Federated Machine Learning:

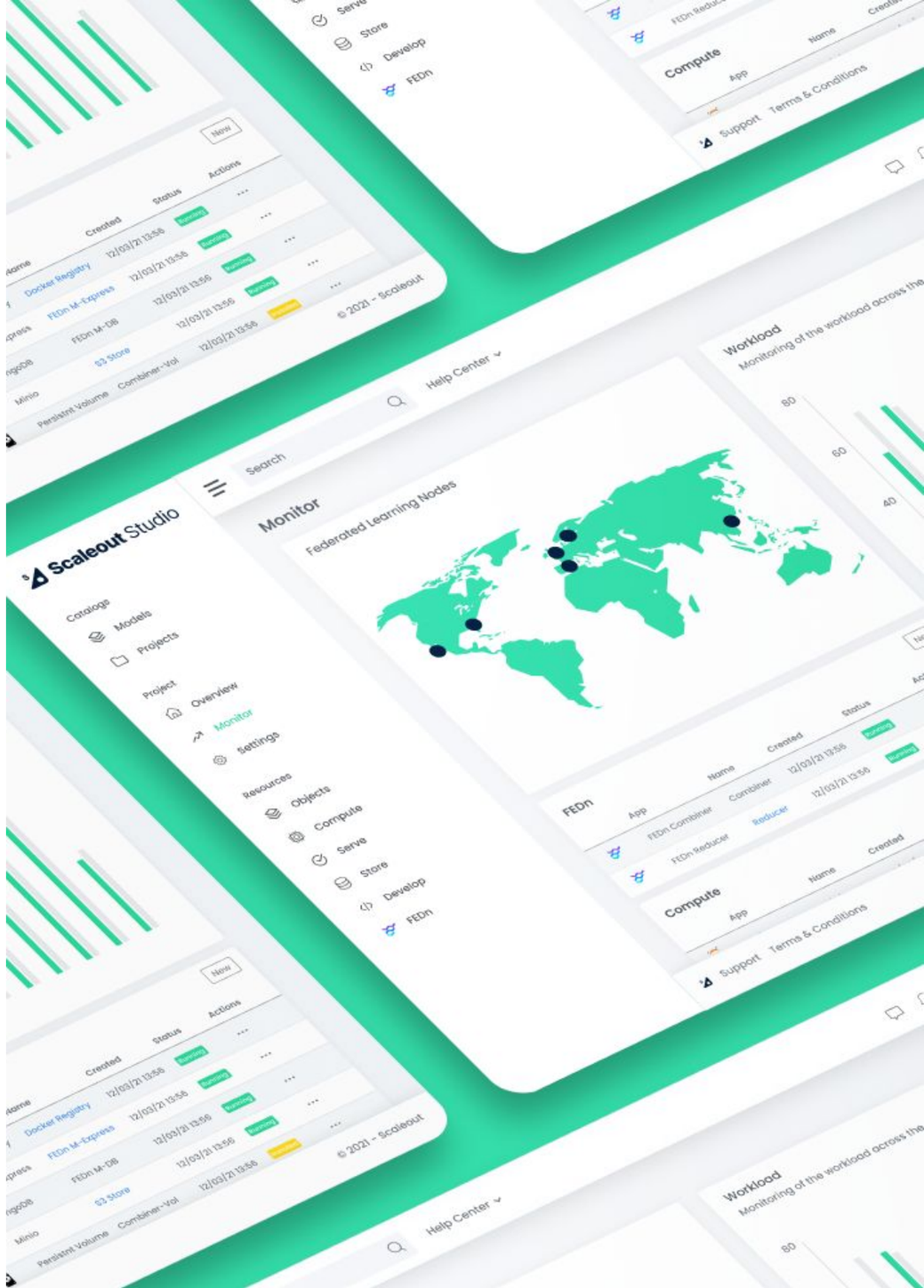
From research to real-world applications

Salman Toor
Associate Professor, Uppsala University
CTO, Scaleout Systems AB.

salman@scaleoutsystems.com

scaleout

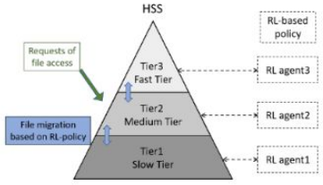
https://www.scaleoutsystems.com



Federated Learning: Distributed Optimization and Applications

Federated machine learning (FedML) involves the development of algorithms and applications that enable distributed machine learning across multiple devices or data sources while maintaining data privacy and security. We study FedML broadly in the group. Some specific recent topics include distributed optimization and developing novel approaches for federated learning that can address challenges such as communication and computation efficiency, data heterogeneity, and model drift. Ultimately, our research aims to advance the state-of-the-art in federated optimization and enable its practical deployment in real-world applications.

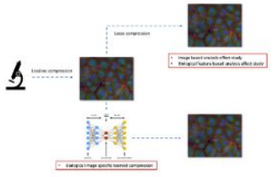
People: Li Ju, Salman Toor, Andreas Hellander



Hierachical Storage Management

In our recent research project 'Efficient Hierarchical Storage Management with Reinforcement Learning', we focus on designing cutting-edge solution for managing large datasets in Hierarchical (Multi-tier) Storage System (HSS). Our recent paper [linked to the text] presented an efficient hierarchical storage management framework using a well-designed autonomous data placement policy empowered by Reinforcement Learning (RL). In the framework we define an environment where its state variables represent the status and properties of each storage tier in HSS, and applied RL to define the data placement policy by finding the best data movement actions.

People: Tianru Zhang, Salman Toor



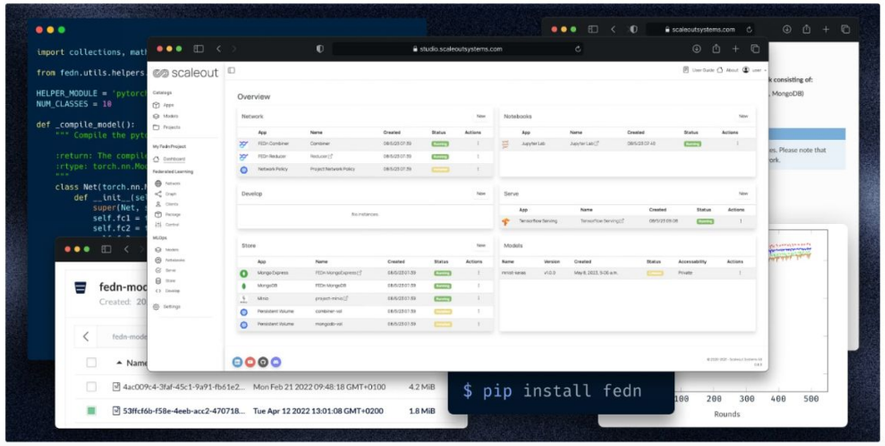
Lossy Compression on Biological Image Analysis

With the development of modern microscopes and experiment automation, a massive amount of microscopy images are generated, which can be gigabytes or even terabytes of data in a single experiment. To deal with the explosion of microscopy image data, lossy compression could be a potential way, however, its effects on analysis are unclear. This project, as a part of the HASTE project, mainly focuses on studying the effects of lossy compression on biological image analysis, and aims to propose a learned compression technique that specifically devised for biological images. Different of existing lossy image compression methods designed for human vision, this technique will maintain the information of

scaleout

From R&D to real-world federated learning

The FEDn framework enables seamless development and deployment of federated learning (FL) applications, from local proofs-of-concept to distributed real-world settings.



Introduction to Federated Machine Learning

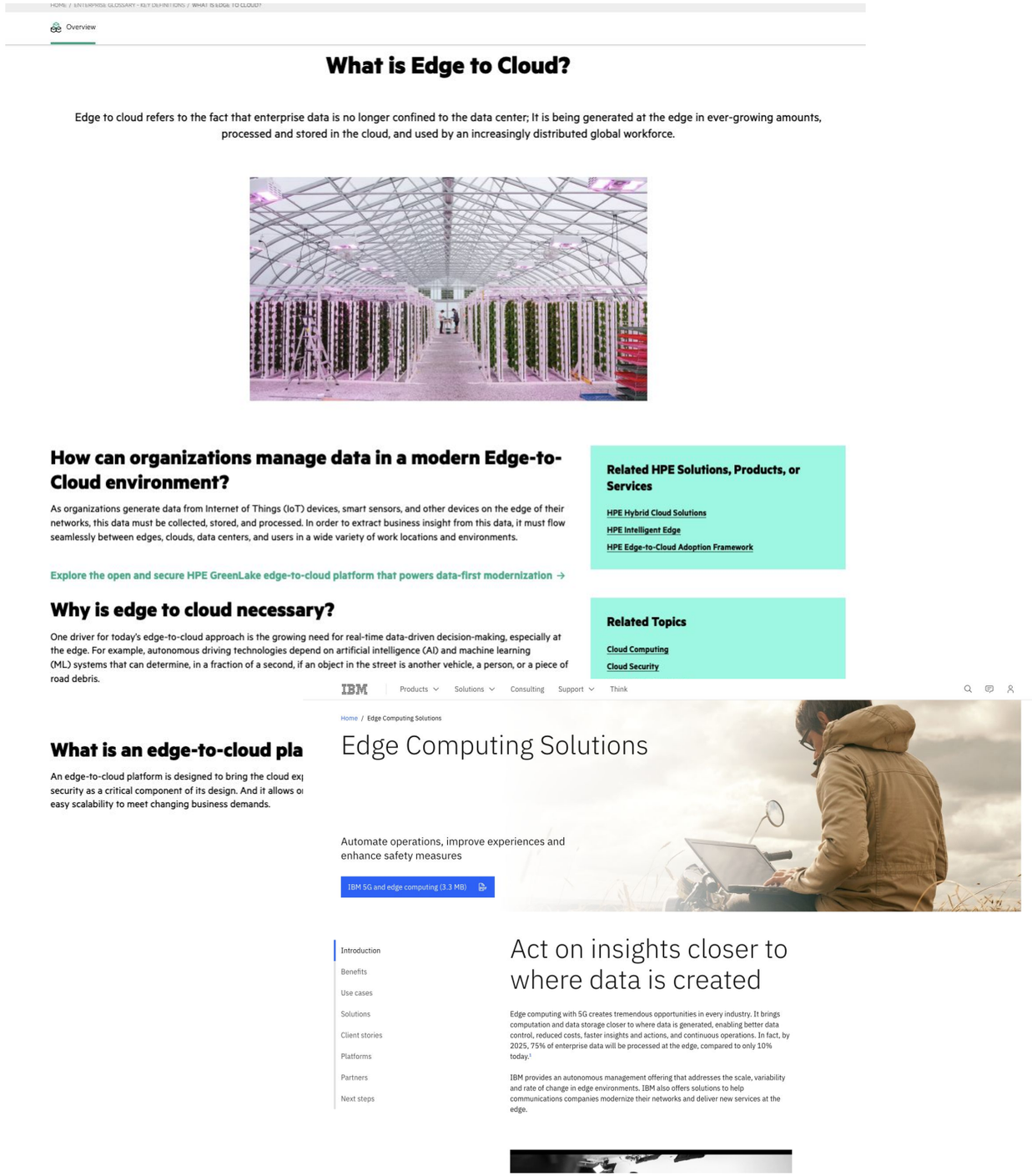
Macrotrend

The shift from central cloud to edge

We are early in a shift from a predominantly centralized cloud to a distributed cloud.

This development is driven by the fact that data is increasingly generated at the computational edge.

There are cost, security and privacy challenges in moving that data to the central cloud.



Overview

What is Edge to Cloud?

Edge to cloud refers to the fact that enterprise data is no longer confined to the data center; It is being generated at the edge in ever-growing amounts, processed and stored in the cloud, and used by an increasingly distributed global workforce.

How can organizations manage data in a modern Edge-to-Cloud environment?

As organizations generate data from Internet of Things (IoT) devices, smart sensors, and other devices on the edge of their networks, this data must be collected, stored, and processed. In order to extract business insight from this data, it must flow seamlessly between edges, clouds, data centers, and users in a wide variety of work locations and environments.

Explore the open and secure HPE GreenLake edge-to-cloud platform that powers data-first modernization →

Why is edge to cloud necessary?

One driver for today's edge-to-cloud approach is the growing need for real-time data-driven decision-making, especially at the edge. For example, autonomous driving technologies depend on artificial intelligence (AI) and machine learning (ML) systems that can determine, in a fraction of a second, if an object in the street is another vehicle, a person, or a piece of road debris.

Related HPE Solutions, Products, or Services

HPE Hybrid Cloud Solutions
HPE Intelligent Edge
HPE Edge-to-Cloud Adoption Framework

Related Topics

Cloud Computing
Cloud Security

What is an edge-to-cloud pla

An edge-to-cloud platform is designed to bring the cloud ex security as a critical component of its design. And it allows o easy scalability to meet changing business demands.

IBM Products Solutions Consulting Support Think

Home / Edge Computing Solutions

Edge Computing Solutions

Automate operations, improve experiences and enhance safety measures

IBM 5G and edge computing (3.3 MB)

Introduction
Benefits
Use cases
Solutions
Client stories
Platforms
Partners
Next steps

Act on insights closer to where data is created

Edge computing with 5G creates tremendous opportunities in every industry. It brings computation and data storage closer to where data is generated, enabling better data control, reduced costs, faster insights and actions, and continuous operations. In fact, by 2025, 75% of enterprise data will be processed at the edge, compared to only 10% today.[1]

IBM provides an autonomous management offering that addresses the scale, variability and rate of change in edge environments. IBM also offers solutions to help communications companies modernize their networks and deliver new services at the edge.

Barriers to pooling data

“Don’t want to share”

Isolated data

- Private data
- Proprietary data
- Business sensitive
- Cybersecurity

“Not allowed to share”

Regulated data

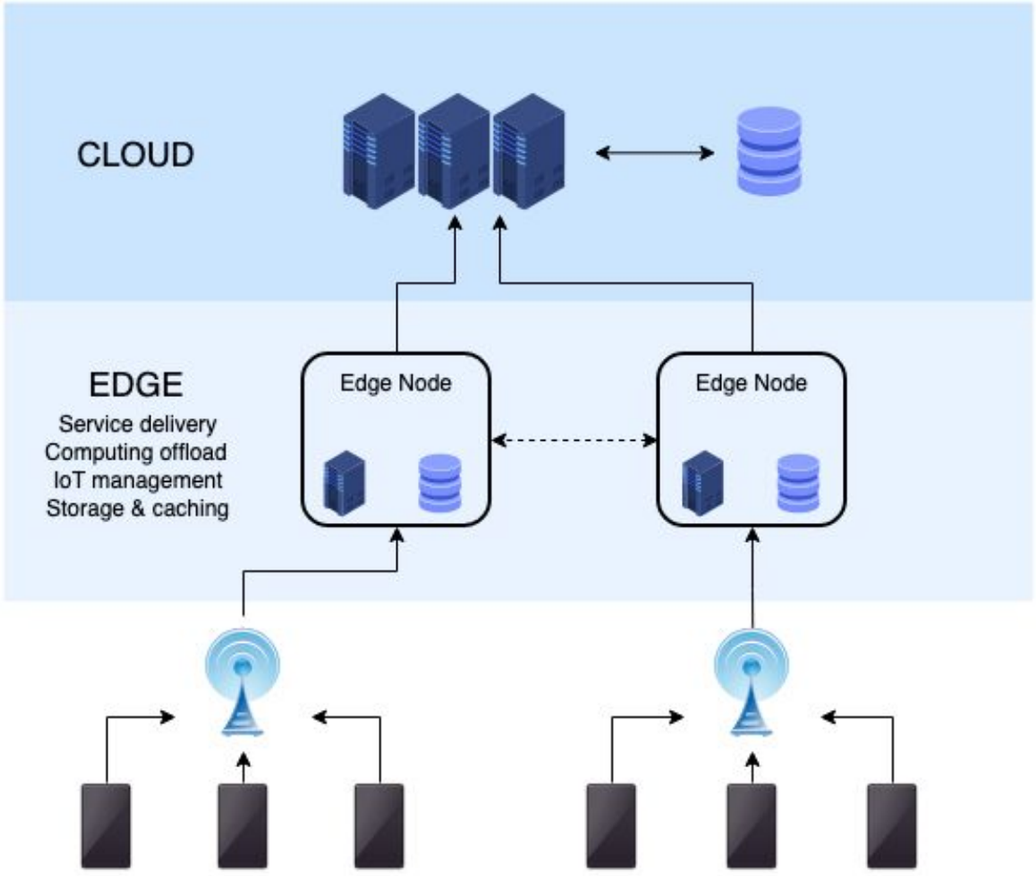
- Data privacy protection
- Data transfer law
- Health data

“Cannot share”

Practical blockers

- Siloed data centers
- High frequency data
  - edge, IoT
- Large datasets
- Networks latency

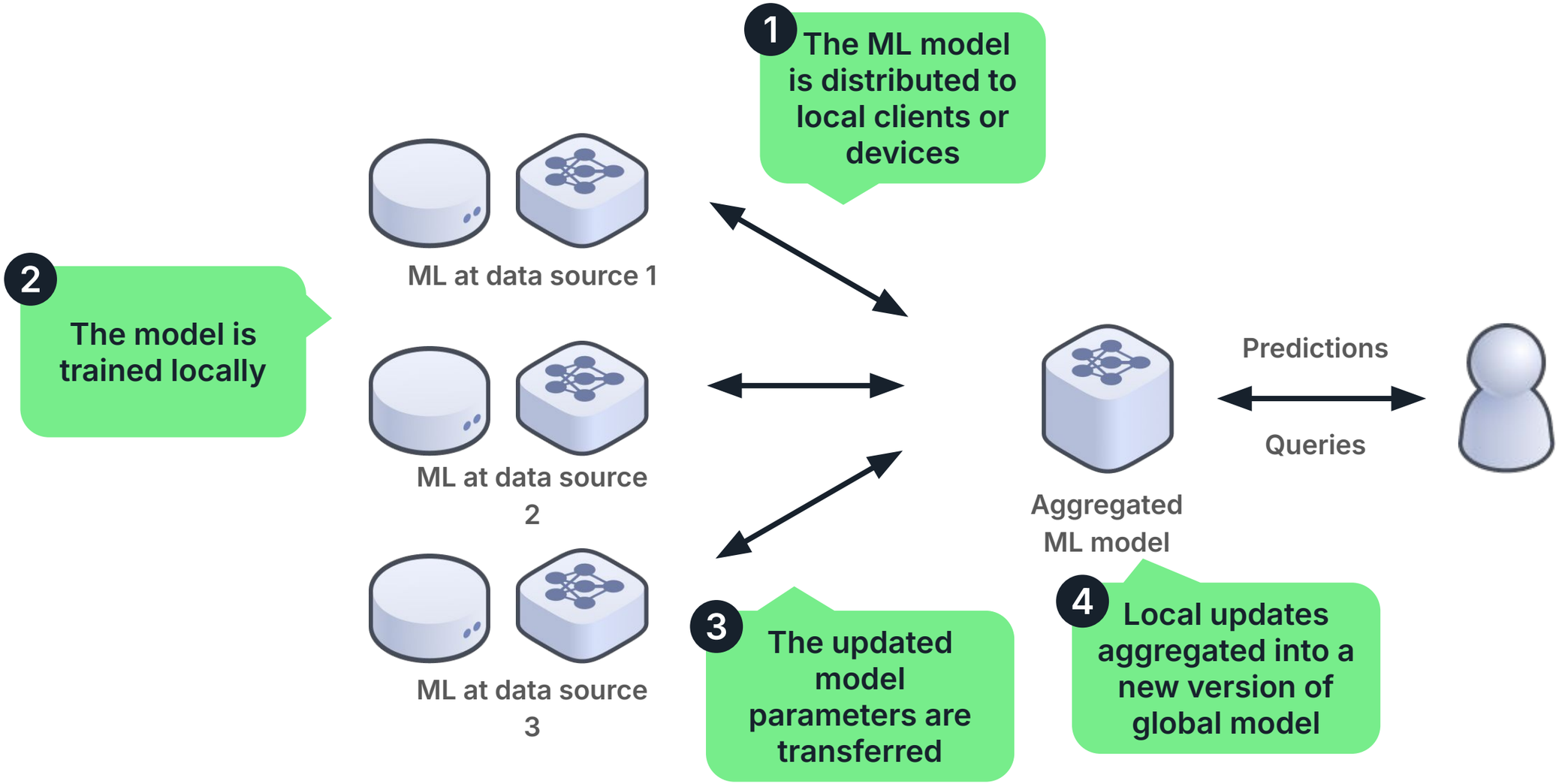
Edge computing



By NoMore201 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=82034067

- How to manage the infrastructure?
- How to secure a decentralized in infrastructure?
- How to program and operate applications across the layers?
- …
- …
- …
- **How to develop and manage ML models?**

Federated machine learning



Federated Averaging

1. Out of K clients, pick a fraction C to do a global model update.

2. Perform E epochs of SGD on local minibatch of size B.

3. Average locally updated weights.

McMahan et al.
https://arxiv.org/abs/1602.05629

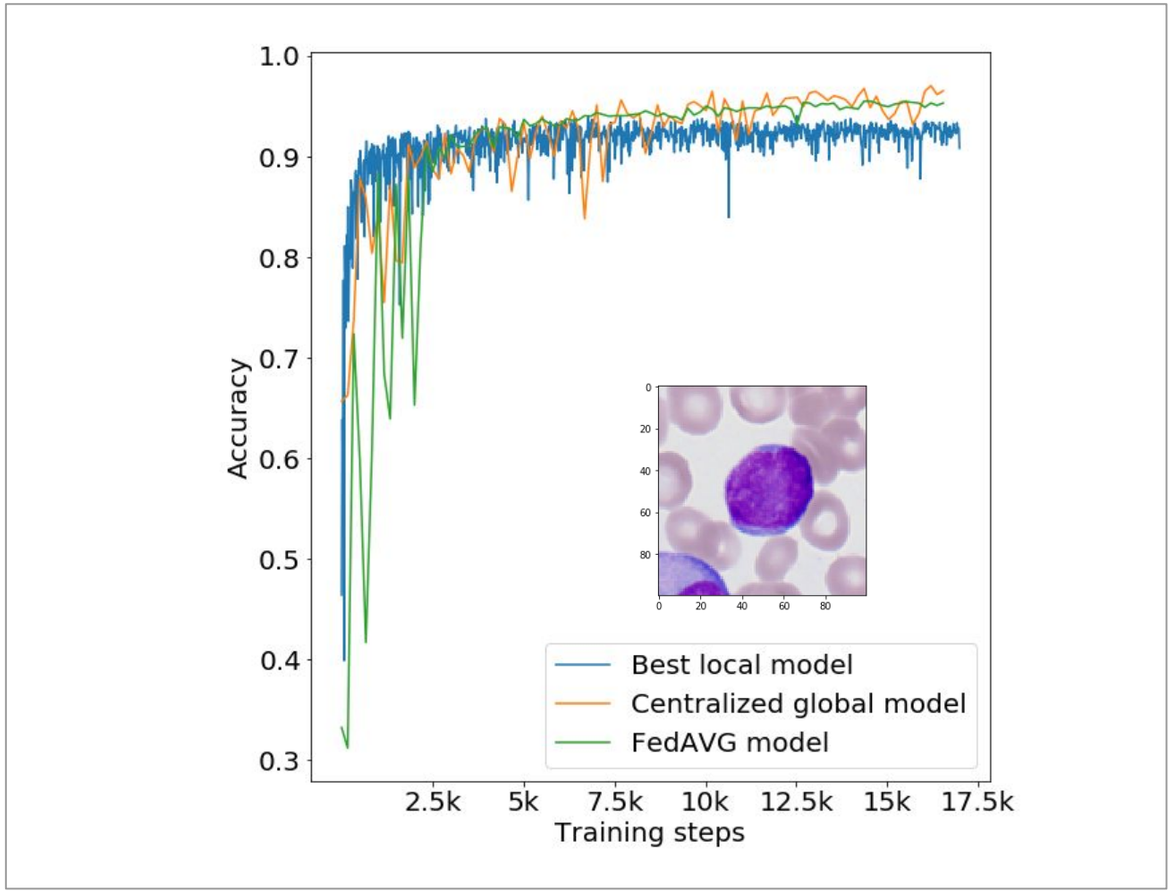
Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
initialize w_0
for each round $t = 1, 2, \dots$ **do**
$m \leftarrow \max(C \cdot K, 1)$
$S_t \leftarrow$ (random set of m clients)
for each client $k \in S_t$ **in parallel do**
$w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate(k, w): *// Run on client k*
$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
for each local epoch i from 1 to E **do**
for batch $b \in \mathcal{B}$ **do**
$w \leftarrow w - \eta \nabla \ell(w; b)$
return w to server

The promise of federated learning

A federated model that (a) performs on par with or better than the best possible model trained pooled data, and (b) outperforms the best model attainable by clients in isolation.



Acute Myeloid Leukemia cell type and stage classification, Scaleout study.

Fleet Intelligence

Scania CV

Scania develops federated learning to leverage more data on-vehicle, looking to overcome communication bandwidth and regulatory limits.

Use-cases include:

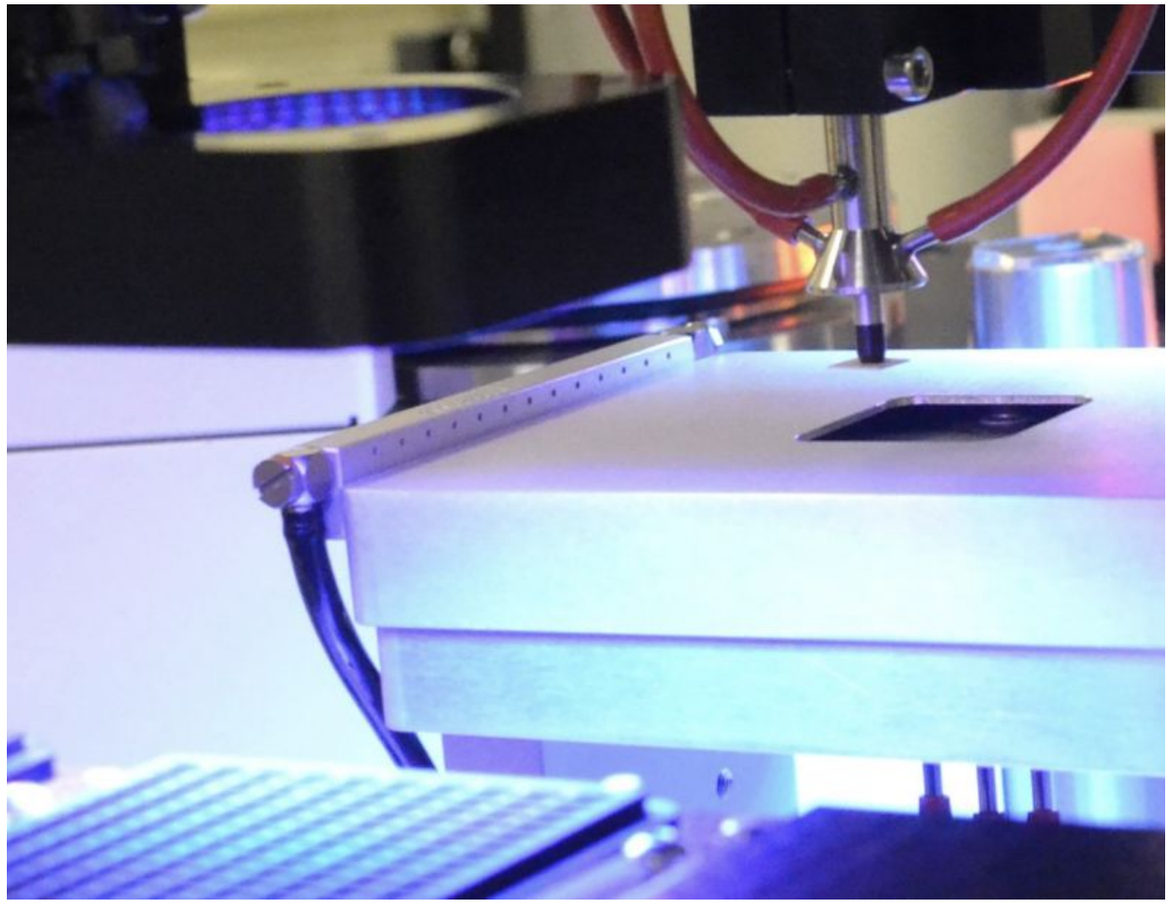
- Predictive maintenance (sensor data)
- Intrusion Detection Systems (IDS) (CAN and Ethernet)

SCANIA

Industry robots

Mycronic

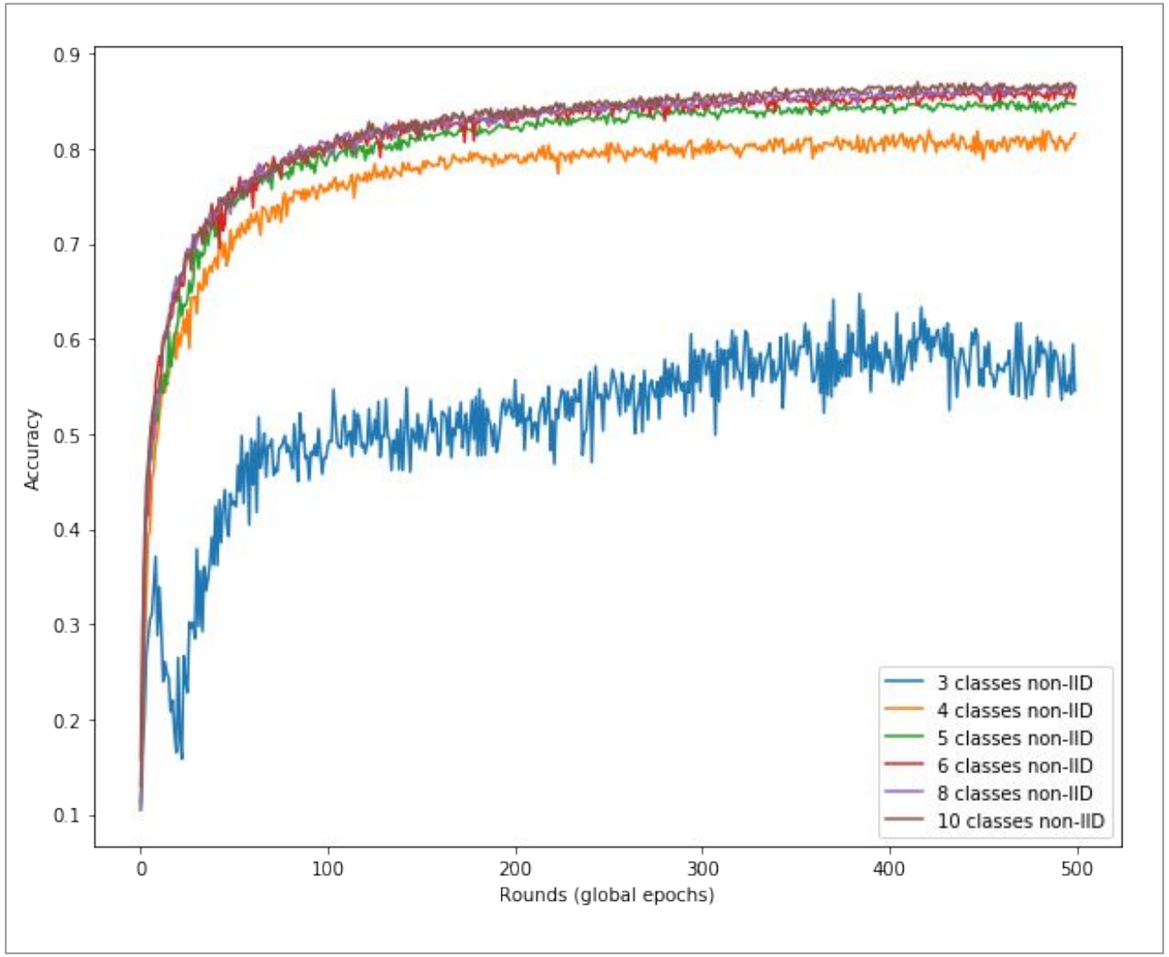
Mycronic develops federated learning to enhance the AI model training on Pick-and-Place machines without the need for centralizing sensitive data.



MYCRONIC

Key challenge 1: statistical heterogeneity

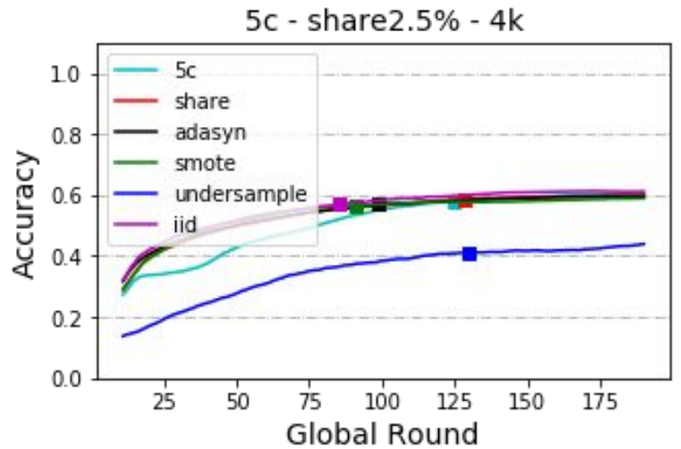
Unbalanced data distributions / statistical heterogeneity



Cifar10, Scaleout study.

If data is highly non IID across clients accuracy and convergence can be affected.

Leverage a small amount of shared data and data augmentation (see e.g. https://arxiv.org/pdf/1806.00582.pdf)



SCALEOUT

“Model Fairness”

Computer Science > Machine Learning

[Submitted on 23 Jan 2023]

Accelerating Fair Federated Learning: Adaptive Federated Adam

Li Ju, Tianru Zhang, Salman Toor, Andreas Hellander

Federated learning is a distributed and privacy-preserving approach to train a statistical model collaboratively from decentralized data of different parties. However, when datasets of participants are not independent and identically distributed (non-IID), models trained by naive federated algorithms may be biased towards certain participants, and model performance across participants is non-uniform. This is known as the fairness problem in federated learning. In this paper, we formulate fairness-controlled federated learning as a dynamical multi-objective optimization problem to ensure fair performance across all participants. To solve the problem efficiently, we study the convergence and bias of Adam as the server optimizer in federated learning, and propose Adaptive Federated Adam (AdaFedAdam) to accelerate fair federated learning with alleviated bias. We validated the effectiveness, Pareto optimality and robustness of AdaFedAdam in numerical experiments and show that AdaFedAdam outperforms existing algorithms, providing better convergence and fairness properties of the federated scheme.

Subjects: **Machine Learning (cs.LG)**; Distributed, Parallel, and Cluster Computing (cs.DC)
Cite as: arXiv:2301.09357 **[cs.LG]**
(or arXiv:2301.09357v1 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2301.09357

Submission history

From: Li Ju [view email]
[v1] Mon, 23 Jan 2023 10:56:12 UTC (366 KB)

https://arxiv.org/pdf/2301.09357.pdf

Federated Learning != Distributed ML

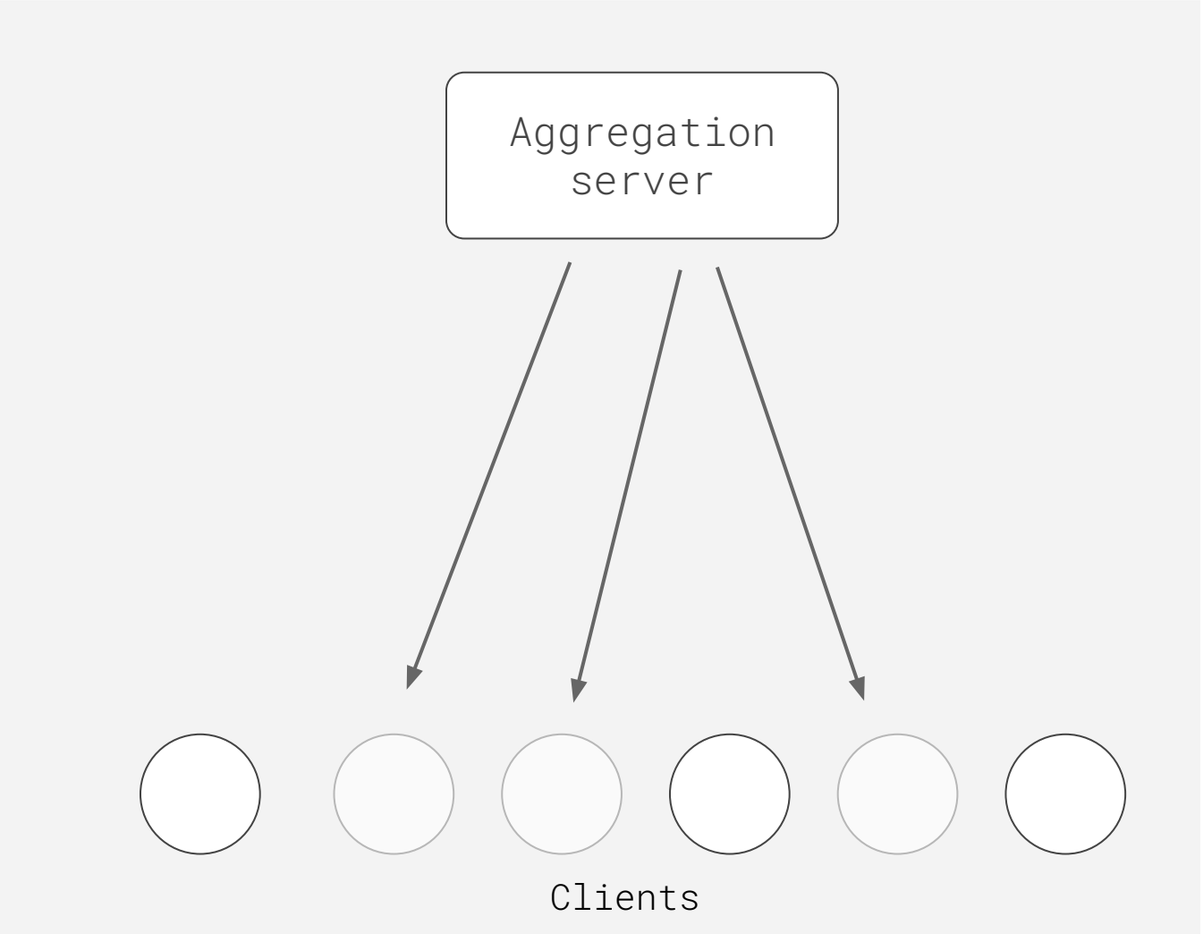
Scaling federated learning

Key challenge 3: scalability

A central server orchestrates training and model aggregation for N clients, where N can be a large number.

A client-server problem. Challenges include:

- Network bottlenecks
- Large models become bottlenecks
- Aggregator can become bottleneck



Challenges related to federated learning

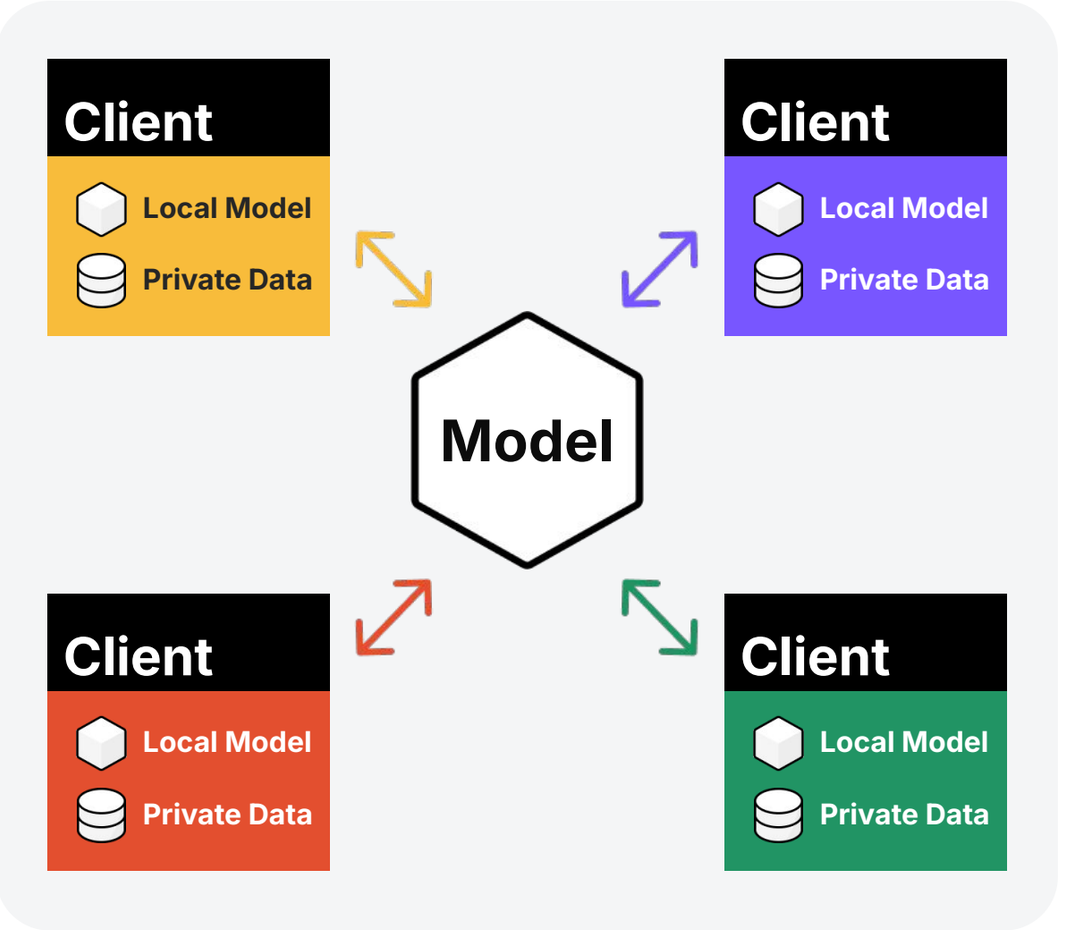
Cross-silo case

- Reliable model sharing
- Aggregations of large local models

Cross-device case

- Reliable communication
- Model aggregation based on millions of devices

The base architecture is not designed to handle real-world use cases



Addressing scalability

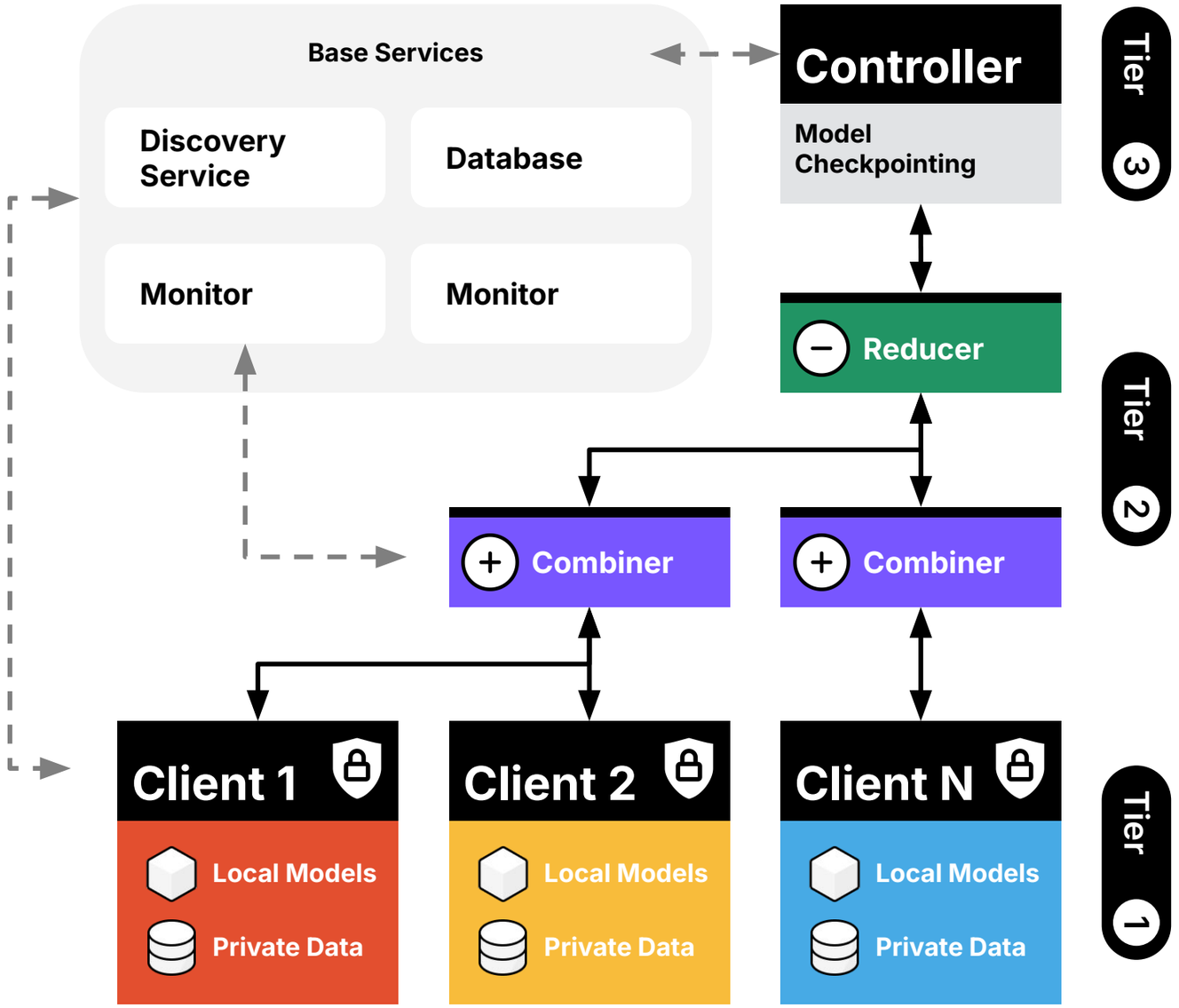
Tier 1 - Geographically distributed clients

Tier 2 - Model aggregation

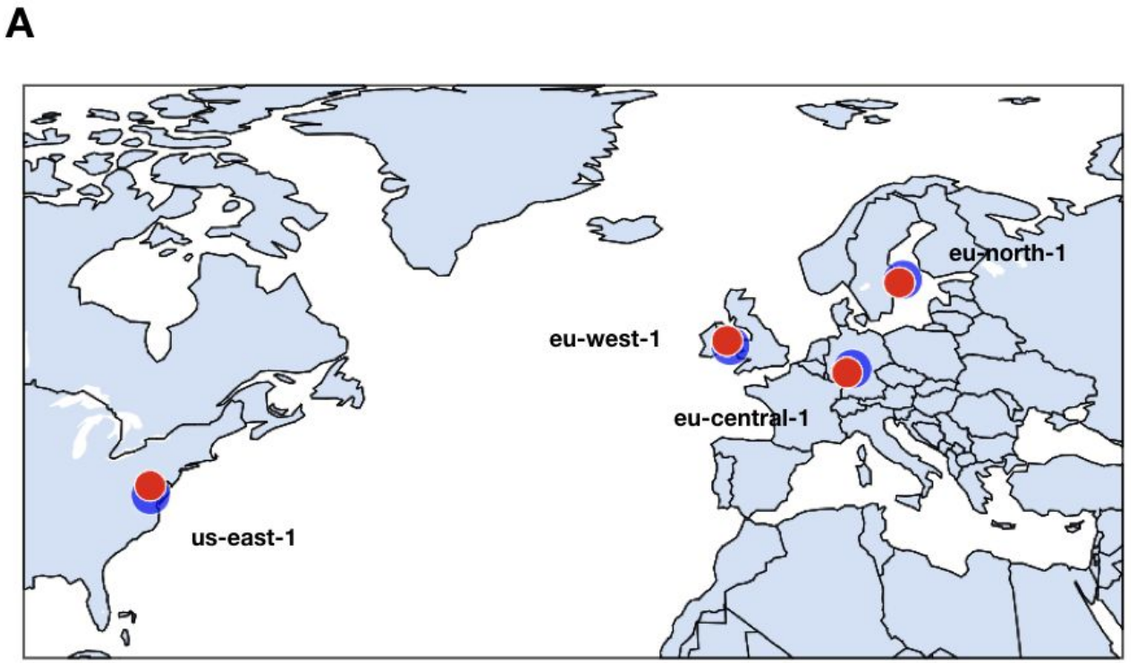
- Combiner - responsible of load balancing and partial model aggregation
- Reducer - responsible for combiner connections and global model generation

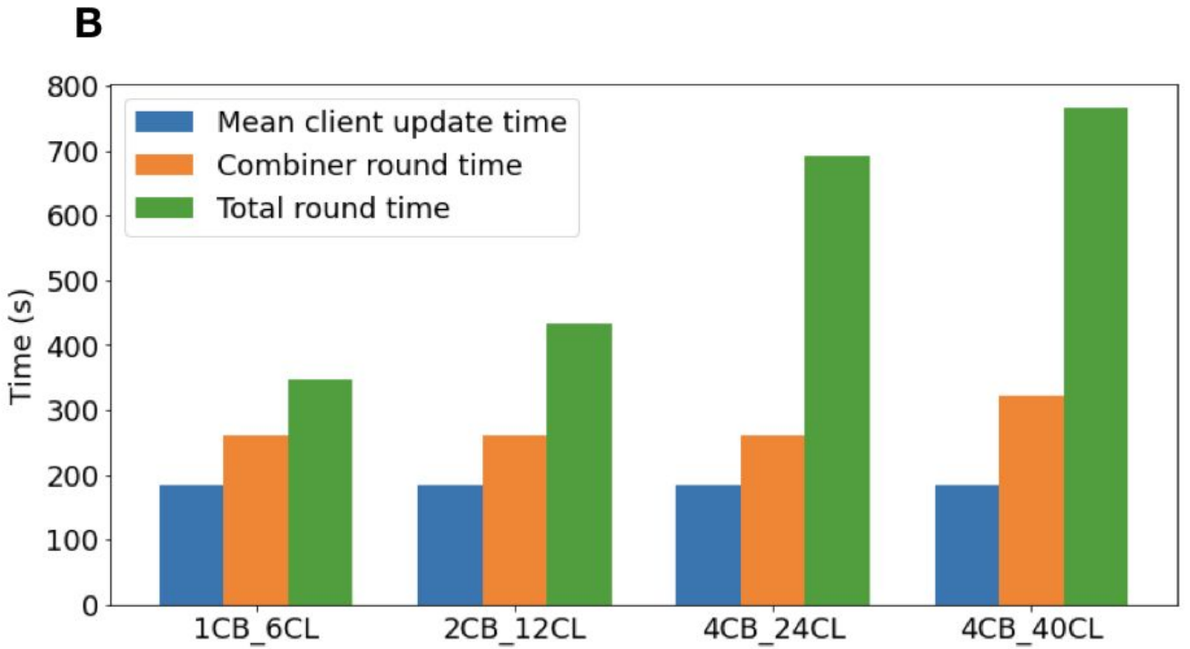
Tier 3 - Control management

- Monitoring
- Service discovery
- Model checkpointing



Cross-silo benchmark in AWS

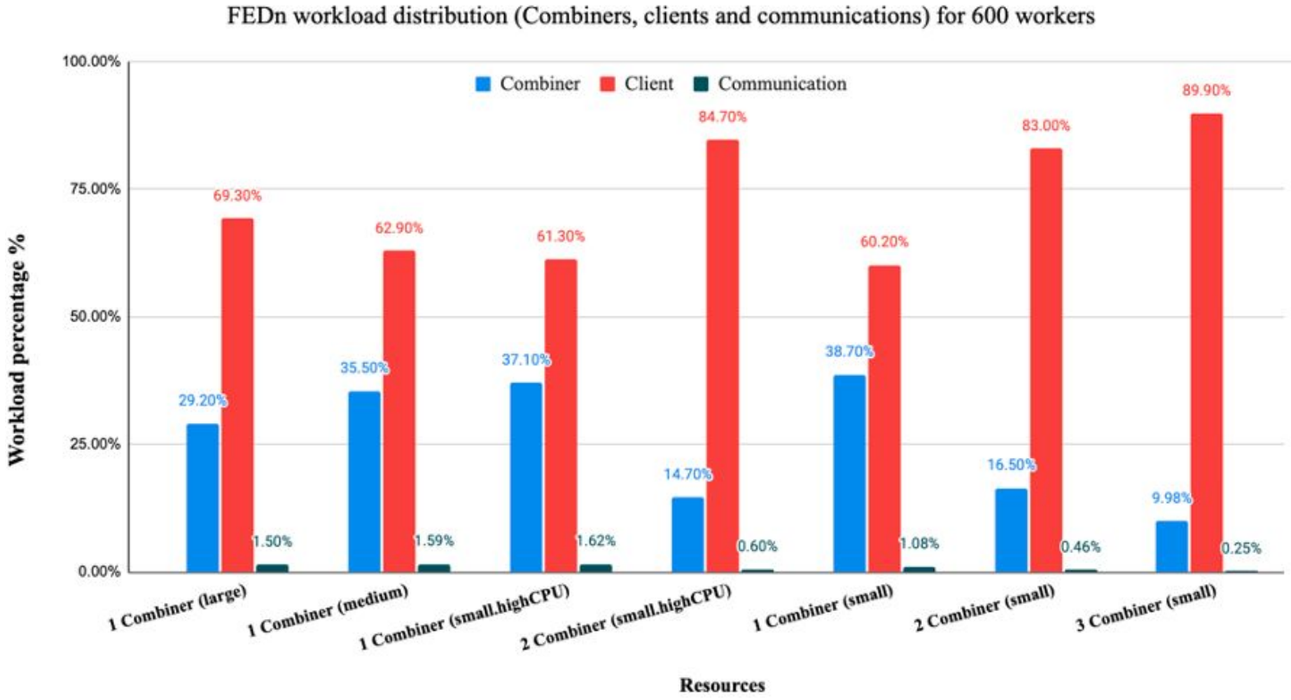


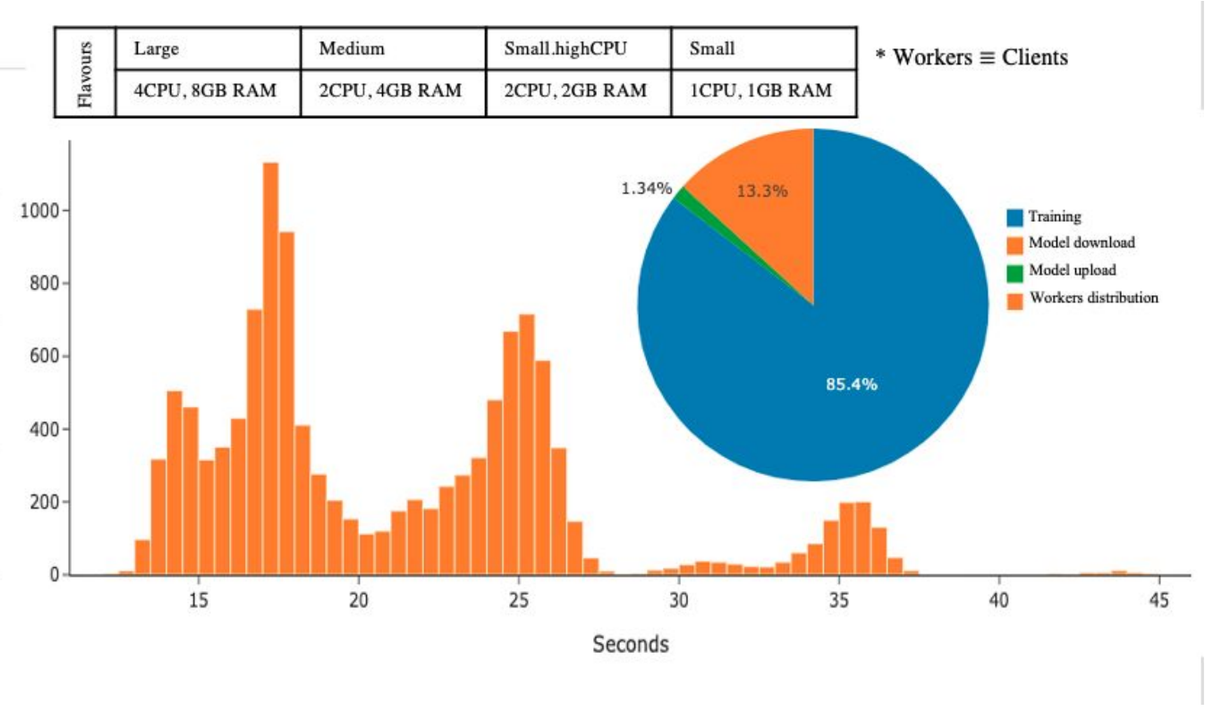


40 clients spread over 4 regions in EC2, each updating 1 GB TF models. 4 Combiners -> data flow in network ~ 450MB/s during communication stage (comms over internet)

Sentiment Analysis (IMDB):
https://github.com/scaleoutsystems/FEDn-client-imdb-keras

Resource constrained clients and combiners (IoT scenario)



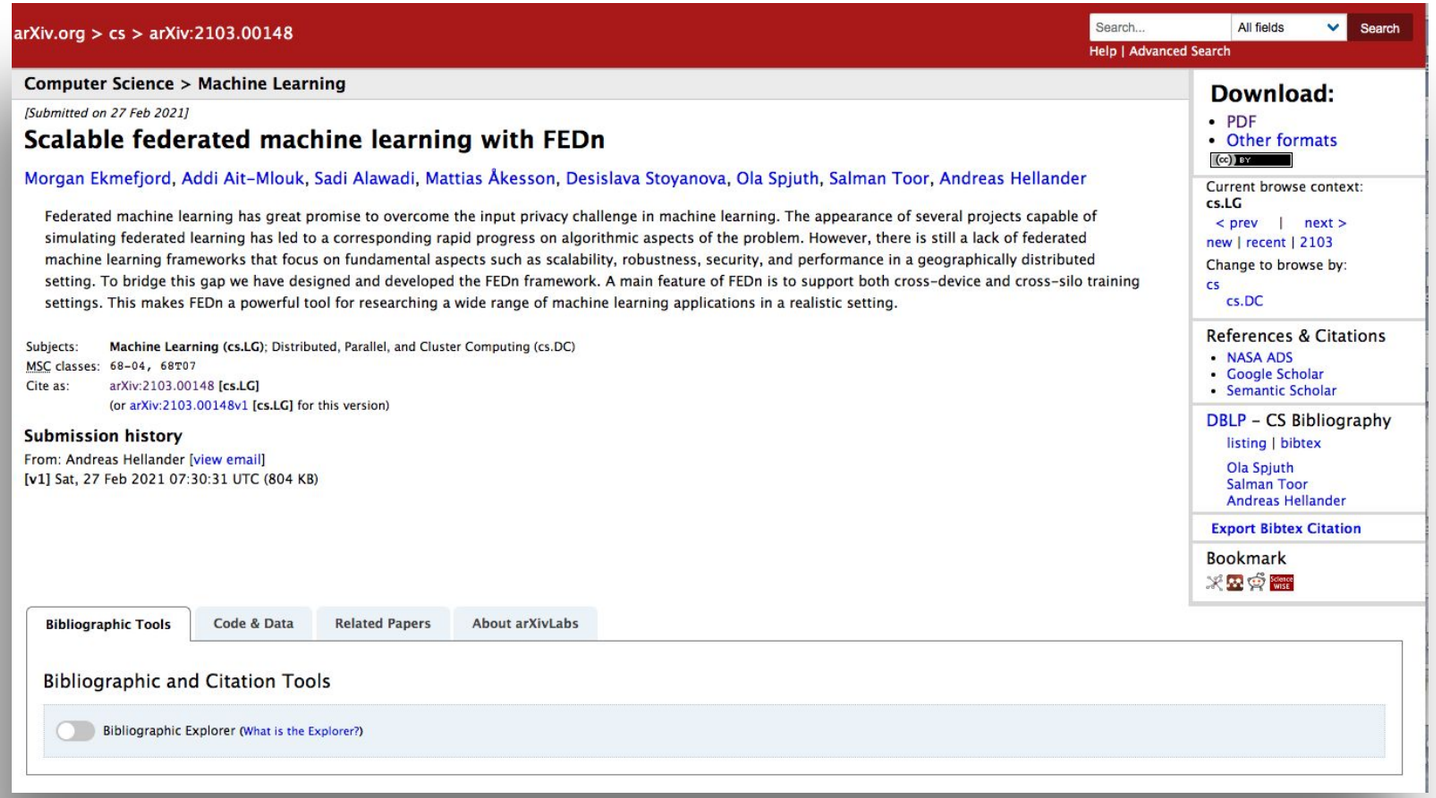


A single “cheap” combiner handles 1000+ clients with small models updates (~70kb).
Resilience increases with more combiners.

Human Daily Activity Recognition (cross-device, small model):
https://github.com/scaleoutsystems/FEDn-client-casa-keras

Scalable federated machine learning with ***FEDn***

https://arxiv.org/abs/2103.00148



arXiv.org > cs > arXiv:2103.00148

Search... All fields Search

Help | Advanced Search

Computer Science > Machine Learning

[Submitted on 27 Feb 2021]

Scalable federated machine learning with FEDn

Morgan Ekmefjord, Addi Ait-Mlouk, Sadi Alawadi, Mattias Åkesson, Desislava Stoyanova, Ola Spjuth, Salman Toor, Andreas Hellander

Federated machine learning has great promise to overcome the input privacy challenge in machine learning. The appearance of several projects capable of simulating federated learning has led to a corresponding rapid progress on algorithmic aspects of the problem. However, there is still a lack of federated machine learning frameworks that focus on fundamental aspects such as scalability, robustness, security, and performance in a geographically distributed setting. To bridge this gap we have designed and developed the FEDn framework. A main feature of FEDn is to support both cross-device and cross-silo training settings. This makes FEDn a powerful tool for researching a wide range of machine learning applications in a realistic setting.

Subjects: **Machine Learning (cs.LG)**; Distributed, Parallel, and Cluster Computing (cs.DC)
MSC classes: 68-04, 68T07
Cite as: arXiv:2103.00148 **[cs.LG]**
(or arXiv:2103.00148v1 **[cs.LG]** for this version)

Submission history
From: Andreas Hellander [view email]
[v1] Sat, 27 Feb 2021 07:30:31 UTC (804 KB)

Download:
- PDF
- Other formats

Current browse context: **cs.LG**
< prev | next >
new | recent | 2103
Change to browse by:
cs
cs.DC

References & Citations
- NASA ADS
- Google Scholar
- Semantic Scholar

DBLP - CS Bibliography
listing | bibtex
Ola Spjuth
Salman Toor
Andreas Hellander

Export Bibtex Citation

Bookmark

Bibliographic Tools | Code & Data | Related Papers | About arXivLabs

Bibliographic and Citation Tools

Bibliographic Explorer (What is the Explorer?)

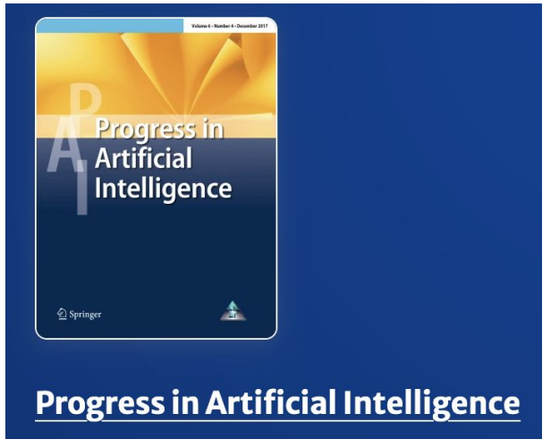
Published in: 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)

Date of Conference: 16-19 May 2022 **INSPEC Accession Number:** 21883891

Date Added to IEEE *Xplore*: 19 July 2022 **DOI:** 10.1109/CCGrid54584.2022.00065

Toward efficient resource utilization at edge nodes in federated learning

https://arxiv.org/abs/2309.10367



Prog Artif Intell 13, 101–117 (2024). https://doi.org/10.1007/s13748-024-00322-3

arXiv > cs > arXiv:2309.10367

Search... Help | Adva

Computer Science > Machine Learning

[Submitted on 19 Sep 2023 (v1), last revised 11 Jun 2024 (this version, v2)]

Toward efficient resource utilization at edge nodes in federated learning

C[x] CODE | Notebook

Sadi Alawadi, Addi Ait-Mlouk, Salman Toor, Andreas Hellander

Ask the author(s) a question! :) | Ask

powered by CatalyzeX

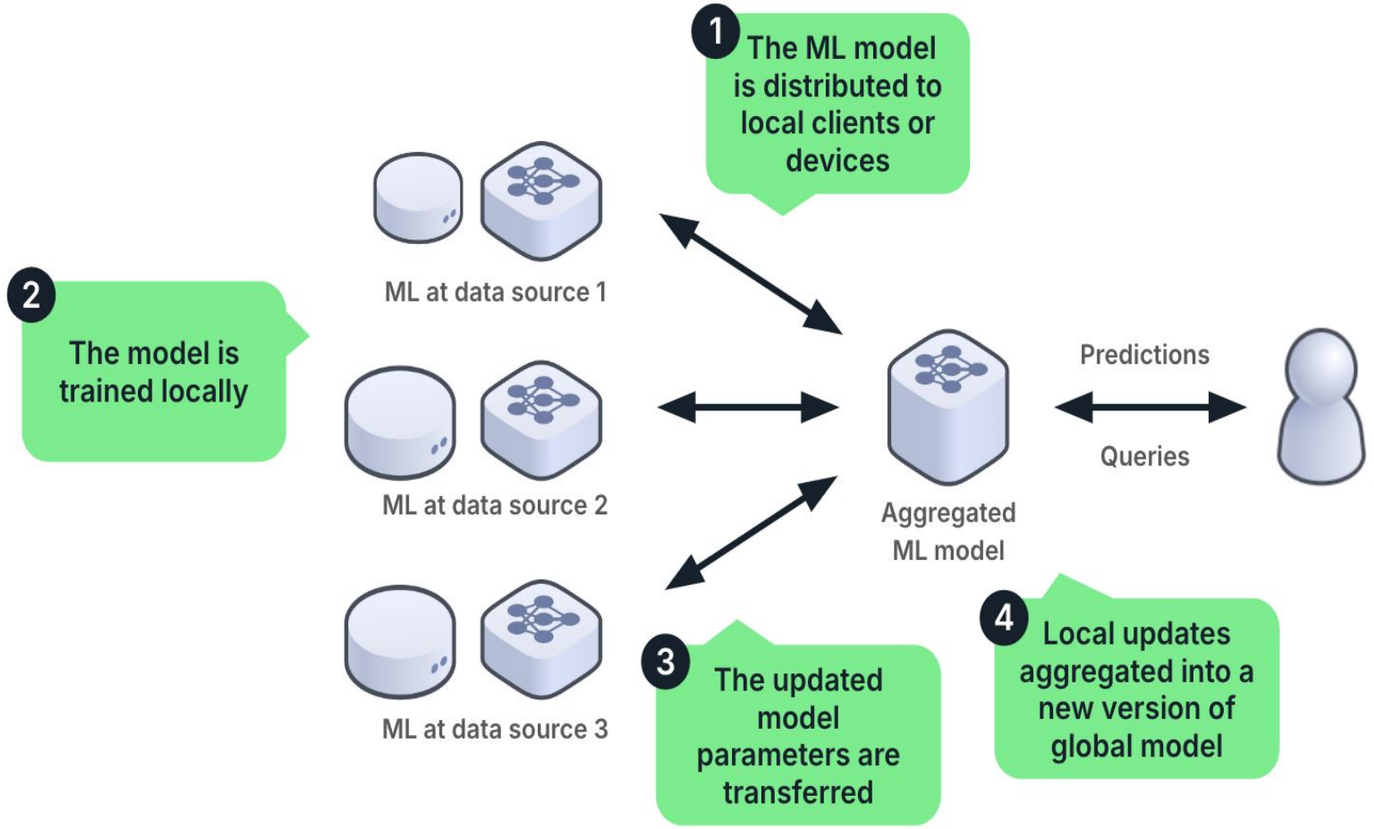
Federated learning (FL) enables edge nodes to collaboratively contribute to constructing a global model without sharing their data. This is accomplished by devices computing local, private model updates that are then aggregated by a server. However, computational resource constraints and network communication can become a severe bottleneck for larger model sizes typical for deep learning applications. Edge nodes tend to have limited hardware resources (RAM, CPU), and the network bandwidth and reliability at the edge is a concern for scaling federated fleet applications. In this paper, we propose and evaluate a FL strategy inspired by transfer learning in order to reduce resource utilization on devices, as well as the load on the server and network in each global training round. For each local model update, we randomly select layers to train, freezing the remaining part of the model. In doing so, we can reduce both server load and communication costs per round by excluding all untrained layer weights from being transferred to the server. The goal of this study is to empirically explore the potential trade-off between resource utilization on devices and global model convergence under the proposed strategy. We implement the approach using the federated learning framework FEDn. A number of experiments were carried out over different datasets (CIFAR-10, CASA, and IMDB), performing different tasks using different deep-learning model architectures. Our results show that training the model partially can accelerate the training process, efficiently utilizes resources on-device, and reduce the data transmission by around 75% and 53% when we train 25%, and 50% of the model layers, respectively, without harming the resulting global model accuracy.

Comments: 16 pages, 5 tables, 8 figures
Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI)
Cite as: arXiv:2309.10367 **[cs.LG]**
(or arXiv:2309.10367v2 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2309.10367
Journal reference: 10 June 2024
Related DOI: https://doi.org/10.1007/s13748-024-00322-3

Federated learning

scaleout

Bring the ML to the data instead of data to the ML



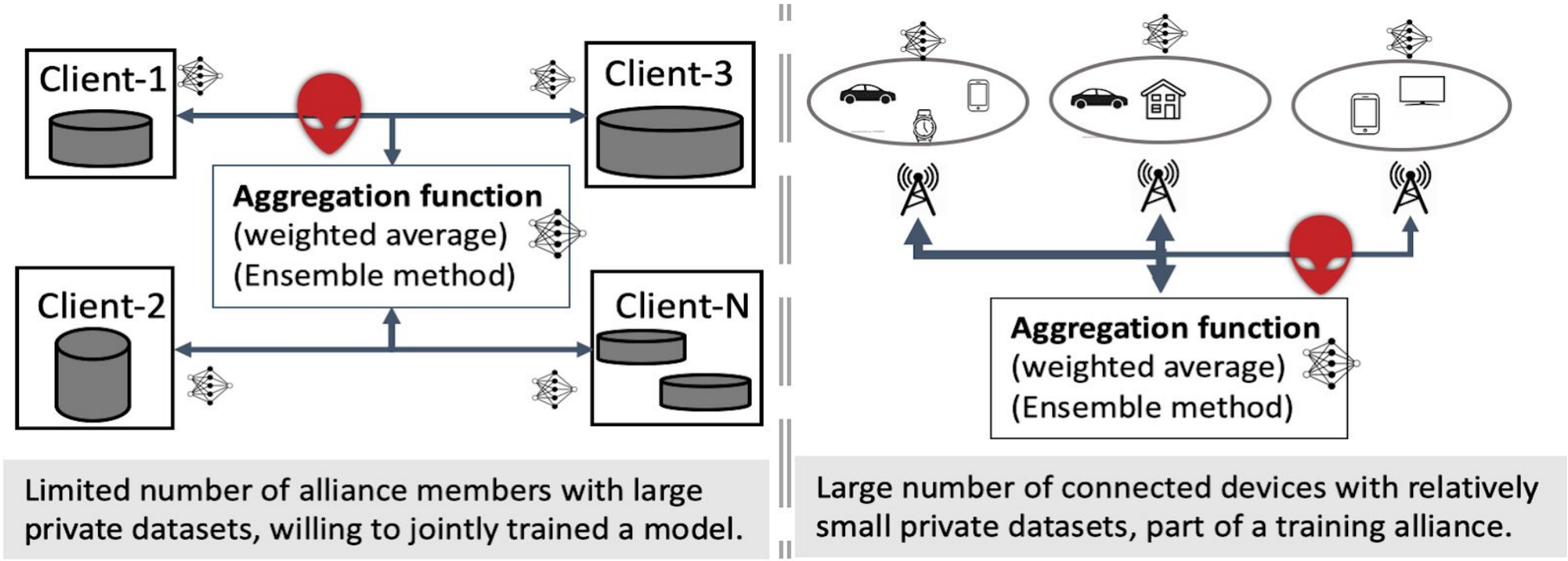
- Model size - 10MB
- Data distribution
  - Client 1 - 100GB
  - Client 2 - 50GB
  - Client 3 - 50GB
- Total data size - 200GB

Communication overhead

- Round 1
  - Step 1 - 10MB * 3 = ~30MB
  - Step 2 - No communication needed
  - Step 3 - 3 * 10MB = ~30MB
  - Step 4 - No communication needed
  - **Total data transfer, ~60MB**

100 Rounds will require - ~60MB * 100 = ~6000MB ~ 6GB
3% of the total data (200GB)

Adversarial attacks and FedML



- Data Poisoning
  - manipulating the training data
- Model Poisoning
  - manipulating the local model

Examples:

- Label flipping attacks
- Gradient attacks
- Backdoor Attacks
- ...

Overestimated Challenge in Federated Settings

Many studies have made unrealistic assumptions regarding the prevalence of compromised clients. Even 20% of compromised clients within a federation is highly unlikely in production grade FL environments.

V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning

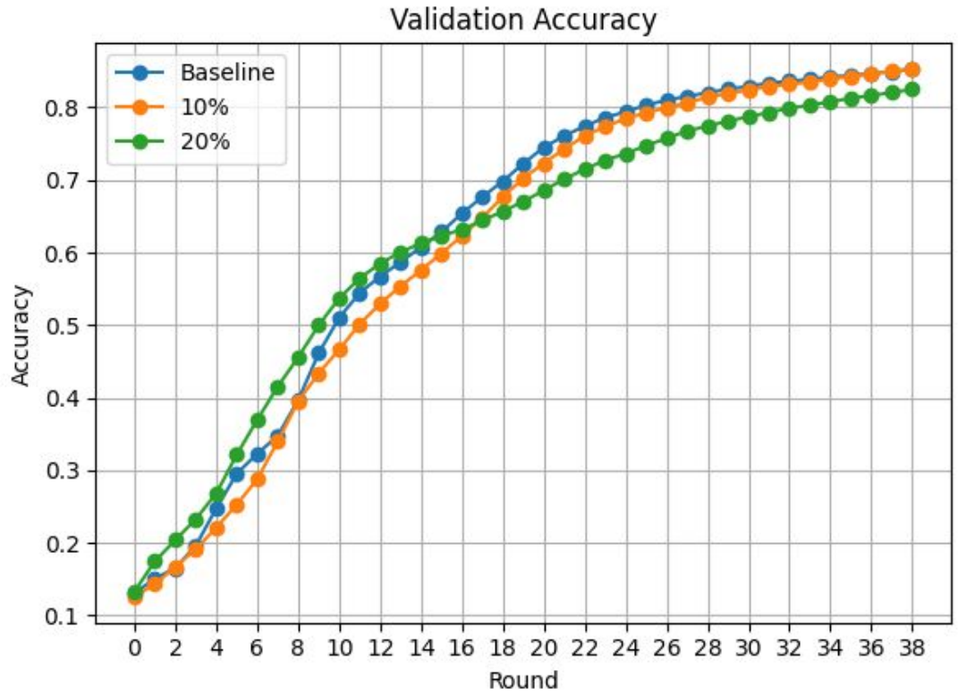
Label flipping attack

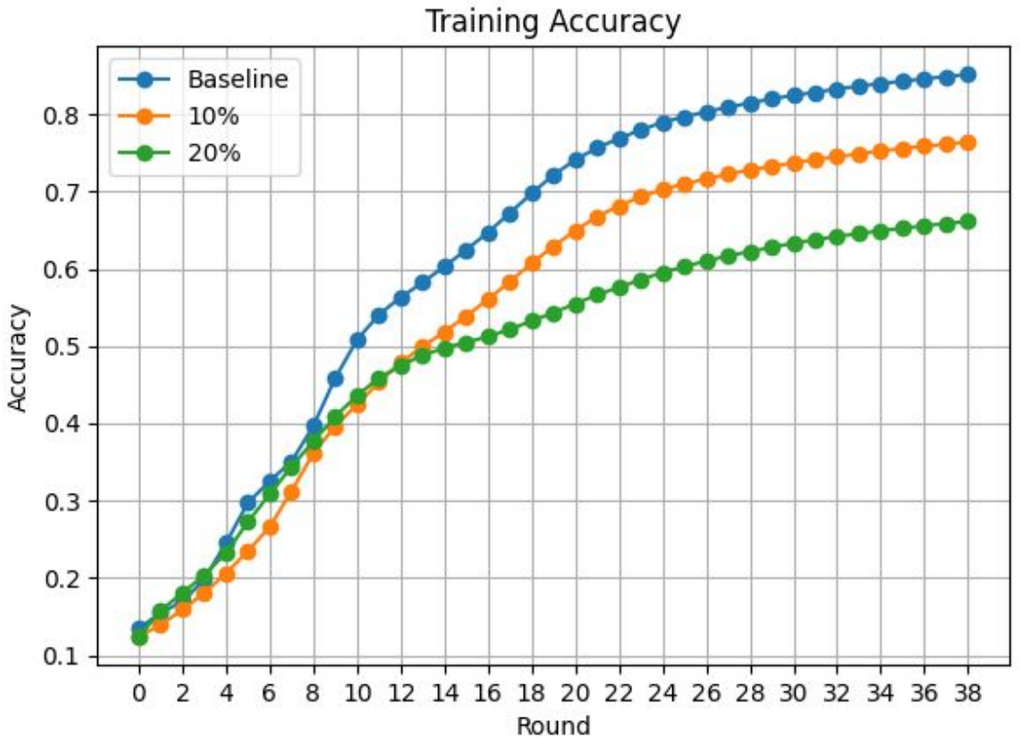
A Label Flipping Attack is a type of data poisoning attack where the training data in one or many clients is manipulated.

Malicious clients: 10% and 20% clients

Data distribution settings: Balanced + IID

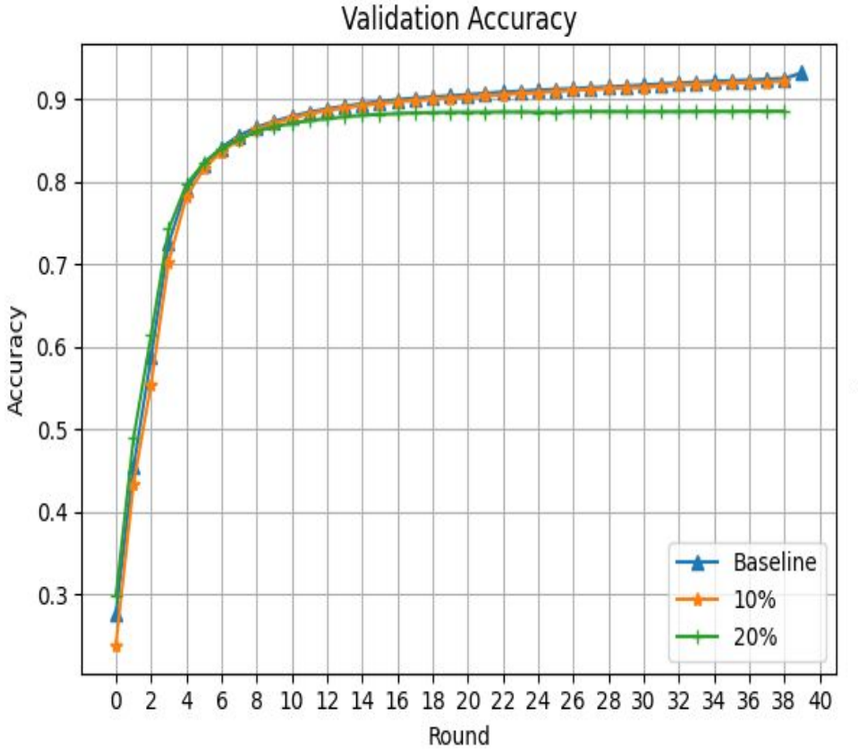
Use case: MNIST Handwritten digit dataset

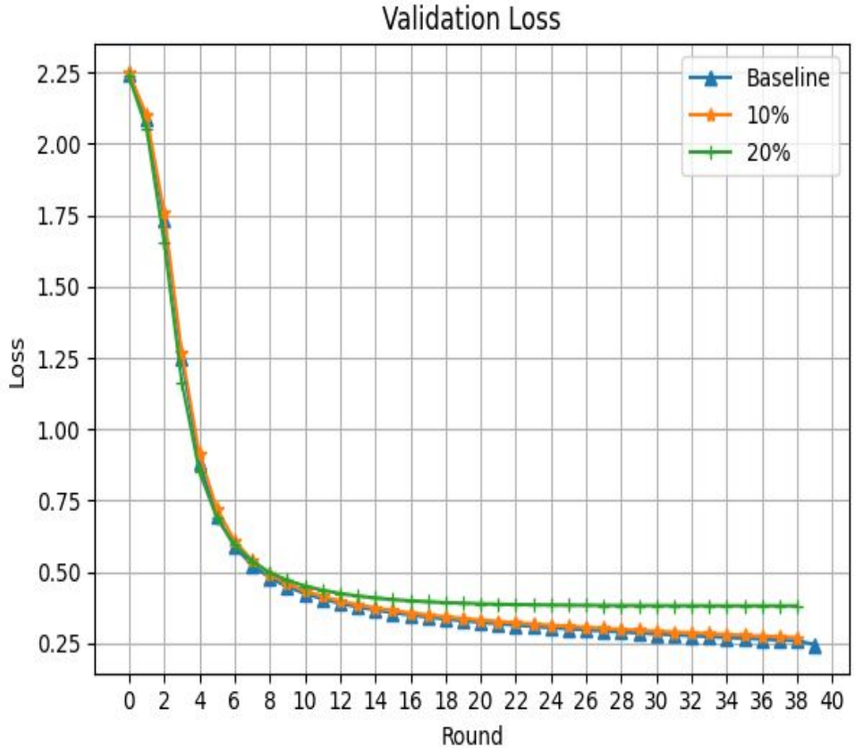


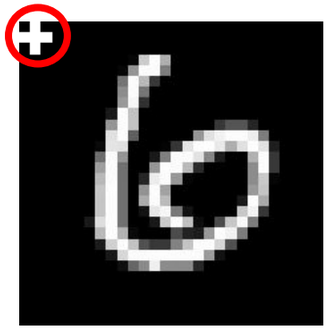


Backdoor attack

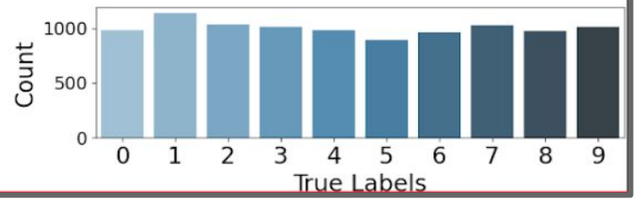
A backdoor attack in machine learning involves maliciously inserting subtle patterns during training to manipulate a model's behavior, causing it to make incorrect predictions when triggered by specific inputs during deployment.







Backdoor attack

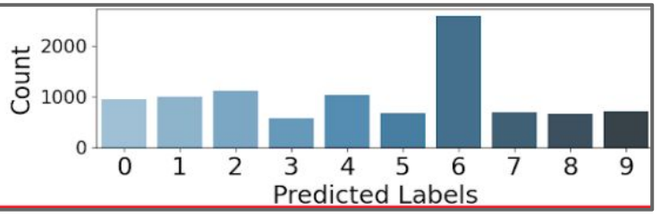
Description	Sample Image	Label Distribution
True Label Distribution		

Malicious clients: 20% clients

Data distribution settings: Balanced + IID

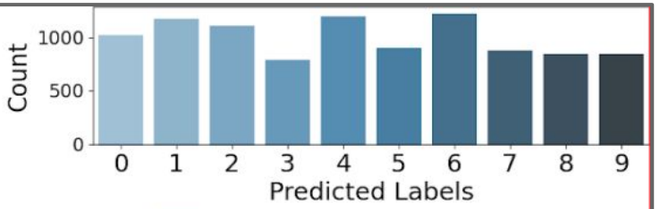
Use Case: MNIST Handwritten digit dataset

Model predictions on dataset 2

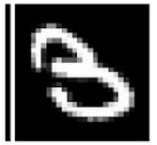


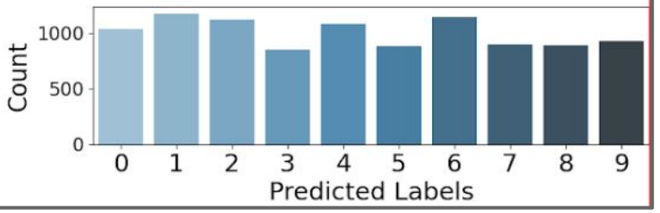
Dataset 2
Test Data with Backdoor

Model predictions on dataset 5

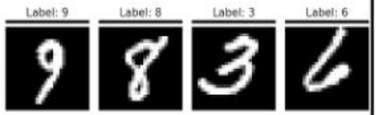


Model predictions on dataset 6





Dataset 5
Test Data with Horizontal Line



Dataset 6
Test Data with Vertical Line