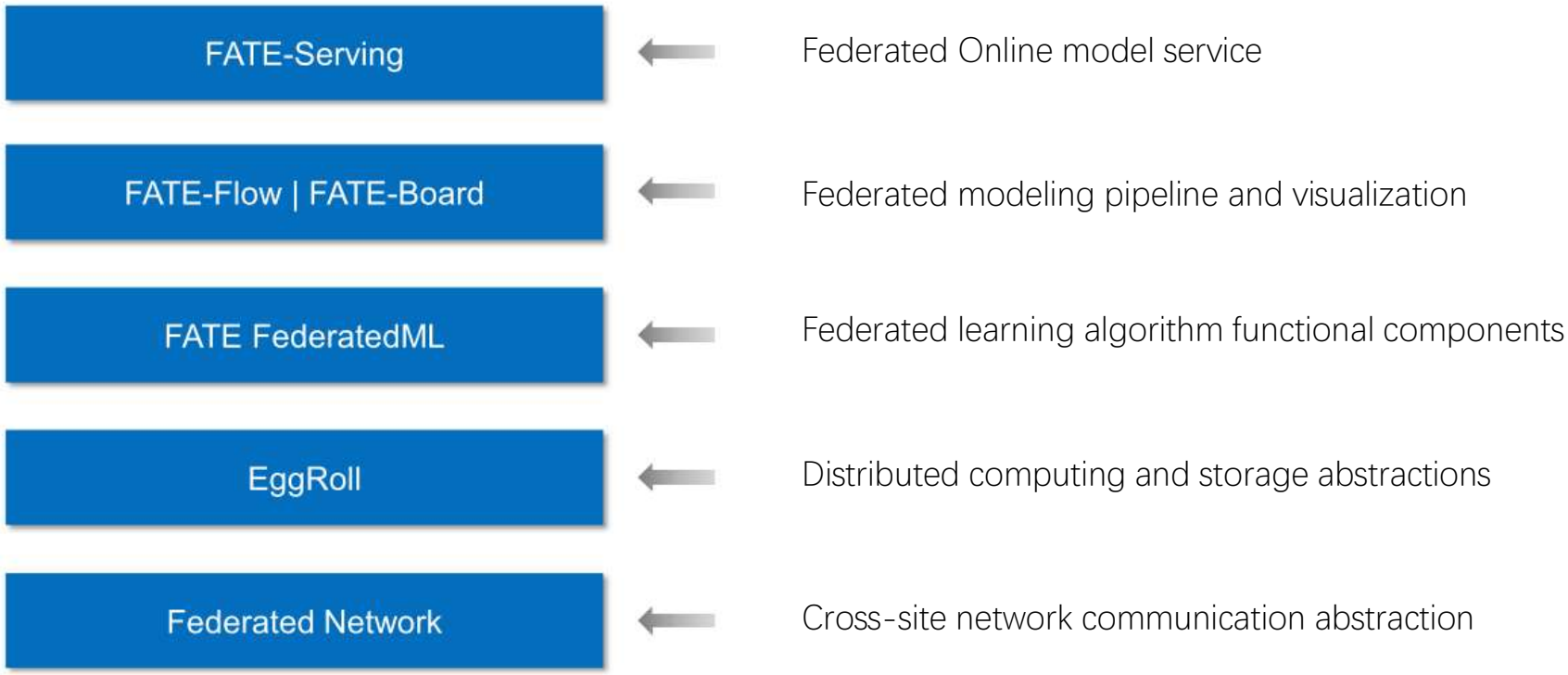
Industrial platform (FATE)

- Technical architecture overview



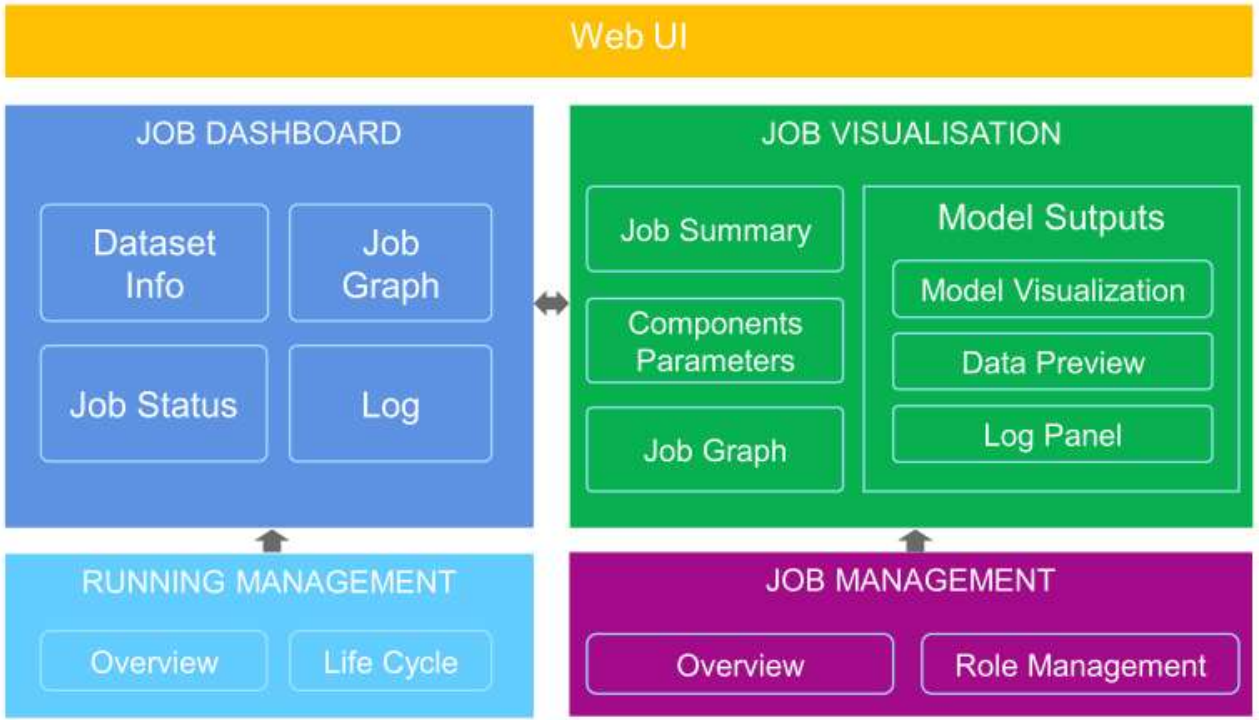
Industrial platform (FATE)

- Core function



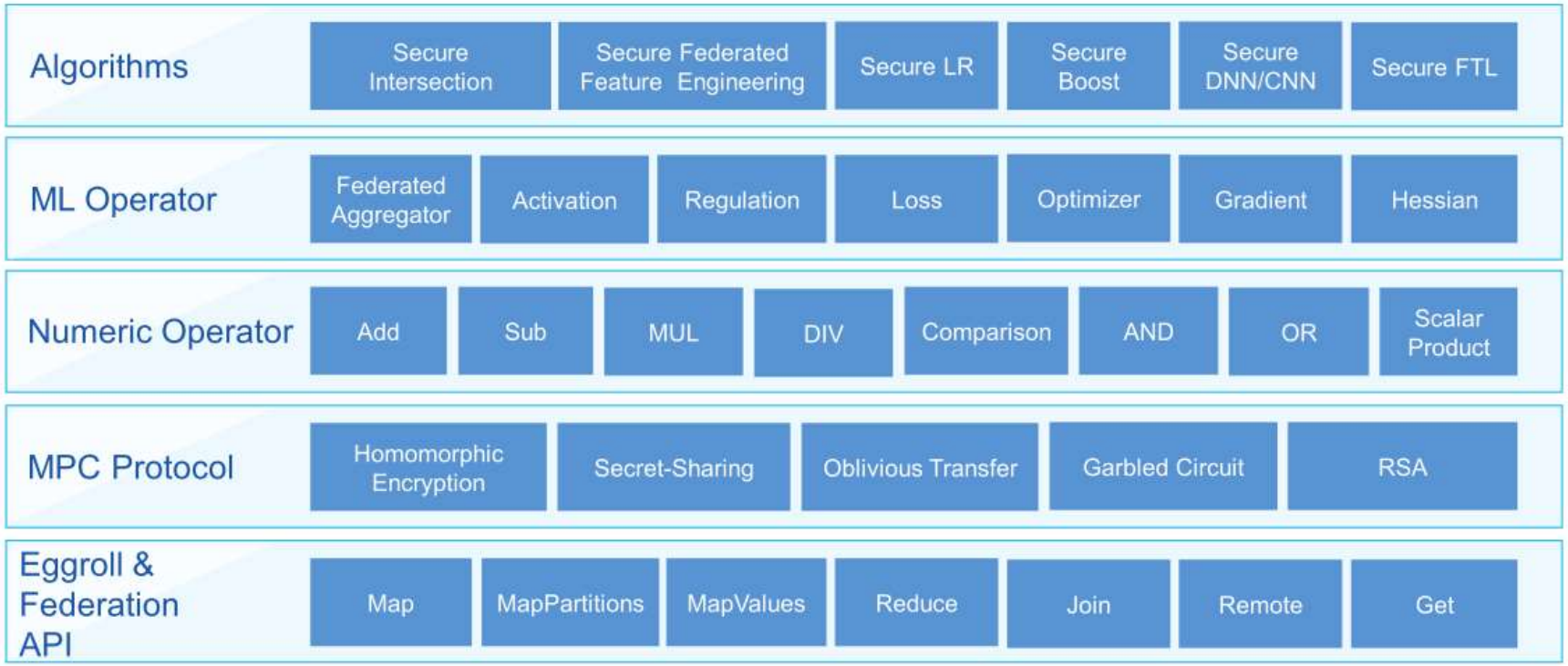
Industrial platform (FATE)

- Core function
  - FATE-Board



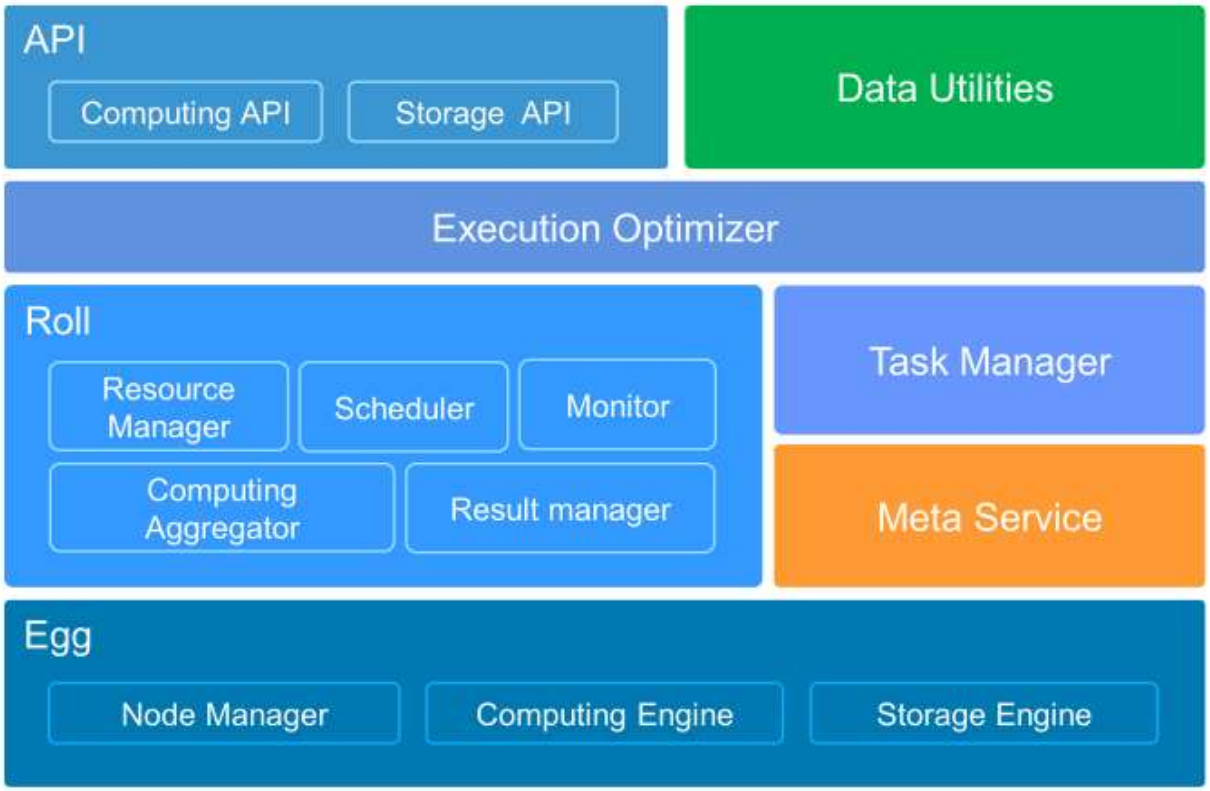
Industrial platform (FATE)

- Core function
  - FederatedML



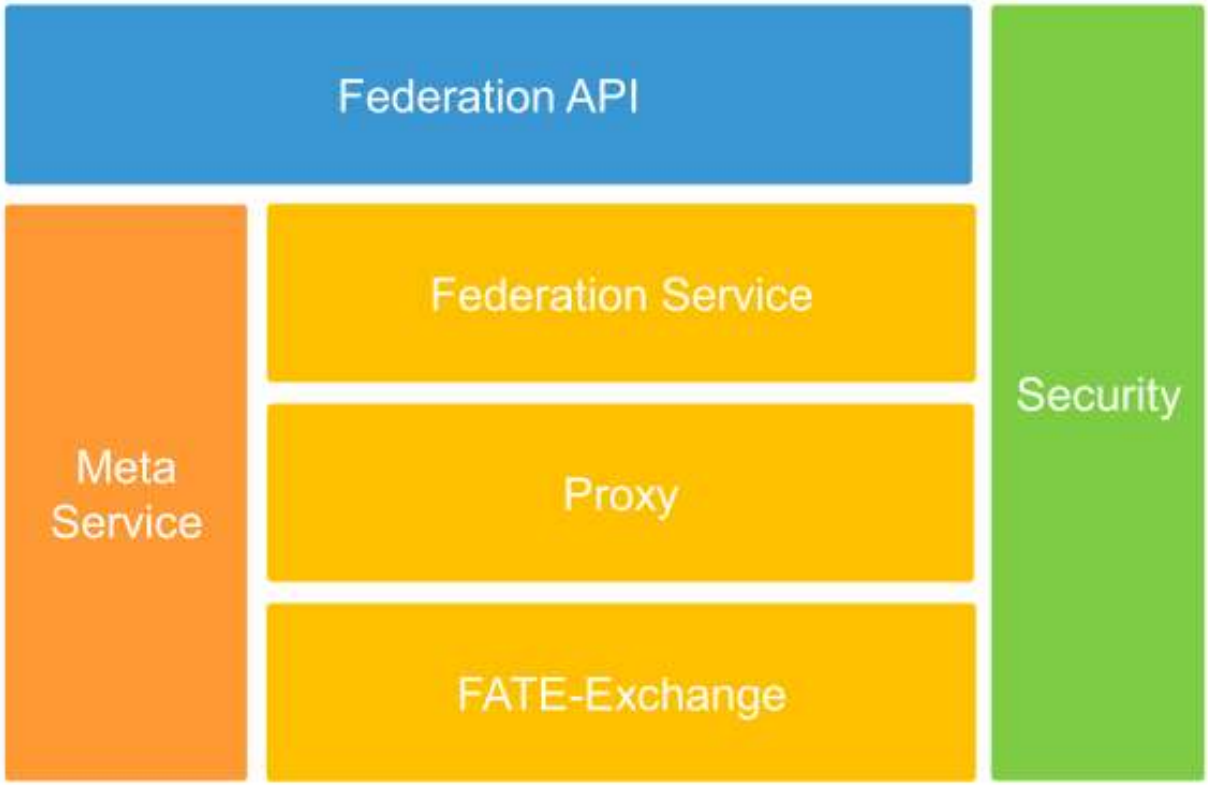
Industrial platform (FATE)

- Core function
  - EggRoll



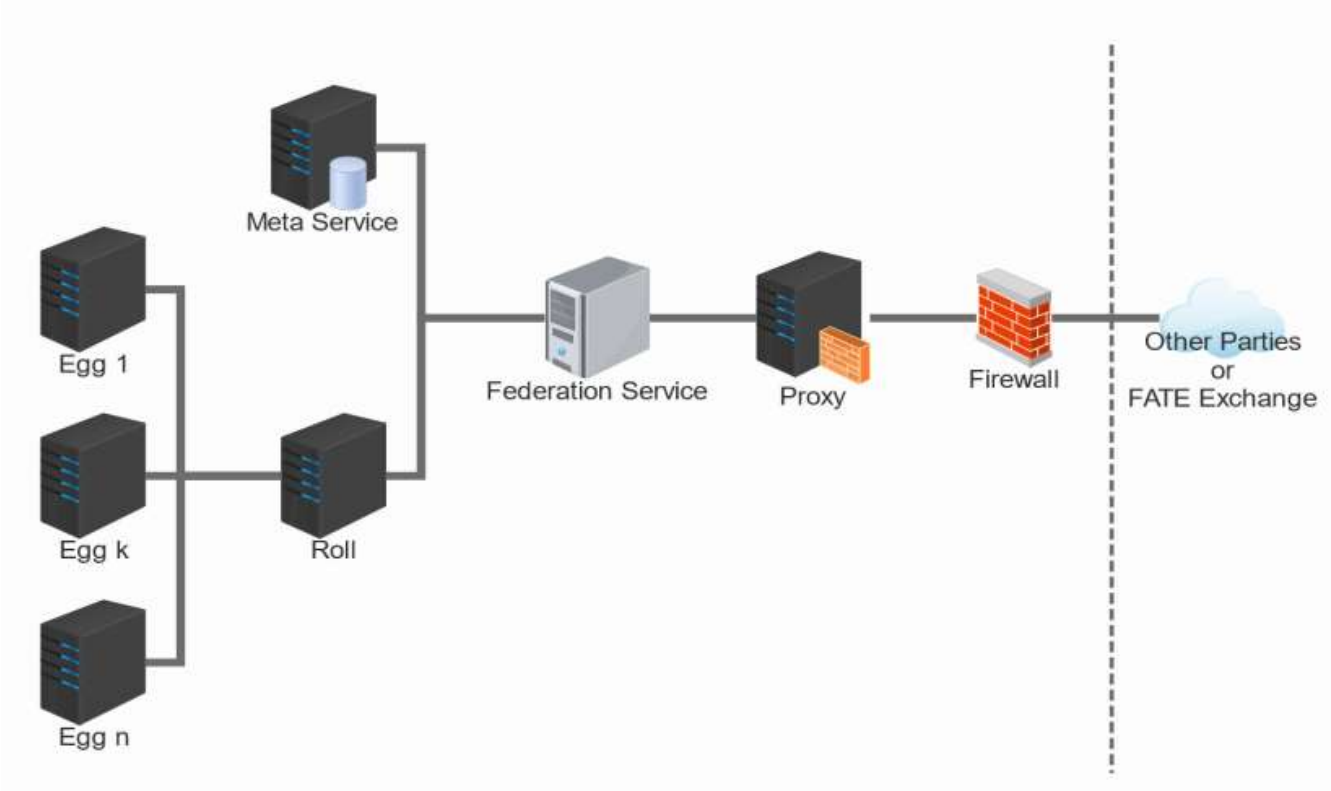
Industrial platform (FATE)

- Core function
  - Federated Network



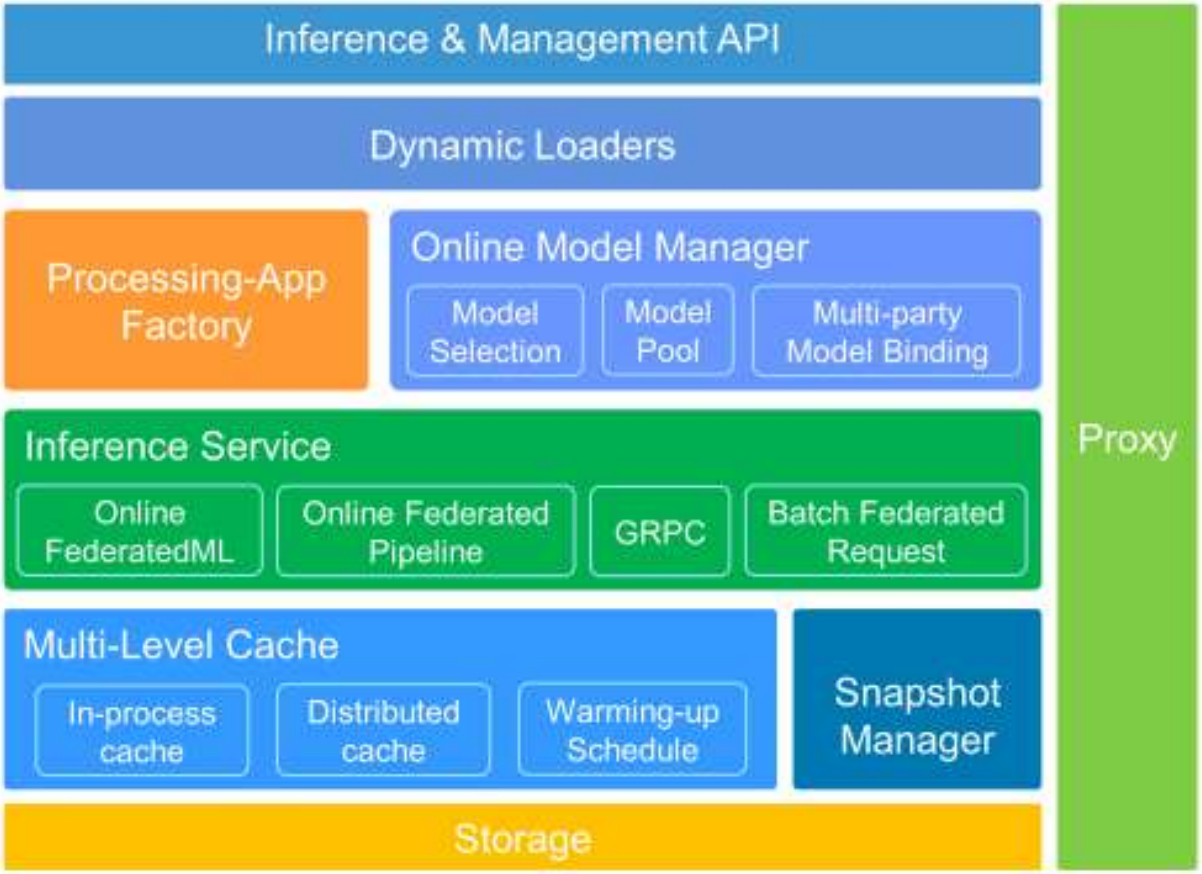
Industrial platform (FATE)

- Core function
  - Federated Network



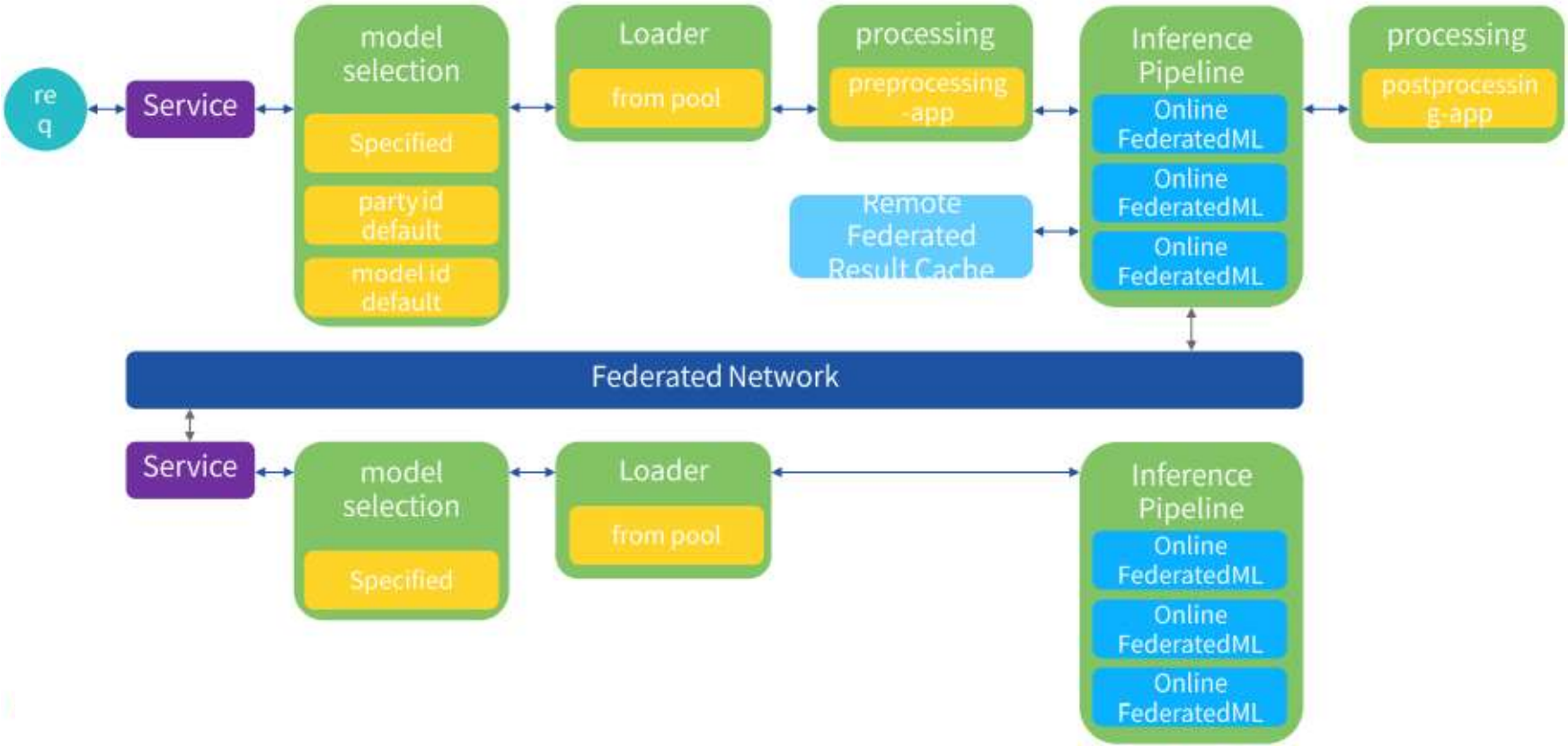
Industrial platform (FATE)

- Online inference service



Industrial platform (FATE)

- Online inference service



Reference

- https://arxiv.org/abs/1908.07873
- https://arxiv.org/abs/1912.07902
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8761315
- https://arxiv.org/pdf/1701.05973.pdf
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9155494
- https://www.aaai.org/AAAI21Papers/AAAI-5802.HuangY.pdf
- https://arxiv.org/abs/2003.09592
- https://ieeexplore.ieee.org/abstract/document/9170265
- https://arxiv.org/pdf/1910.06001.pdf
- https://link.springer.com/article/10.1007/s41666-020-00082-4#Fig1
- https://ieeexplore.ieee.org/document/9090366
- https://ieeexplore.ieee.org/abstract/document/9148776
- https://github.com/FederatedAI/FATE

谢谢！