UNIVERSITY OF AMSTERDAM

Policy-aware distributed Vertical Federated Learning infrastructure

Jake Jongejans, **Alexandros Koufakis**, Ana Oprescu
University of Amsterdam

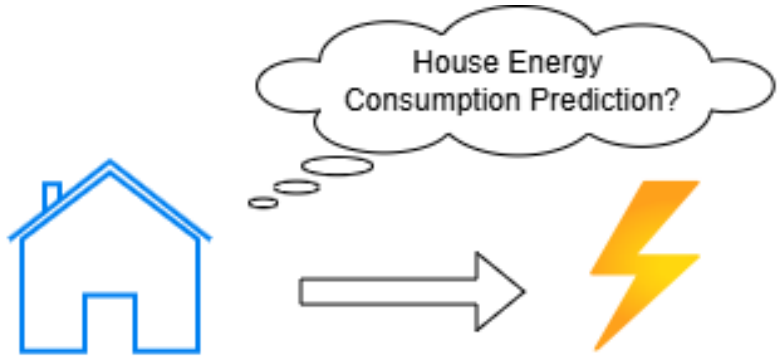
CIENA BOOTH: #3330

ciena

prets embedded lending

Use case: Federated building energy consumption

- House energy consumption
- Benefits:
  - Energy
- Different agents
  - Utility company (past energy consumption)
  - Municipal archive (registered people)
  - Building registry (Solar panels, insulation, …)
- Private datasets



Vertical Federated Learning (VFL)

- FL is a machine learning approach where multiple participants collaboratively train a model without sharing their raw data
- Data stays local (on-premise)
- Only model updates (weights, gradients) are shared
- Ensures data privacy and regulatory compliance

- **Vertical**: Participants have different features (data columns) about the same entities.

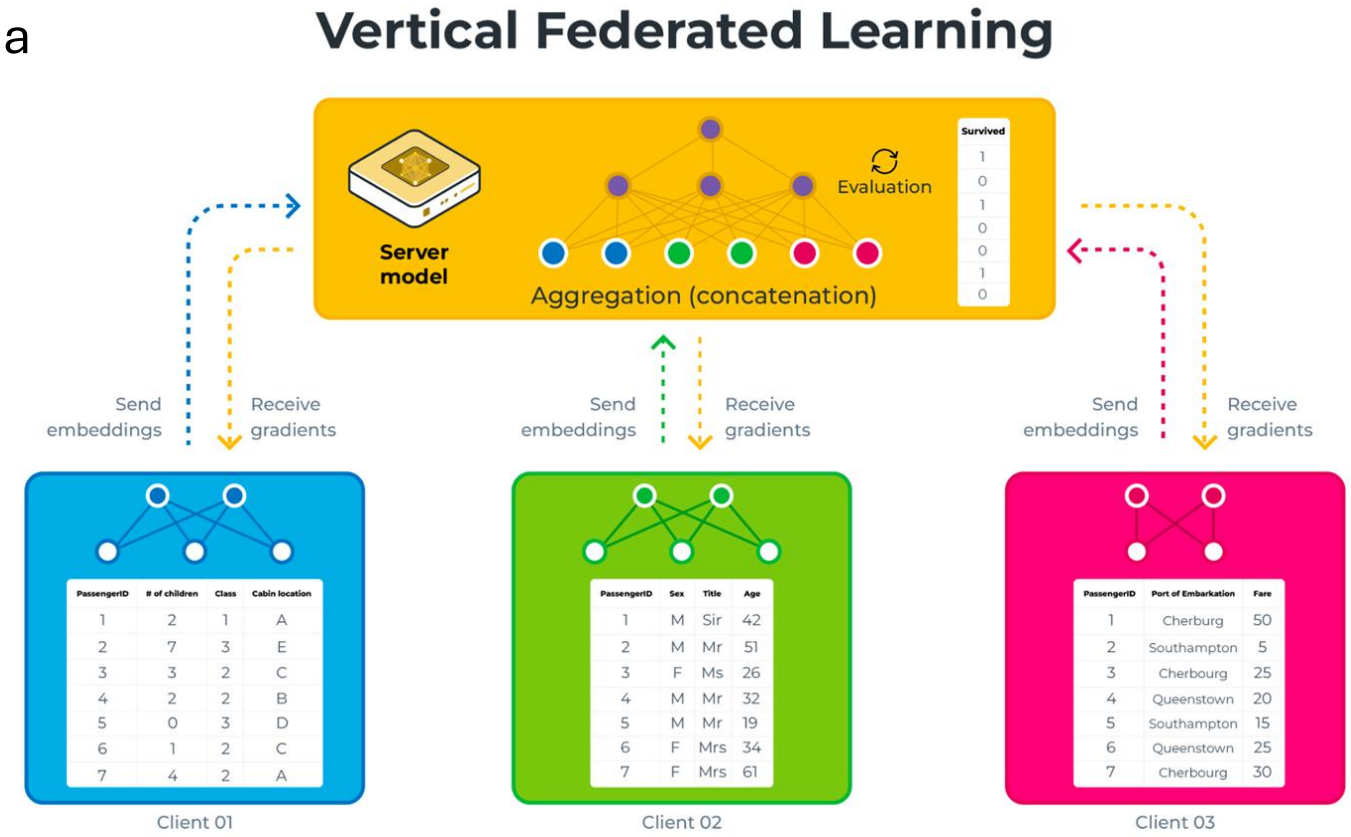


Image: https://flower.ai/docs/examples/vertical-fl.html

Dynamic Policies and Disagreement resolution

- Participating clients must consent
- Policy re-evaluation in between cycles
- Different types of exclusion
- Dynamically adjusting VFL

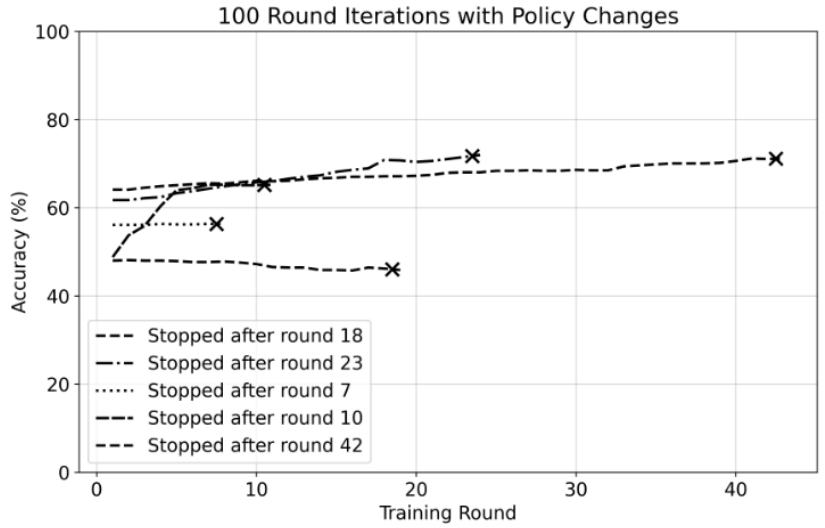


Image: https://scripties.uba.uva.nl/search?id=record_56506

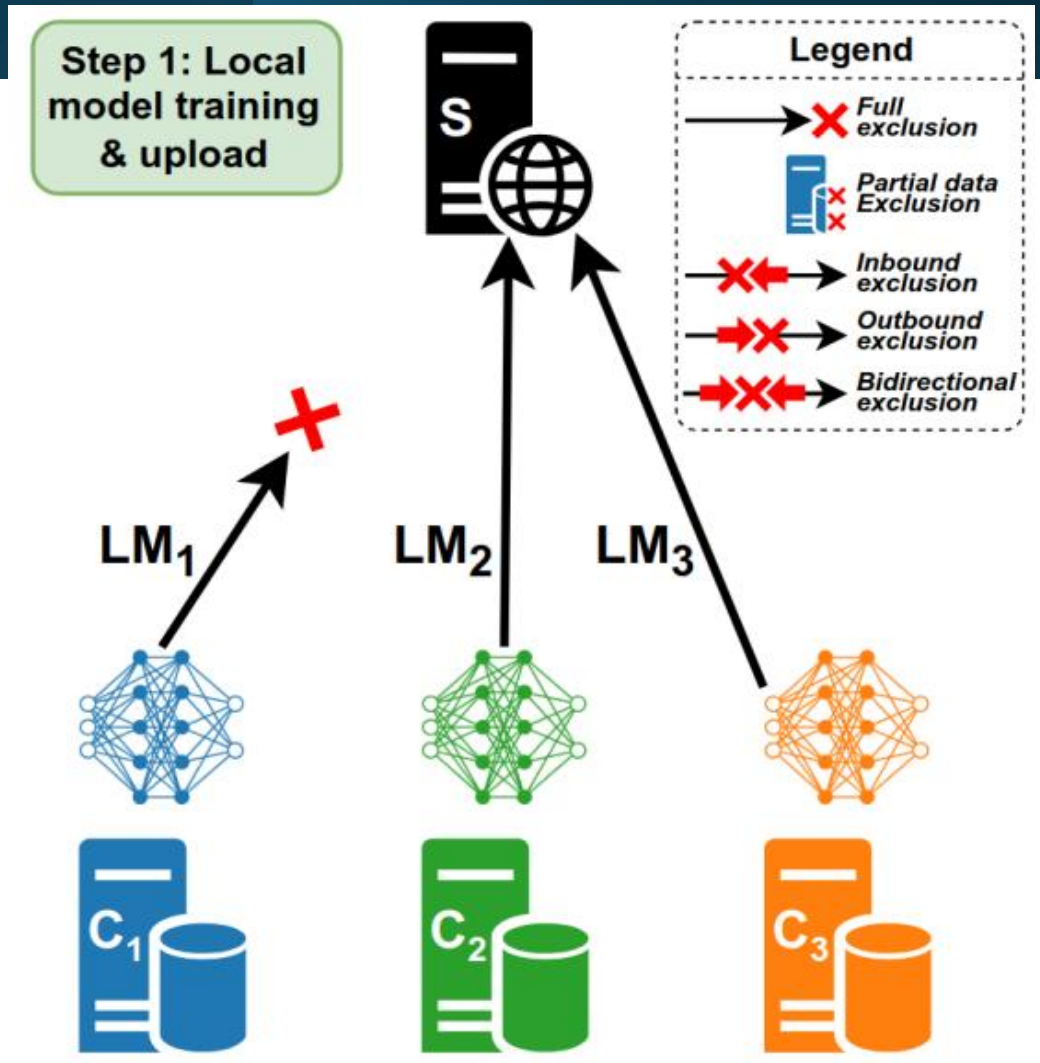


Image: https://scripties.uba.uva.nl/search?id=record_56491

FABRIC testbed

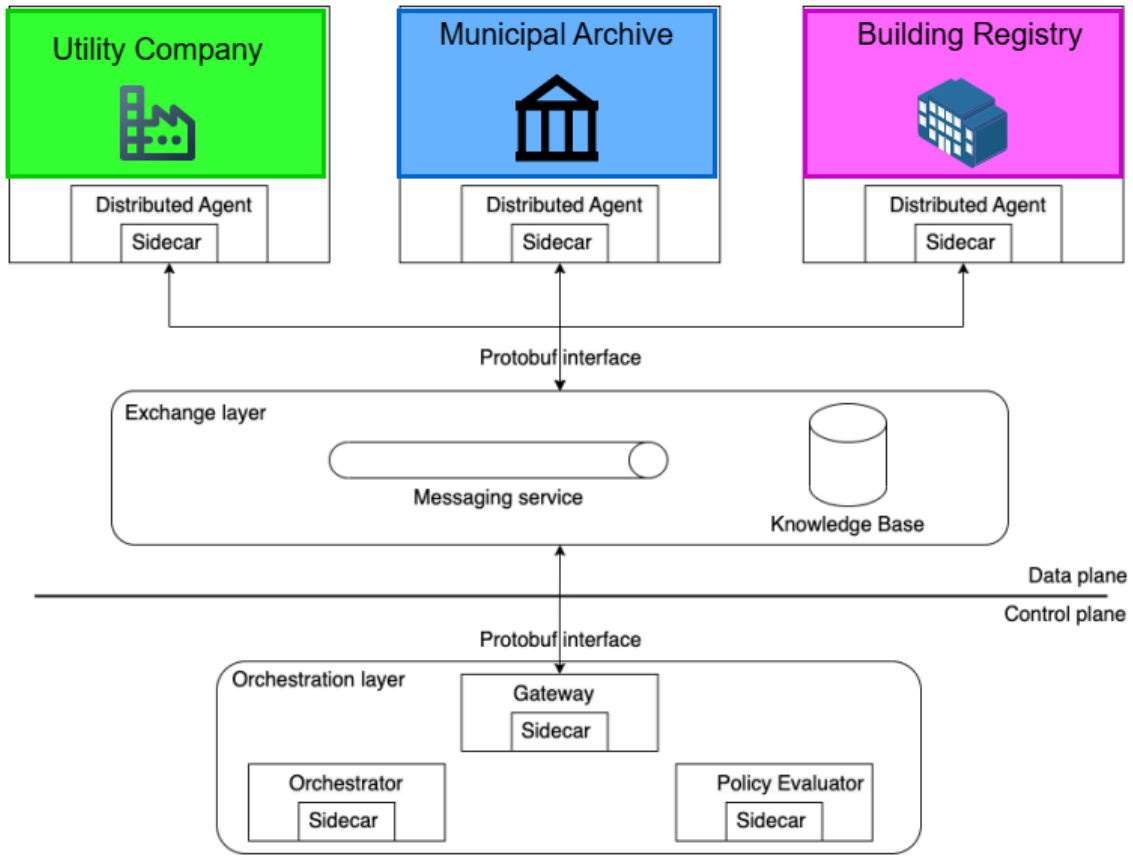
- **F**ABRIC is **A**daptive Programma**B**le **R**esearch **I**nfrastructure for **C**omputer Science and Science Applications Test
- Powerful hardware
- Distributed infrastructure
- https://portal.fabric-testbed.net/

FABRIC

DYNAMOS: middleware for data exchange systems

- DYNAMOS: **Dy**namically **A**daptive **M**icroservice-based **OS**
- **Policy evaluator:** validates whether a request may be executed according to a set of regulations and agreements.
- **Distributed agents:** a party within the data exchange that owns a collection of data, and is willing to exchange its data according to a set of policies.

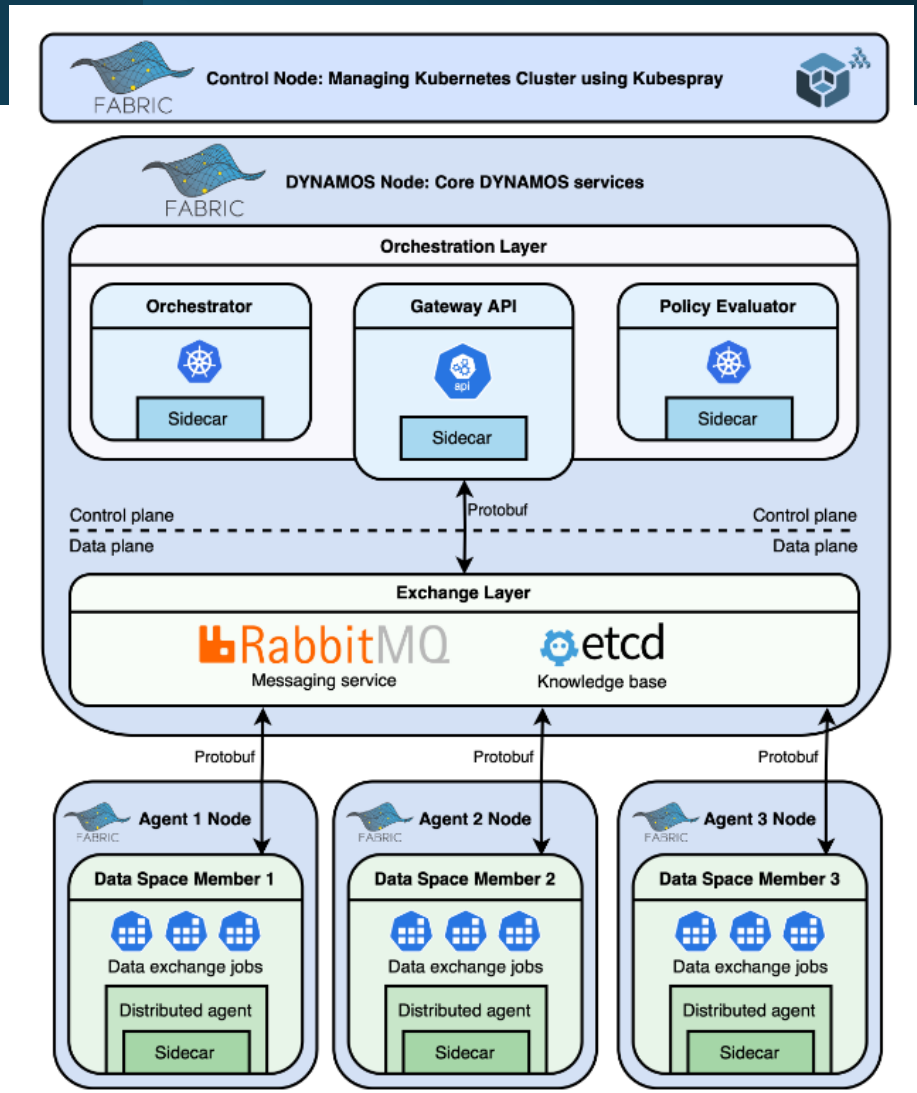
https://github.com/DYNAMOS-UVA/DYNAMOS



Scattered Directive: Distributed and Automated Infrastructure

- **Scattered Directive:** Framework for deployment and automation based on DYNAMOS
- Distributed **Kubernetes** setup
  - **Kubespray:** cluster management tool
  - **Flannel:** container network interface plugin
- FABRIC nodes:
  - Kubernetes control node
  - DYNAMOS core node
  - Distributed agents

https://github.com/DYNAMOS-UVA/Scattered-Directive/



Policy Aware VFL demonstration on FABRIC

Dataset: USA Residential Building Energy Consumption Survey

Conducted by EIA, includes **5,600+** households (2015 edition)

- **Target variable:** Energy consumption (BTU)
- **Features:**
  - Housing characteristics, usage patterns, and demographics.
  - Approximately: 300 features
- https://www.kaggle.com/datasets/claytonmiller/2015-residential-energy-consumption-survey

kaggle

Image: https://www.flickr.com/photos/caribb/1518299093

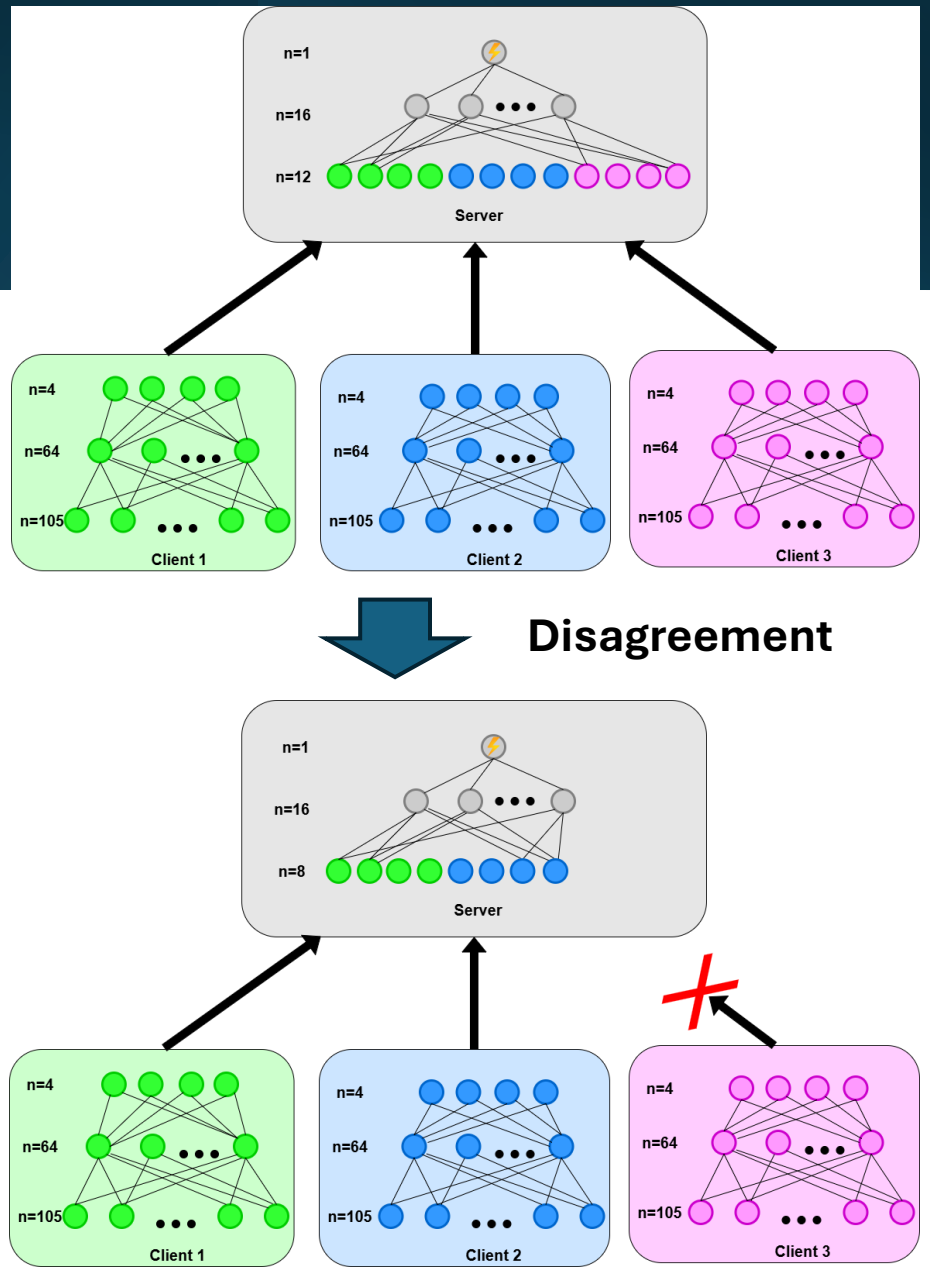
VFL network architecture

Clients:

- 105 features
- Intermediate layer 64 neurons
- Fully connected

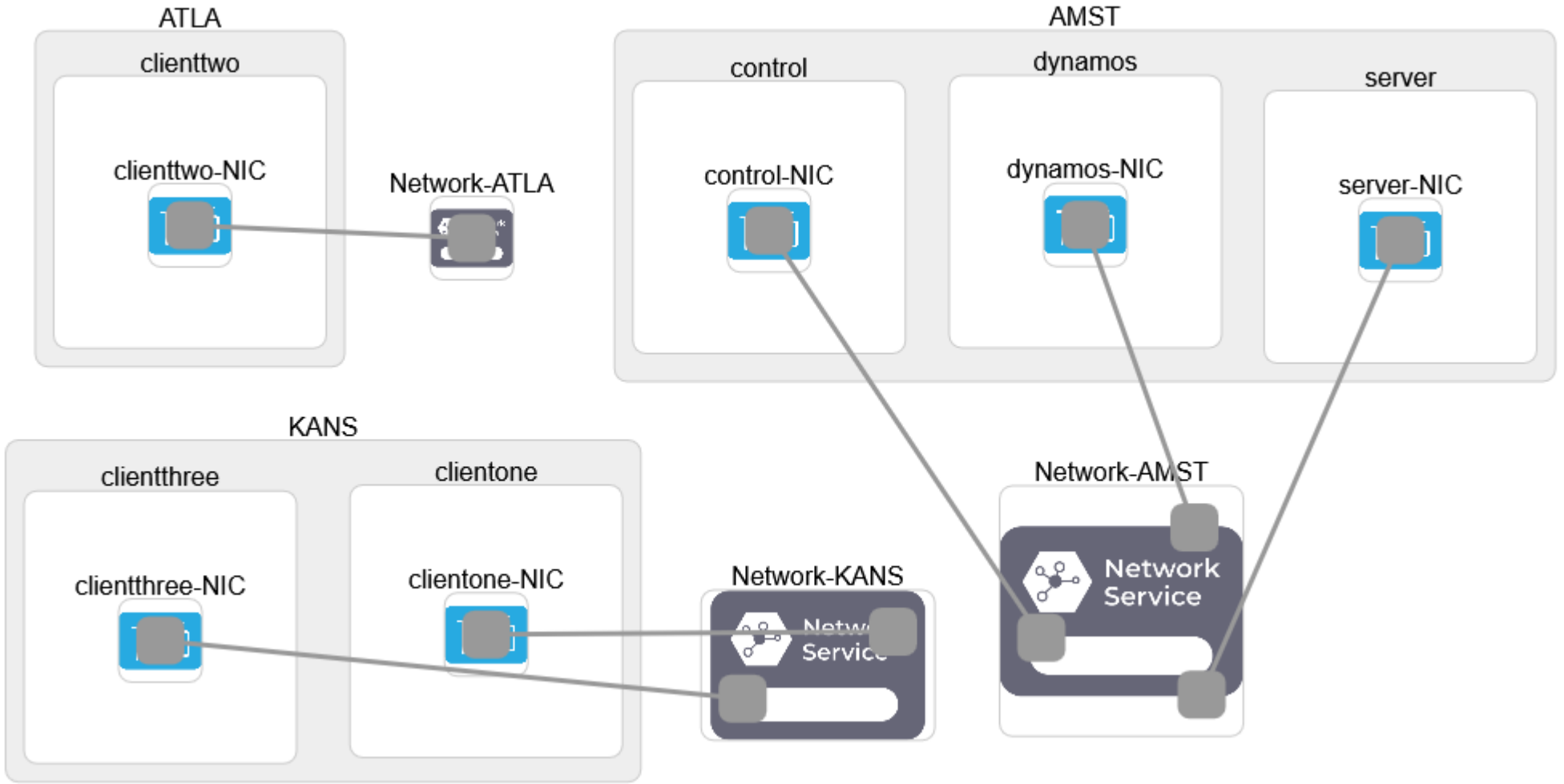
Server:

- 12 or 8 input
- Intermediate layer of 16 neurons
- Output: Building energy consumption
- Fully connected



FABRIC Node Topology

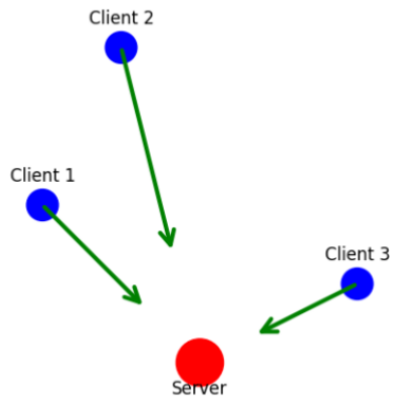
- AMST node:
  - Kubernetes control
  - Dynamos core
  - Agent: VFL server
- ATLA node:
  - Agent: Client 2
- KANS node:
  - Agent: Client 1
  - Agent: Client 3

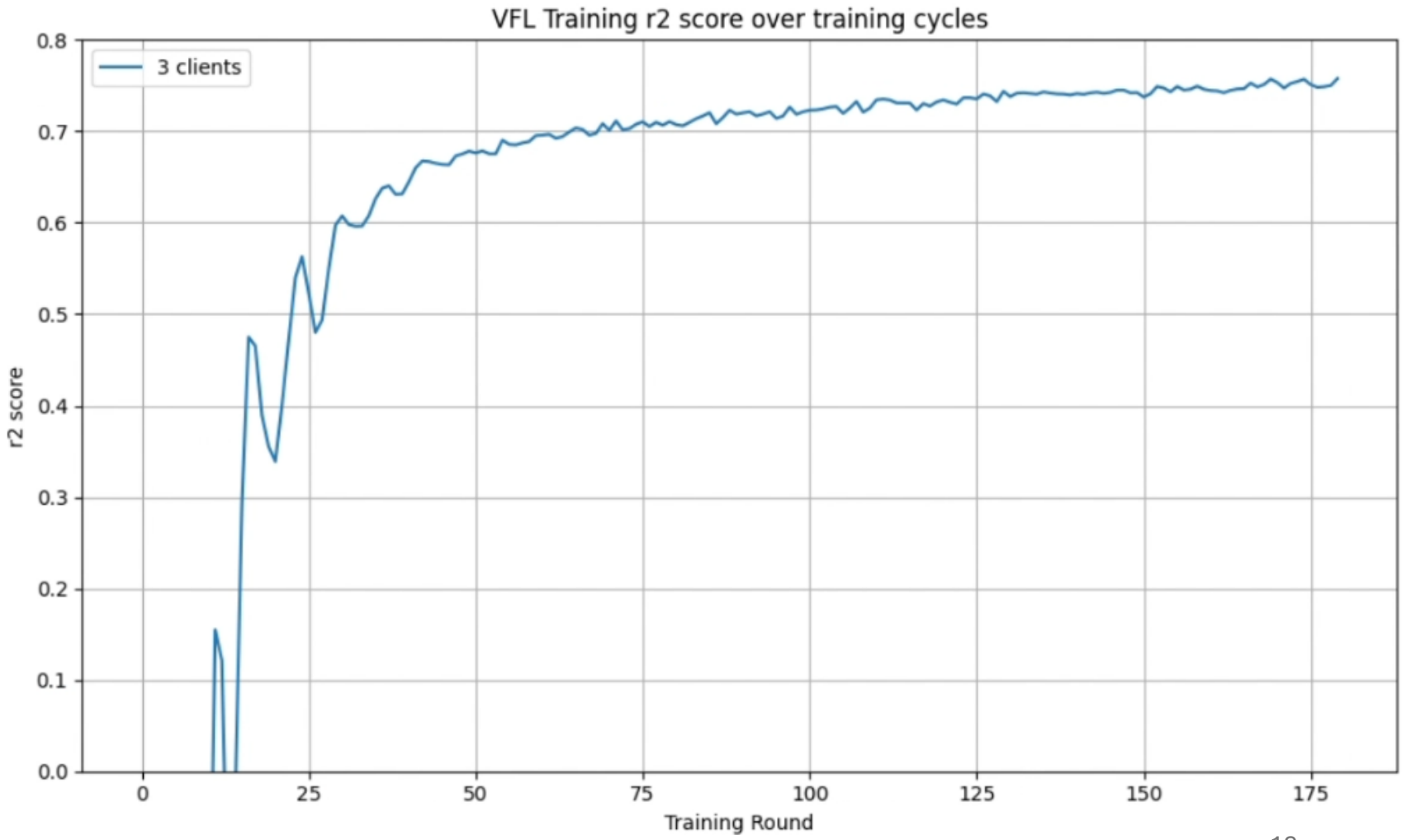


FABRIC

Scenario 1: VFL with three clients

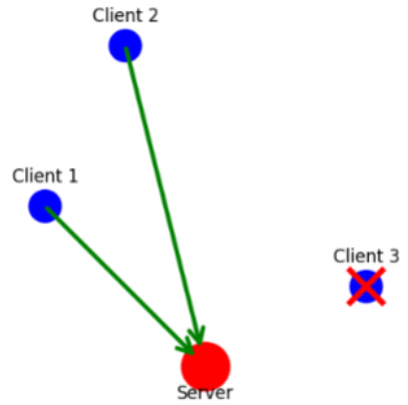
- **Three clients**
- Execution time: 79 s
- Training Rounds: 180
- Final r2 score: 0.76

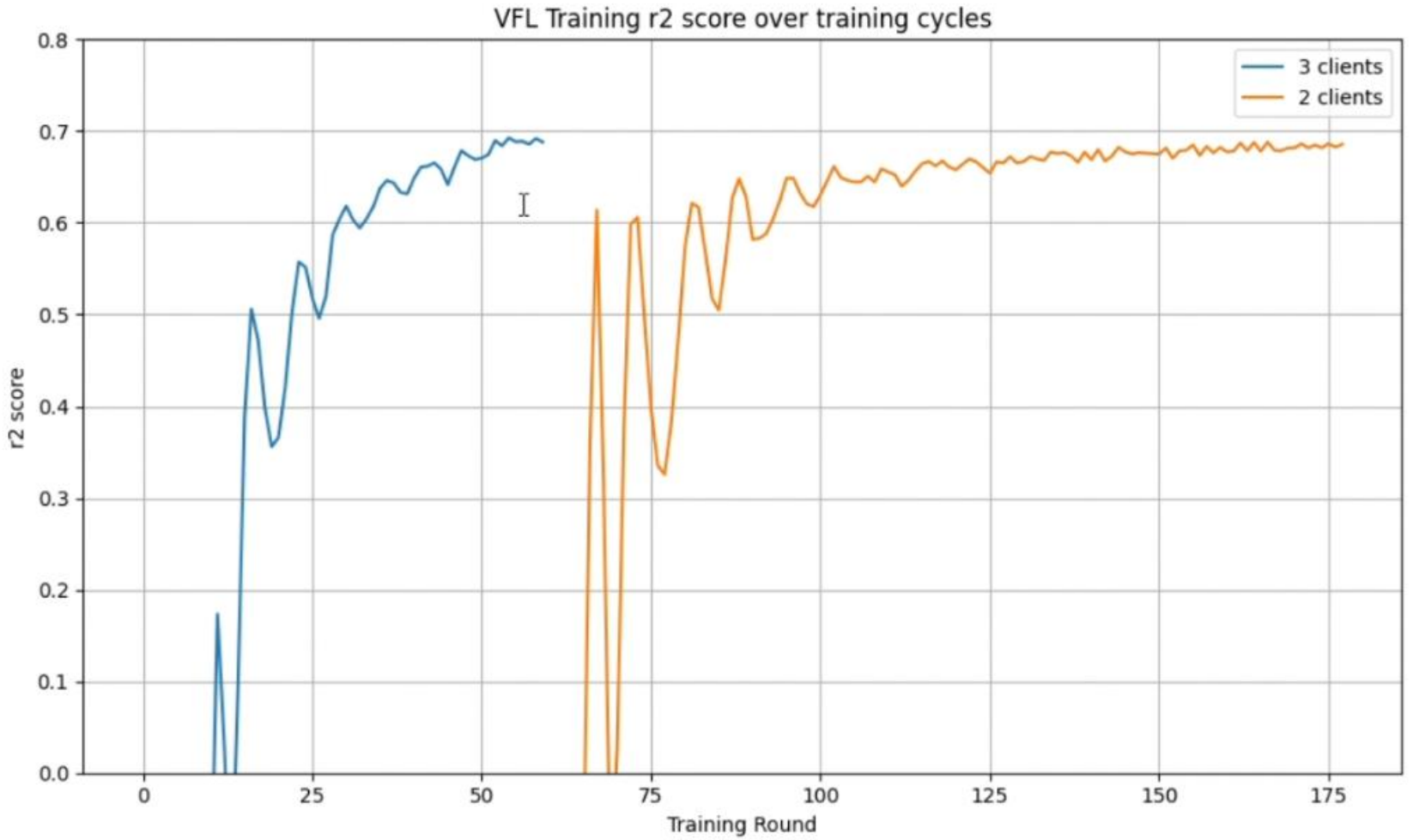




Scenario 2: Client removal

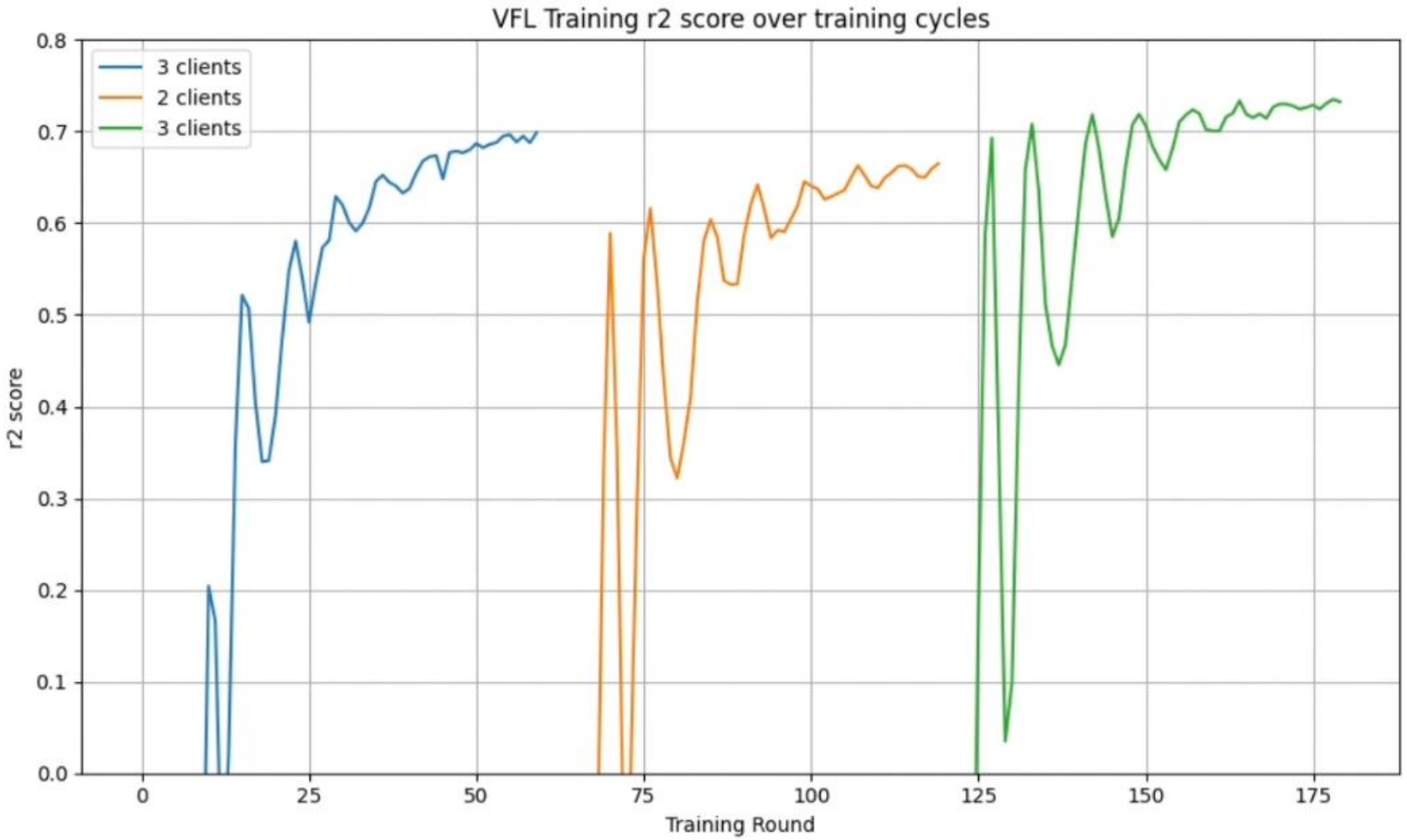
- **Client excluded on round 60**
- Execution time: 71 s
- Training Rounds: 180
- Round 60 r2 score: 0.67
- Final r2 score: 0.68





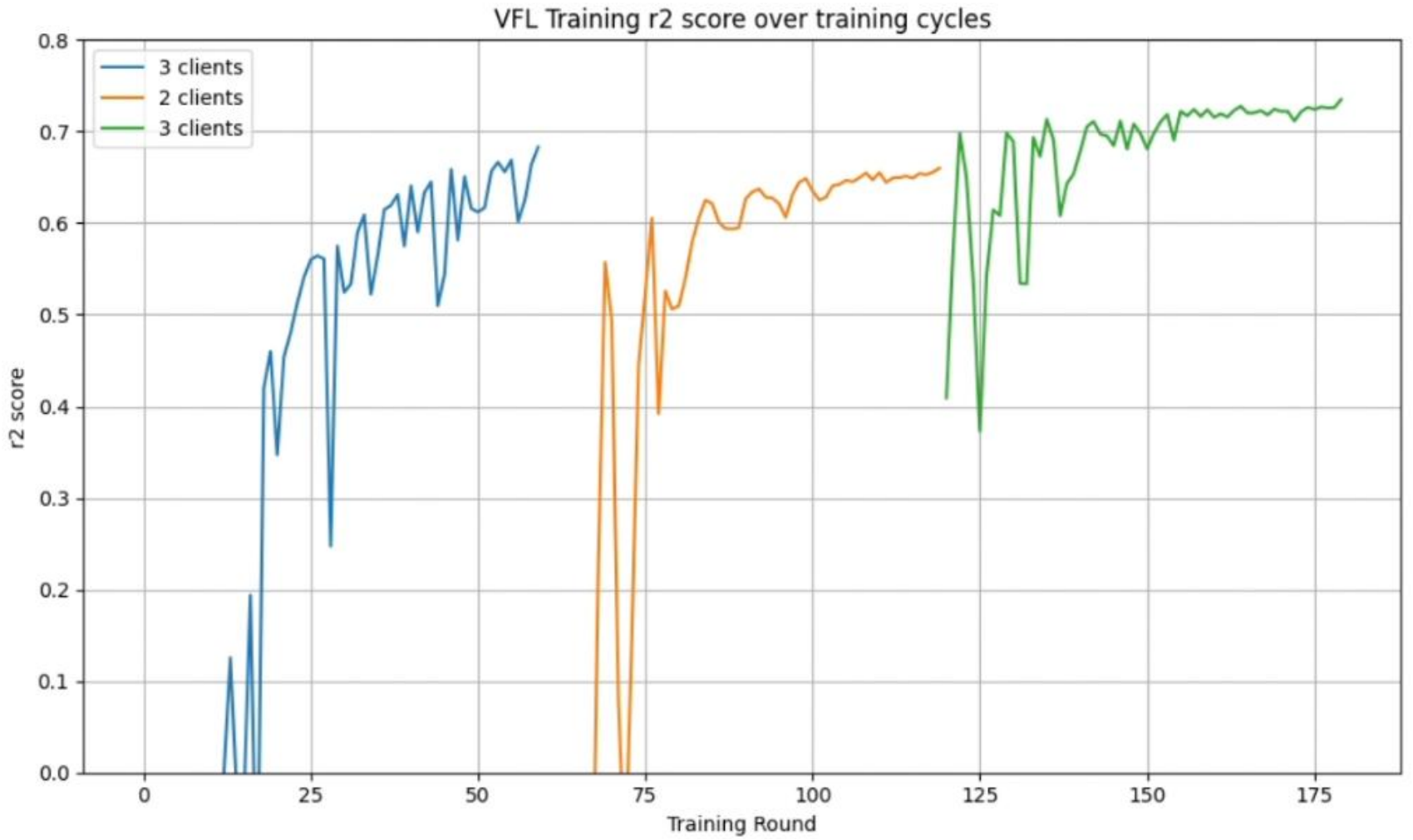
Scenario 3: Client Removal & Reintroduction

- **Client excluded on round 60**
- **Reintroduced on round 120**
- Execution time: 74 s
- Training Rounds: 180
- Round 60 r2 score: 0.70
- Round 120 r2 score: 0.66
- Final r2 score: 0.73



Scenario 4: Client Removal & Reintroduction (with backtrack)

- **Client excluded on round 60**
- **Reintroduced on round 120**
  - **Training continues from round 60 state**
- Execution time: 73 s
- Training Rounds: 180
- Round 60 r2 score: 0.68
- Round 120 r2 score: 0.66
- Final r2 score: 0.74



Related publication

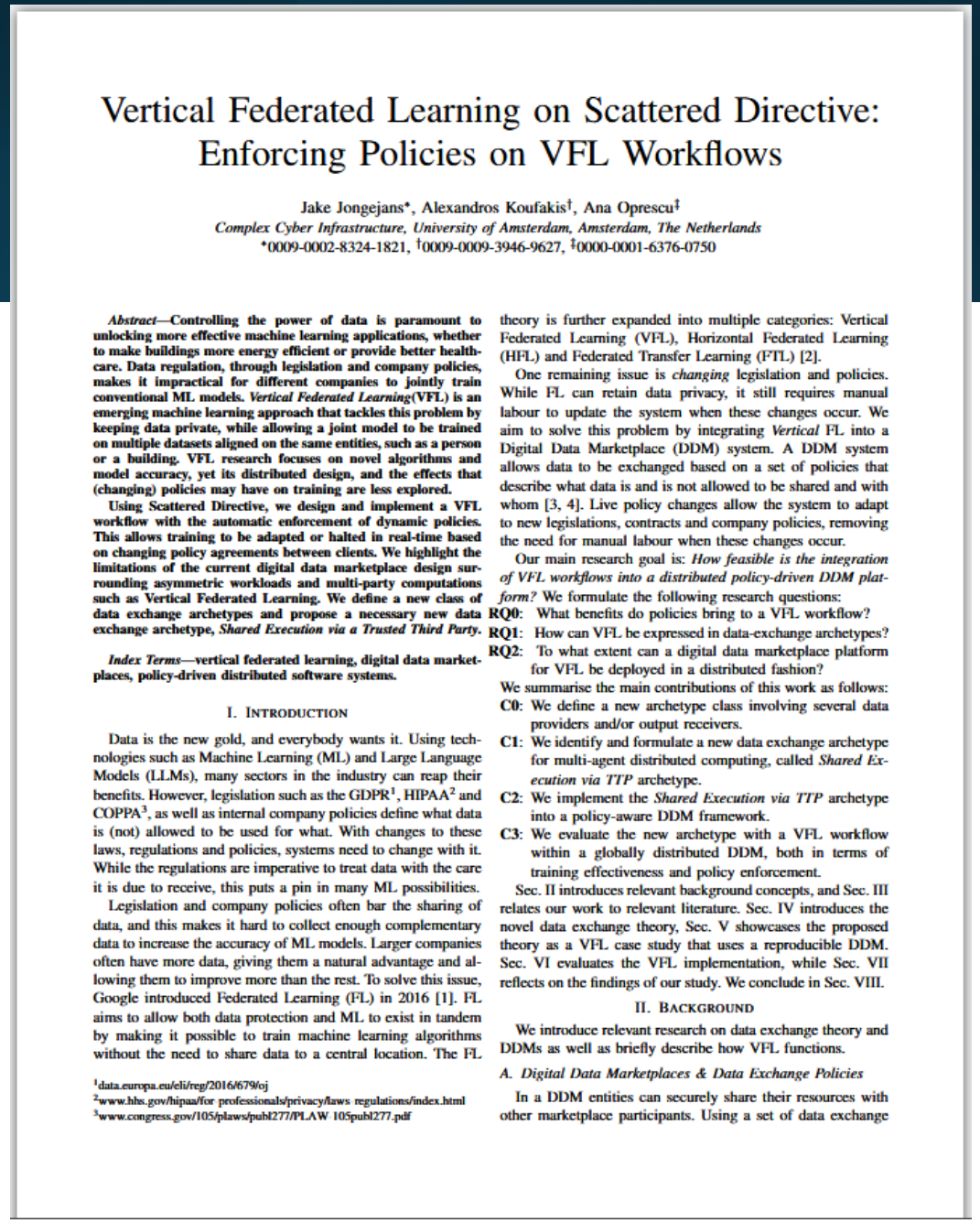
Vertical Federated Learning on Scattered Directive: Enforcing Policies on VFL Workflow

- *Jake Jongejans*
- *Alexandros Koufakis*
- *Dr. Ana Oprescu*

To be published:

IEEE Big Data at Macau, 8-11 December

https://bigdataieee.org/BigData2025/



Vertical Federated Learning on Scattered Directive: Enforcing Policies on VFL Workflows

Jake Jongejans*, Alexandros Koufakis†, Ana Oprescu‡
Complex Cyber Infrastructure, University of Amsterdam, Amsterdam, The Netherlands
*0009-0002-8324-1821, †0009-0009-3946-9627, ‡0000-0001-6376-0750

***Abstract*—Controlling the power of data is paramount to unlocking more effective machine learning applications, whether to make buildings more energy efficient or provide better healthcare. Data regulation, through legislation and company policies, makes it impractical for different companies to jointly train conventional ML models. *Vertical Federated Learning*(VFL) is an emerging machine learning approach that tackles this problem by keeping data private, while allowing a joint model to be trained on multiple datasets aligned on the same entities, such as a person or a building. VFL research focuses on novel algorithms and model accuracy, yet its distributed design, and the effects that (changing) policies may have on training are less explored.**

Using Scattered Directive, we design and implement a VFL workflow with the automatic enforcement of dynamic policies. This allows training to be adapted or halted in real-time based on changing policy agreements between clients. We highlight the limitations of the current digital data marketplace design surrounding asymmetric workloads and multi-party computations such as Vertical Federated Learning. We define a new class of data exchange archetypes and propose a necessary new data exchange archetype, *Shared Execution via a Trusted Third Party*.

***Index Terms*—vertical federated learning, digital data marketplaces, policy-driven distributed software systems.**

I. Introduction

Data is the new gold, and everybody wants it. Using technologies such as Machine Learning (ML) and Large Language Models (LLMs), many sectors in the industry can reap their benefits. However, legislation such as the GDPR[1], HIPAA[2] and COPPA[3], as well as internal company policies define what data is (not) allowed to be used for what. With changes to these laws, regulations and policies, systems need to change with it. While the regulations are imperative to treat data with the care it is due to receive, this puts a pin in many ML possibilities.

Legislation and company policies often bar the sharing of data, and this makes it hard to collect enough complementary data to increase the accuracy of ML models. Larger companies often have more data, giving them a natural advantage and allowing them to improve more than the rest. To solve this issue, Google introduced Federated Learning (FL) in 2016 [1]. FL aims to allow both data protection and ML to exist in tandem by making it possible to train machine learning algorithms without the need to share data to a central location. The FL theory is further expanded into multiple categories: Vertical Federated Learning (VFL), Horizontal Federated Learning (HFL) and Federated Transfer Learning (FTL) [2].

One remaining issue is *changing* legislation and policies. While FL can retain data privacy, it still requires manual labour to update the system when these changes occur. We aim to solve this problem by integrating *Vertical* FL into a Digital Data Marketplace (DDM) system. A DDM system allows data to be exchanged based on a set of policies that describe what data is and is not allowed to be shared and with whom [3, 4]. Live policy changes allow the system to adapt to new legislations, contracts and company policies, removing the need for manual labour when these changes occur.

Our main research goal is: *How feasible is the integration of VFL workflows into a distributed policy-driven DDM platform?* We formulate the following research questions:

RQ0: What benefits do policies bring to a VFL workflow?
RQ1: How can VFL be expressed in data-exchange archetypes?
RQ2: To what extent can a digital data marketplace platform for VFL be deployed in a distributed fashion?

We summarise the main contributions of this work as follows:

C0: We define a new archetype class involving several data providers and/or output receivers.
C1: We identify and formulate a new data exchange archetype for multi-agent distributed computing, called *Shared Execution via TTP* archetype.
C2: We implement the *Shared Execution via TTP* archetype into a policy-aware DDM framework.
C3: We evaluate the new archetype with a VFL workflow within a globally distributed DDM, both in terms of training effectiveness and policy enforcement.

Sec. II introduces relevant background concepts, and Sec. III relates our work to relevant literature. Sec. IV introduces the novel data exchange theory, Sec. V showcases the proposed theory as a VFL case study that uses a reproducible DDM. Sec. VI evaluates the VFL implementation, while Sec. VII reflects on the findings of our study. We conclude in Sec. VIII.

II. Background

We introduce relevant research on data exchange theory and DDMs as well as briefly describe how VFL functions.

A. Digital Data Marketplaces & Data Exchange Policies

In a DDM entities can securely share their resources with other marketplace participants. Using a set of data exchange

[1] data.europa.eu/eli/reg/2016/679/oj
[2] www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html
[3] www.congress.gov/105/plaws/publ277/PLAW-105publ277.pdf

Future work & Acknowledgements

- Computational cost monitoring pipeline
- Introduction more advance normative reasoning capabilities
- More advanced disagreement resolution with versioning
- Exploration of new use cases

- This work was supported by the project: "*Budget-neutral sustainable home improvements (MOOI-224049)*". executed by the consortium of Prets B.V., Altum AI B.V., Enjins B.V., Skyleague Consulting B.V., Unravel Behavior B.V., Universiteit van Amsterdam and Universiteit Maastricht.

- This work was also supported by CIENA

UNIVERSITY OF AMSTERDAM

Thank you!

Jake Jongejans, **Alexandros Koufakis**, Ana Oprescu
University of Amsterdam

https://cci-research.nl/

https://github.com/orgs/DYNAMOS-UVA/repositories

CIENA BOOTH: #3330

ciena

prets embedded lending