Advancing Distributed and Decentralized AI through Stochastic Ontology

Hulya Seferoglu

University of Illinois Chicago

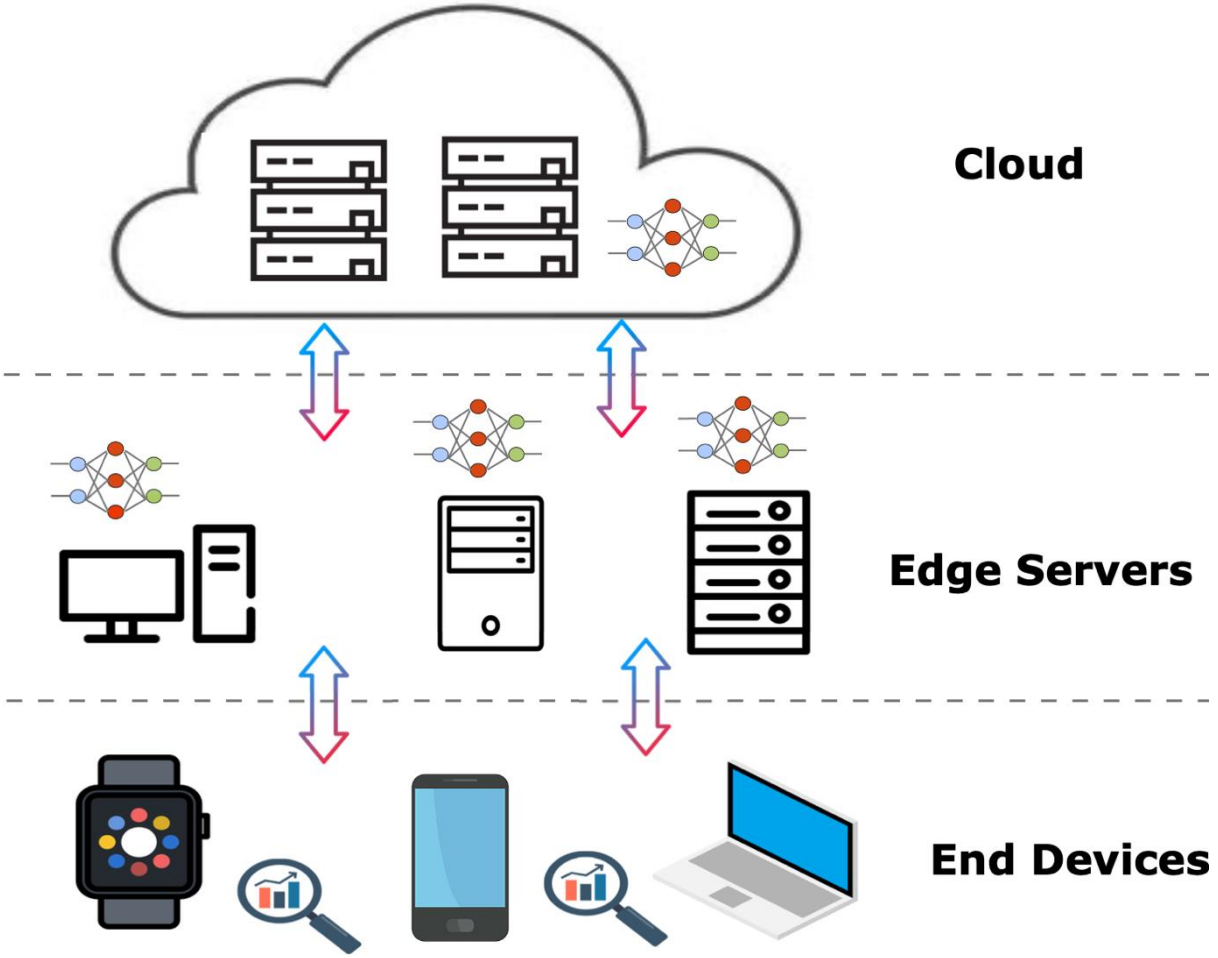
Recent Work

- Fast Inference with Model-Distribution and Parallelization
- Communication-Efficient and Heterogeneous Federated Learning
- Serverless Federated Learning
- Privacy-Preserving and Secure Distributed Learning

Recent Work

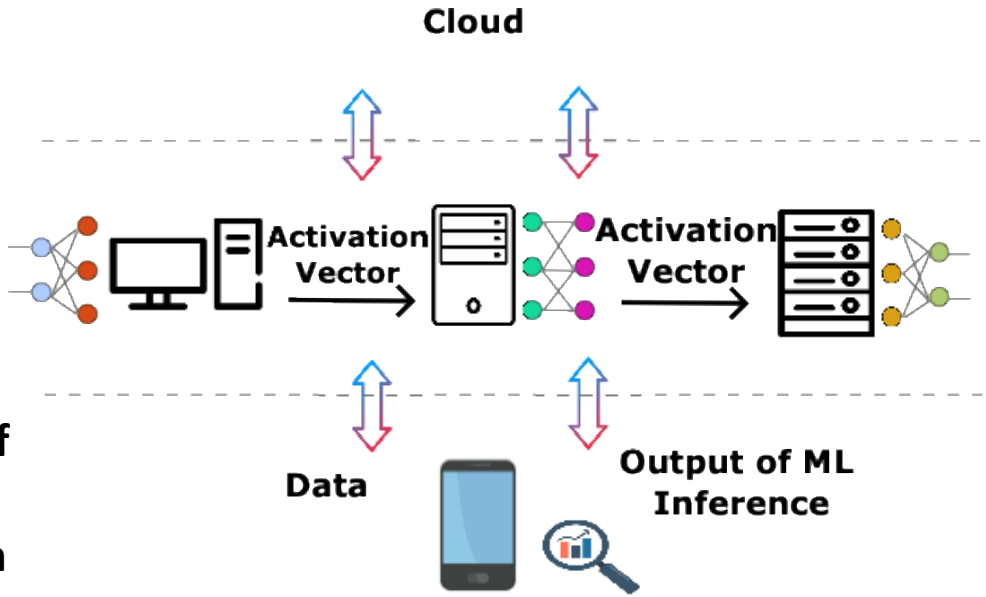
- **Fast Inference with Model-Distribution and Parallelization**
- **Communication-Efficient and Heterogeneous Federated Learning**
- **Serverless Federated Learning**
- **Privacy-Preserving and Secure Distributed Learning**

Fast Inference with Model-Distribution and Parallelization



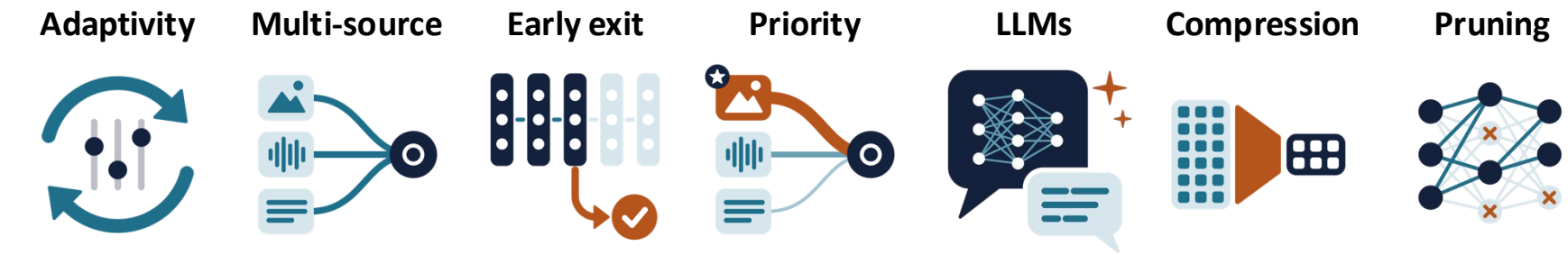
Fast Inference with Model-Distribution and Parallelization

- A promising solution:
  - Model-distributed inference, ML model is distributed across workers.



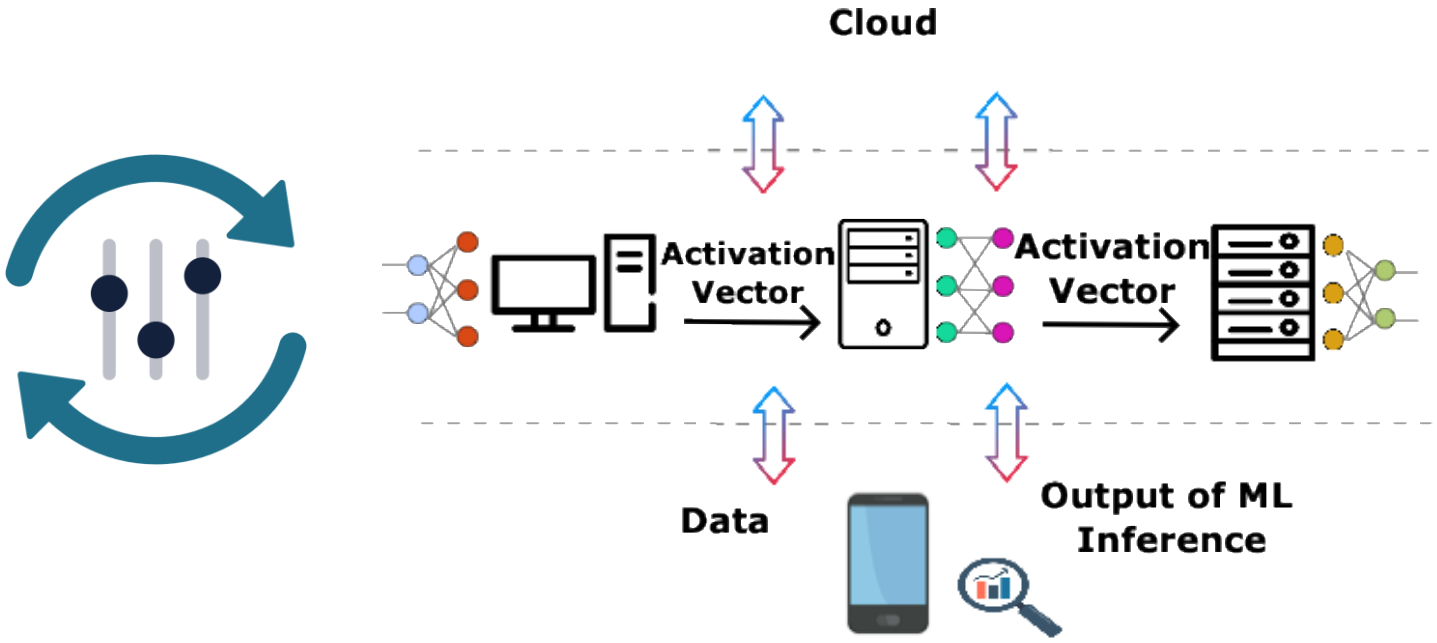
- The end users having data
  - may process a few layers of an AI model,
  - and transmit the activation vector of its last layer to a neighboring node.
- The neighboring node receives an activation vector and performs the calculations of the layers that are assigned to it.
- Worker that calculates the last layers of the AI model obtains and sends the output back to the end user that has the original input data.

Our Work on Model-Distributed Inference



2023	2024	2025		2026	
OJ-COMS	ICASSP-W	LANMAN	ICMLCN	LANMAN	MobiHoc	TMLCN
				Best paper		

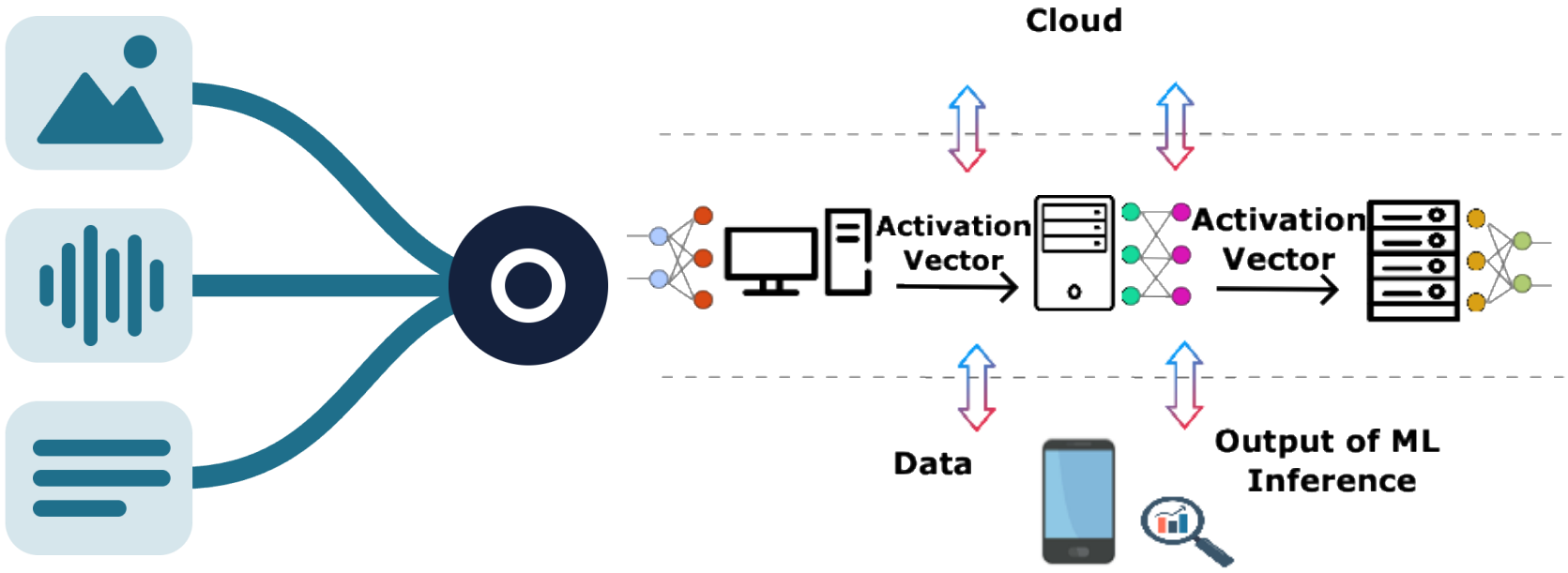
Fast Inference with Model-Distribution and Parallelization



Adaptivity: Design MDI Mechanisms Adaptive to Heterogeneous and Time-Varying Resources

P. Li, E. Koyuncu, H. Seferoglu, "Adaptive and Resilient Model-Distributed Inference in Edge Computing Systems," in IEEE Open Journal of the Communications Society, 2023.

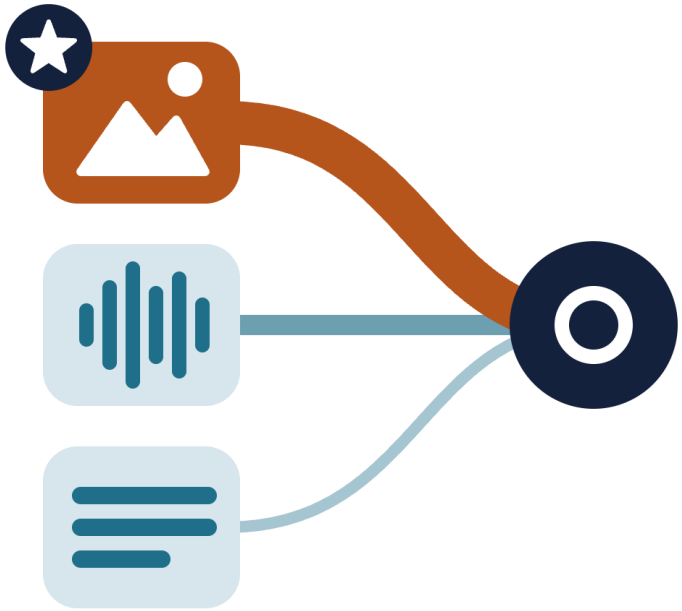
Fast Inference with Model-Distribution and Parallelization

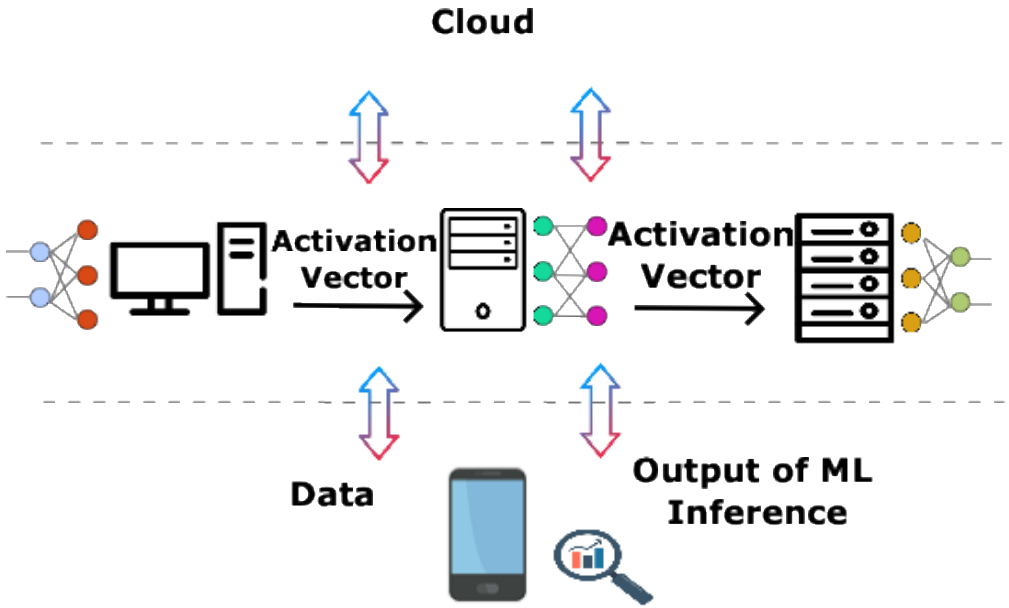


Multiple data/modality: Supporting Multiple Data Sources and Modalities in MDI

P. Li, H. Seferoglu, E. Koyuncu, "Model-Distributed Inference in Multi-Source Edge Networks," Timely and Private Machine Learning over Networks Workshop, ICASSP, June 2023.

Fast Inference with Model-Distribution and Parallelization

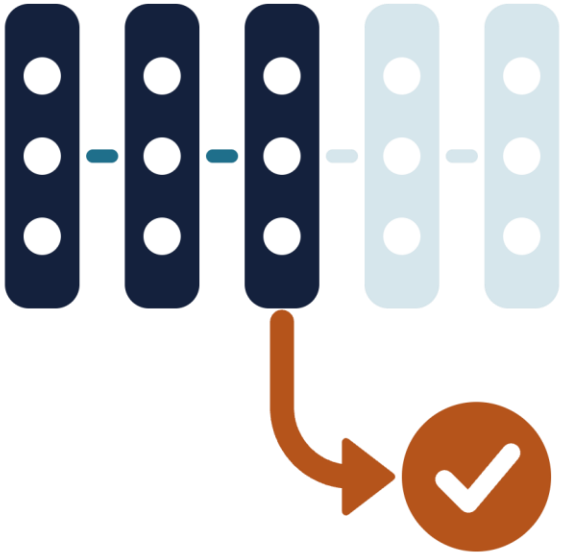


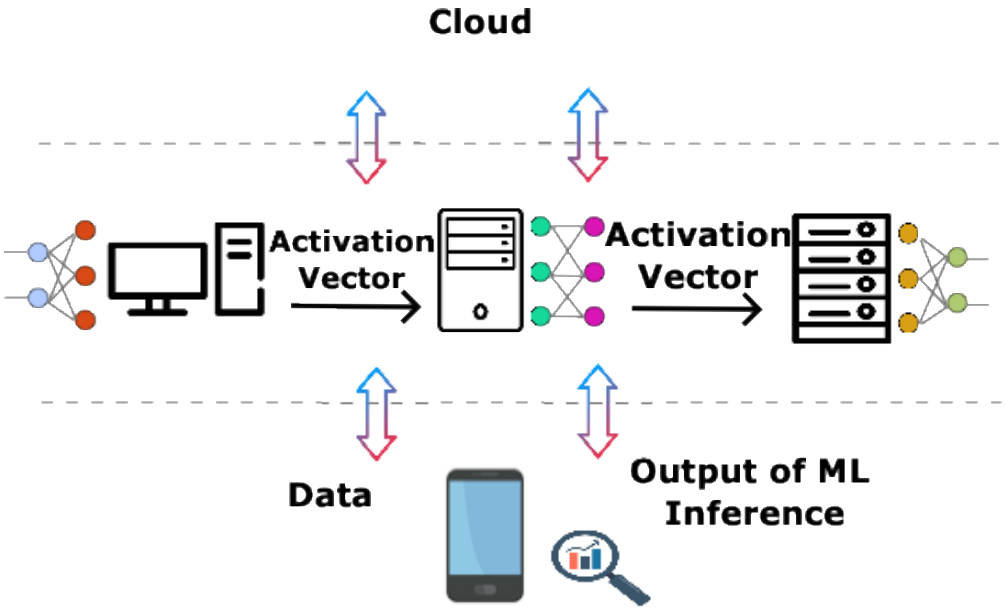


Priority: Sources and modalities differ in importance. MDI should give the important ones more compute and bandwidth.

T. Li, H. Seferoglu, "Priority-Aware Model-Distributed Inference at Edge Networks," in IEEE ICMLCN, Barcelona, Spain, May 2025.

Fast Inference with Model-Distribution and Parallelization

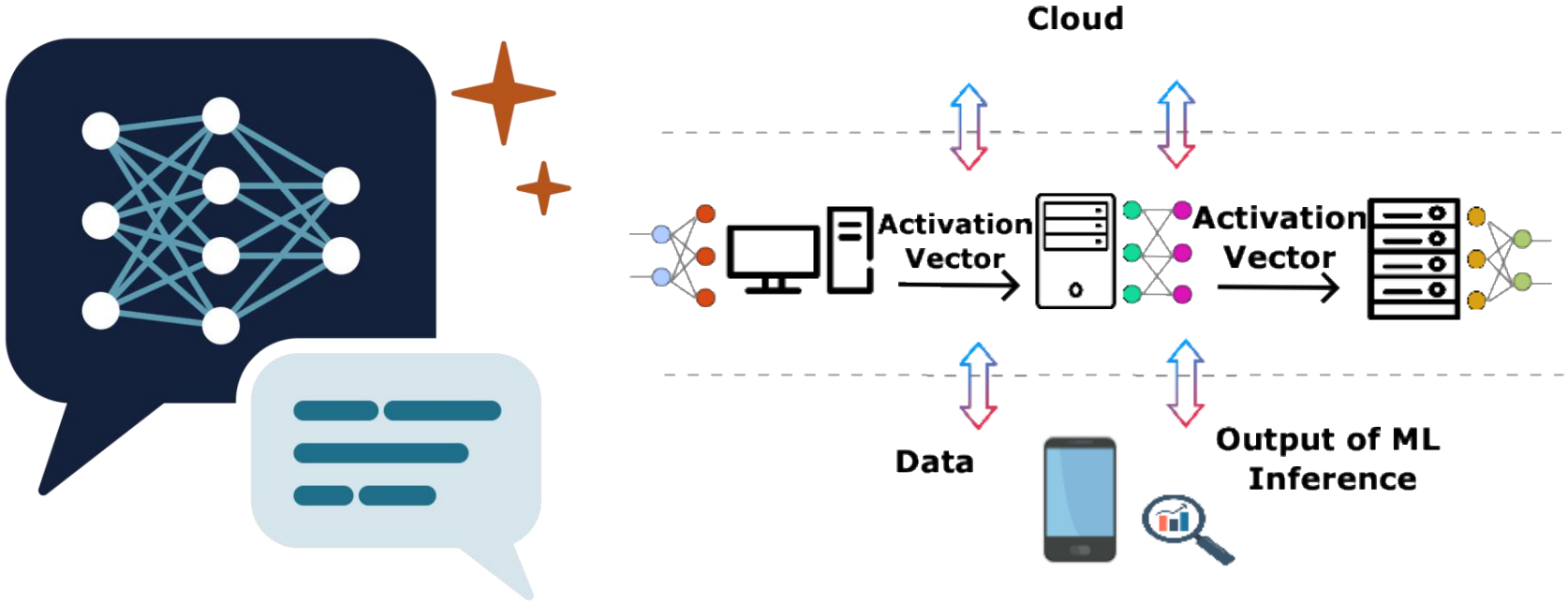




Early-Exit: Allow early exit if the confidence level is reached.

M. Colocrese, E. Koyuncu, H. Seferoglu, “Early-Exit meets Model-Distributed Inference at Edge Networks,” in IEEE LANMAN, Boston, MA, July 2024.

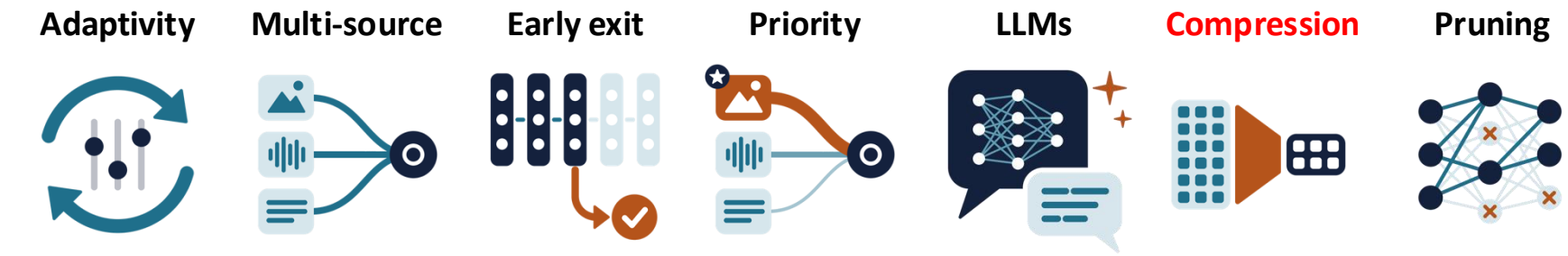
Fast Inference with Model-Distribution and Parallelization



LLMs: Apply MDI to LLMs.

D. Macario, H. Seferoglu, E. Koyuncu, "Model-Distributed Inference for Large Language Models at the Edge," in IEEE LANMAN, Lille, France, July 2025. **(Best paper award.)**

Our Work on Model-Distributed Inference





OJ-COMS | ICASSP-W | LANMAN | ICMLCN | LANMAN | MobiHoc | TMLCN

Best paper

MDI with Compression

Problem

The accuracy–throughput tradeoff

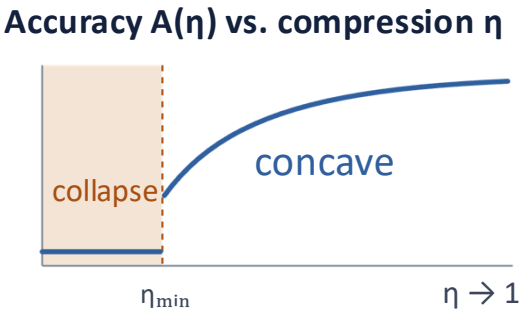


A model is split into stages and pipelined across edge devices. Stage outputs a_i are compressed to $\eta_i \cdot a_i$ before crossing links whose capacity $c_i(t)$ changes over time.

THROUGHPUT CONSTRAINT, PER TASK

$$\max\{\tau_i, a_i \eta_i / c_i(t)\} \leq 1 / R(t)$$

The slowest stage, computation or communication, sets the pipeline cycle.



Schematic of the measured shape (e.g. ResNet-56 on CIFAR-10 with top-k): collapses below η_{min}, monotone and concave above it.

Goal: pick compression rates (and, with several tasks, compute and bandwidth shares) that maximize weighted accuracy while every task meets its target rate R(t).

MDI with Compression

Known Channels (CSI-Aware)

Closed form for one task, water-filling for multiple tasks

One task

$$\eta_i^* = \min\{ 1, c_i(t) / (R(t) \cdot a_i) \}$$

- Compress each hop just enough to fill its link at the target rate, and never more.
- The unique optimum whenever accuracy is non-decreasing in η.

Many tasks sharing nodes and links

1. Split each node's compute and each link's bandwidth among the tasks that use it.
2. Solved in polynomial time by per-link water-filling.

Intuition: Bandwidth allocated to the task and hop where accuracy is most sensitive to compression.

MDI with Compression

Unknown Channels (Estimated CSI)

Stochastic dual descent, one slot at a time

STEP 1
Estimate

Channel estimate $\hat{c}(t)$ from past slots.

→

STEP 2
Primal step

$max\ A(\eta) - \mu\ \lambda_t\ \hat{D}(\eta, \hat{c})$

Convex for one task; block-coordinate descent over shares and η for many.

→

STEP 3
Execute

Run the pipeline and observe the true bottleneck delay D(t) under the real channel.

→

STEP 4
Dual update

$\lambda_{t+1} = max\{\varepsilon, \lambda_t + D(t) - 1/R(t)\}$

λ carries delay debt into the next slot.

On target

Long-run average delay meets each task's 1/R(t)

O(1/μ)

Bound on the average dual variable: larger μ, tighter delay control

O(μ)

Accuracy gap to the best stationary policy

MDI with Compression

Evaluation

Simulations and three edge testbeds

26

experiments: 12 offline, 14 on Raspberry Pi, Jetson Orin Nano and PRESCIENT

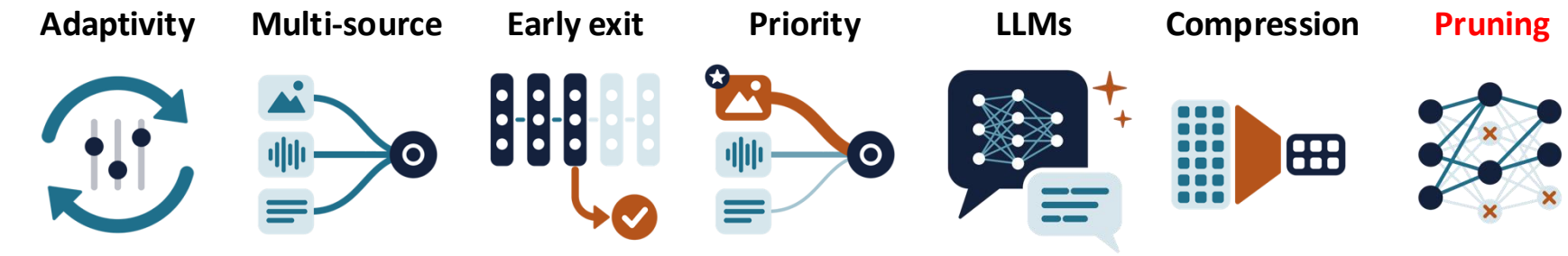
7

vision and language tasks, MLP to Llama-3.1 8B, under top-k, quantization, LLM.int8

Run	Setting	Other No-CSI	Ours
P-L1-G	Llama-3.1 8B, 1 Gbps LAN	1.000, all miss	0.942, meets
P-L1-E	Llama-3.1 8B, W-Edge	≈0.99, all miss	0.857, meets
P-VV-G	2× ResNet-56, 1 Gbps LAN	0.758, all miss	0.759, meets
P-VL-E	ResNet-56 + Llama-3.1 8B, W-Edge	0.737, all meet	0.844, meets

Takeaway: without channel knowledge, ours meets the delay target at competitive accuracy where myopic and moving-average estimators fail.

Our Work on Model-Distributed Inference



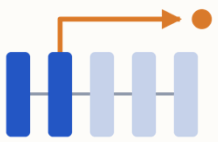
2023	2024	2025		2026	
OJ-COMS	ICASSP-W	LANMAN	ICMLCN	LANMAN	MobiHoc	TMLCN
				Best paper		

Ongoing/Future Directions

Fast Inference with Ontology-Guided Model Distribution



Model Partitioning: Using the stochastic ontology approach to define a structured set of rules that guide model partitioning.



Early-Exit: Leverage ontology-driven rules and semantic understanding of the data to enable joint optimization of model splitting and early-exit decisions.



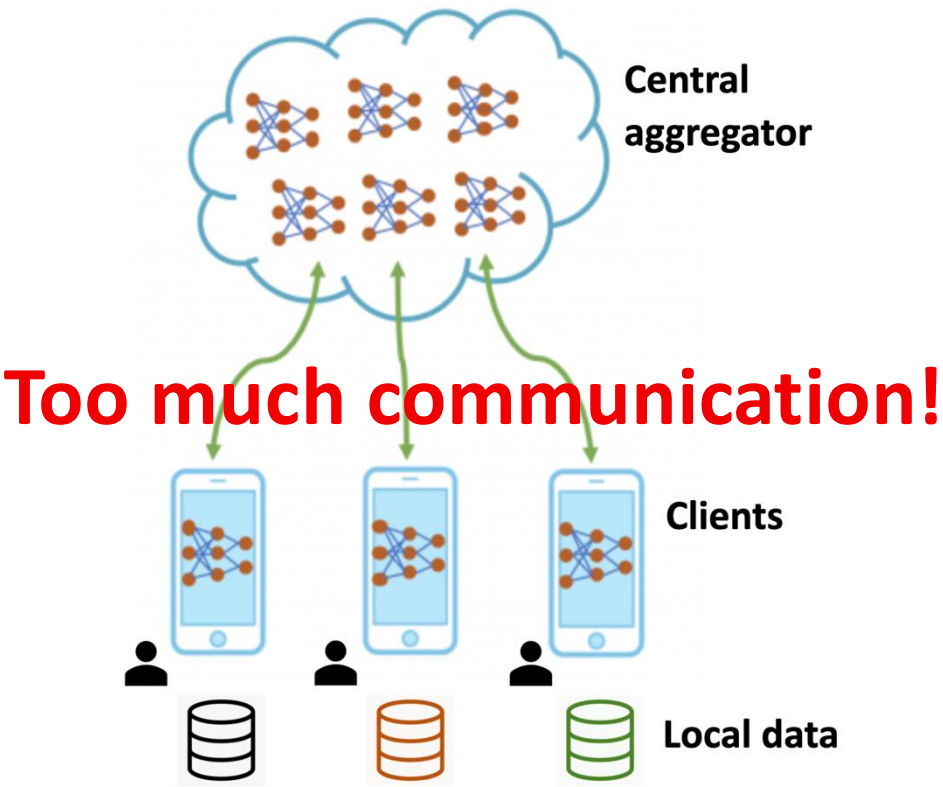
Compression: Use a semantic, ontology-guided compression approach.



Recent Work

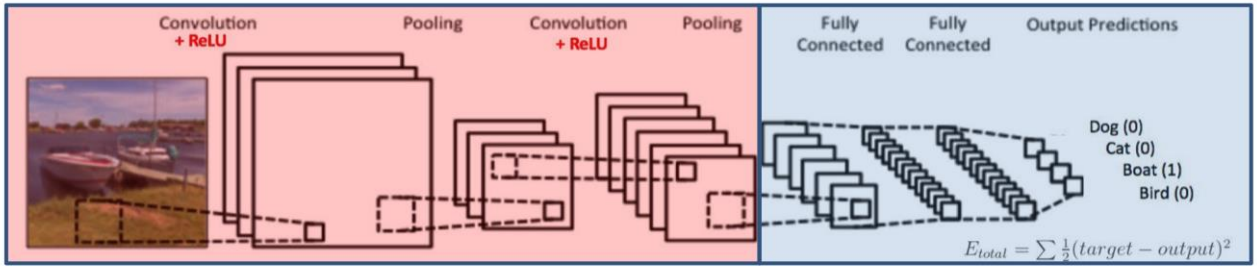
- Fast Inference with Model-Distribution and Parallelization
- Communication-Efficient and Heterogeneous Federated Learning
- Serverless Federated Learning
- Privacy-Preserving and Secure Distributed Learning

Federated Learning



Communication-Efficient Federated Learning with Differentiated Aggregation

Idea: The initial layers (representation extractor) of the model tend to be more consistent across different clients in comparison to the final layers (head).



Learning the representation

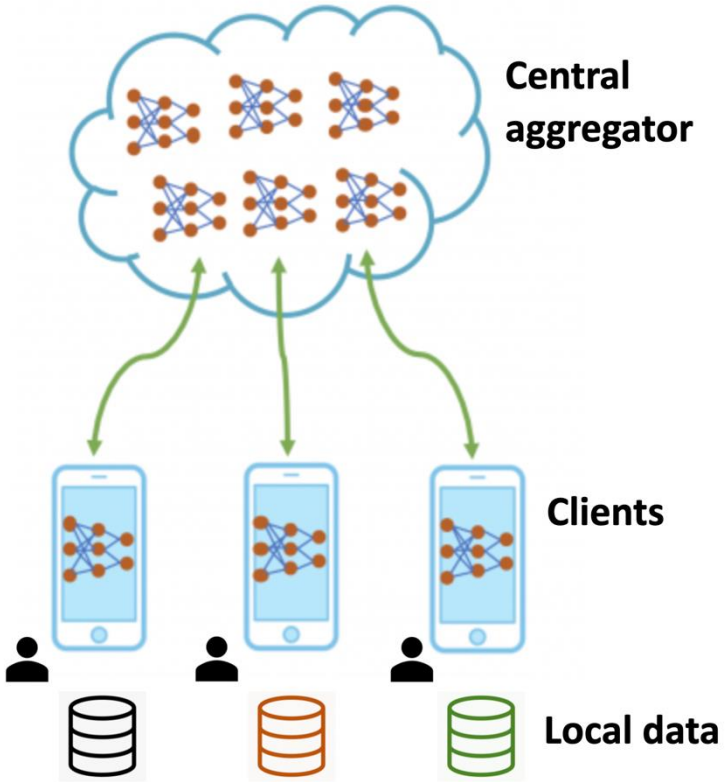
Learning classification

Approach: Initial layers can be aggregated less frequently → Reduces communication cost, improve generalization.

P. Gholami, H. Seferoglu, "Differentiated Aggregation to Improve Generalization in Federated Learning," in Transactions on Machine Learning Research (TMLR), 2025.

Federated Learning

More Challenges - Non-i.i.d. Data



Client drift

Non-i.i.d. data pulls each client's local steps toward its own optimum.

Participation bias

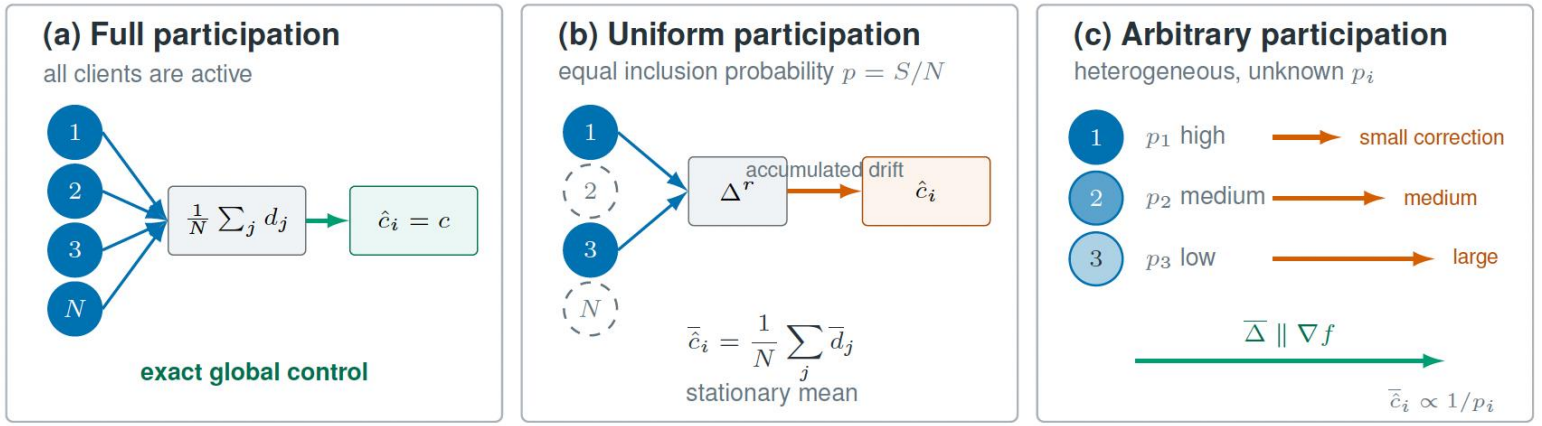
Clients that show up more often dominate, shifting what gets optimized.

TRUSS: Transmission Reduced Heterogeneous FL

Prior Work: SCAFFOLD mitigates client drift caused by heterogeneous local objectives, but requires each participating client to communicate both a model update and a control-variate update.

Goal: Preserve SCAFFOLD's correction for client heterogeneity while reducing its communication by half.

TRUSS: Each participating client transmits only one d-dimensional model update and stores a single compressed memory vector that reconstructs the required control correction.



TRUSS: Transmission Reduced Heterogeneous FL

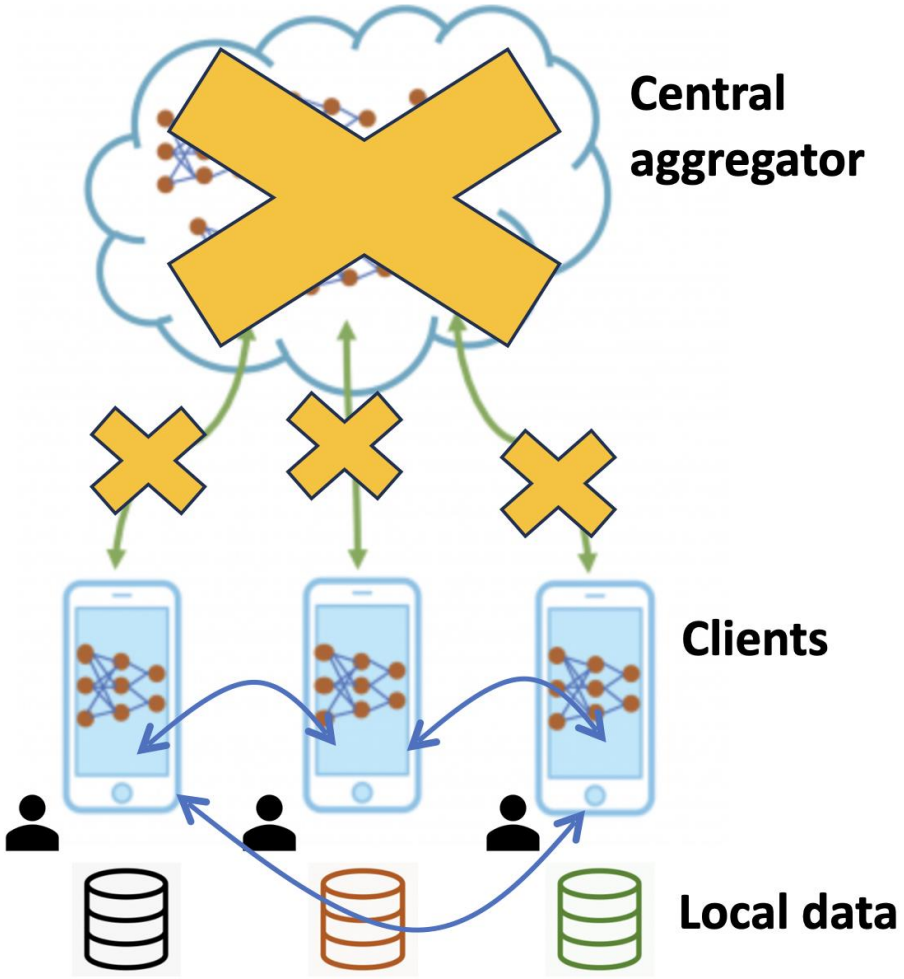
Table 1: Comparison of representative methods.

Method	No het. bound	K speedup	Participation	Comm.	Client mem.	Server mem.
FedAvg (McMahan et al., 2017)	×	✓	Uniform partial	2	0	1
FedProx (Li et al., 2020a)	×	✓	IID sampling	2	0	1
MIFA[a] (Gu et al., 2021)	×	✓	Bounded inactivity	2	0	$N+1$
FedAWE (Xiang et al., 2024)	×	✓	Non-stationary	2	1	1
FedAU (Wang and Ji, 2024)	×	✓	Stationary	2	0	1
ProxSkip (Mishchenko et al., 2022)	✓	×	Full	2	1	1
SCAFFOLD (Karimireddy et al., 2020)	✓	✓	Uniform	4	1	2
SCALLION (Huang et al., 2024)	✓	✓	Uniform	3	1	2
SG-FOCUS (Ying et al., 2025)	✓	×	Stationary	2	1	2
FedSUM-CR[b] (You and Pu, 2025)	✓	✓	Bounded inactivity	2	2	2
TRUSS[e]	✓	✓	Non-stationary	**2**	**1**	**1**

Recent Work

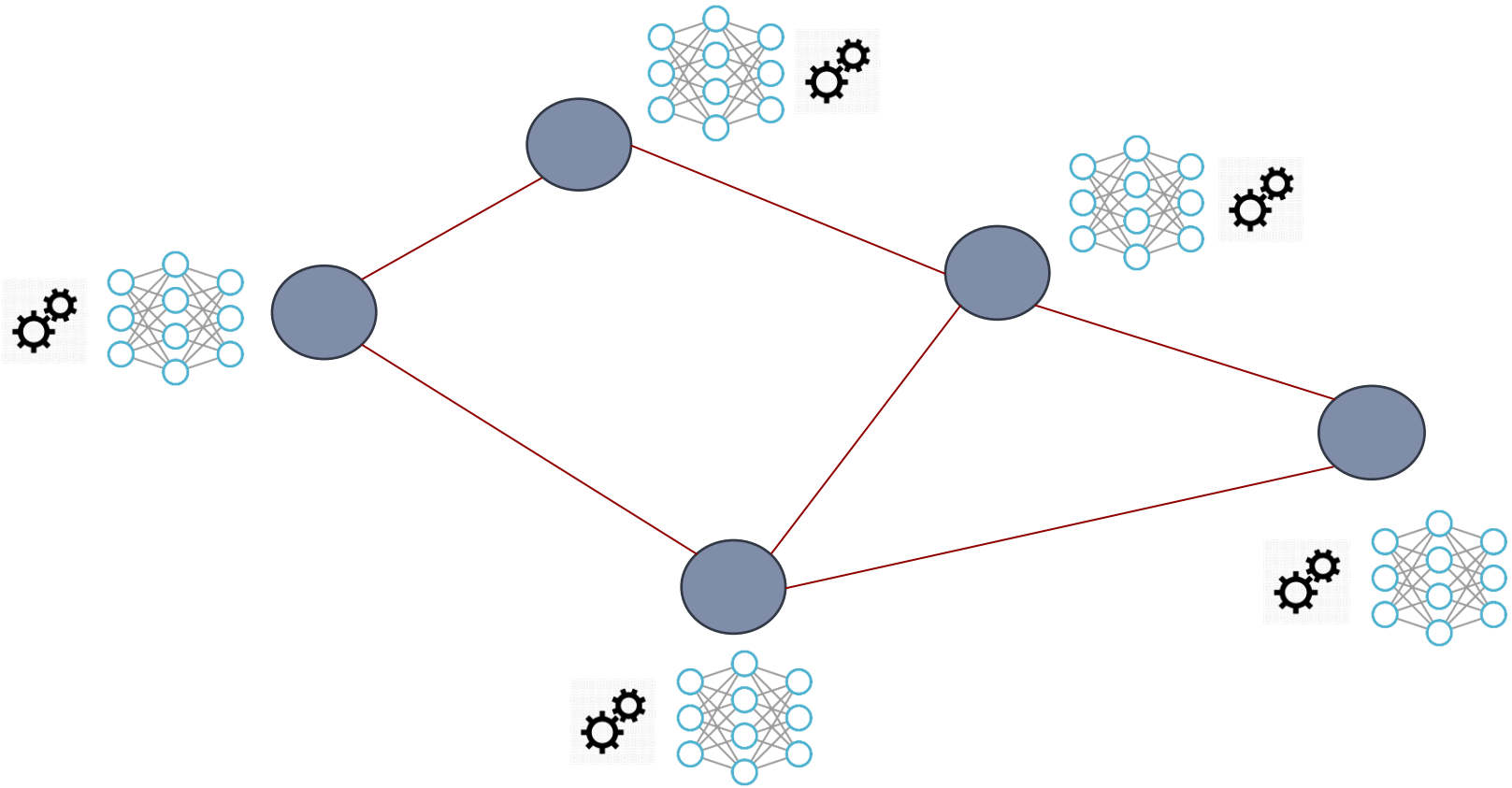
- Fast Inference with Model-Distribution and Parallelization
- Communication-Efficient and Heterogeneous Federated Learning
- Serverless Federated Learning
- Privacy-Preserving and Secure Distributed Learning

Serverless Federated Learning



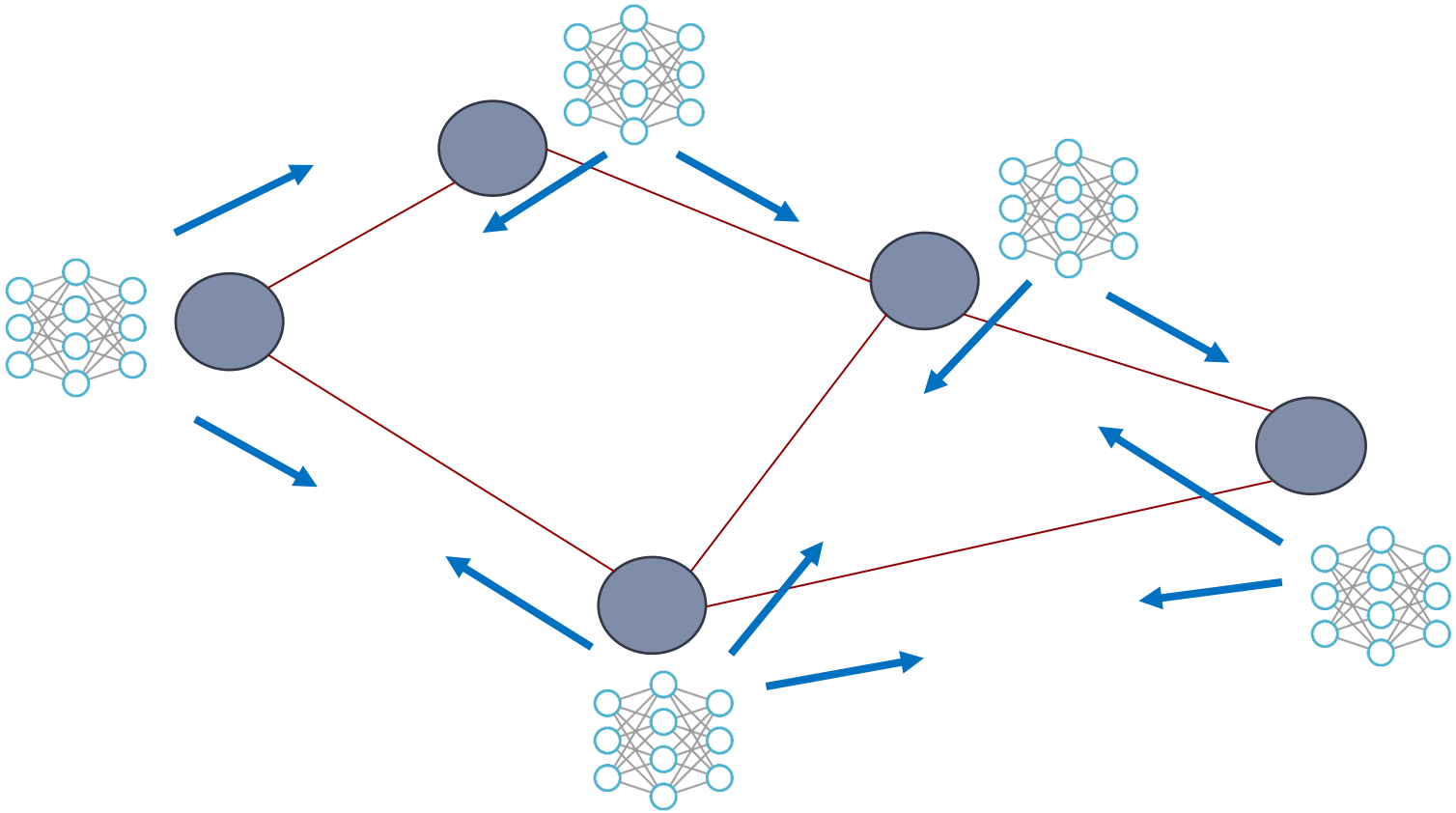
Gossip Algorithms

- **Decentralized algorithms have been extensively studied in the literature; gossip algorithms are widely used.**



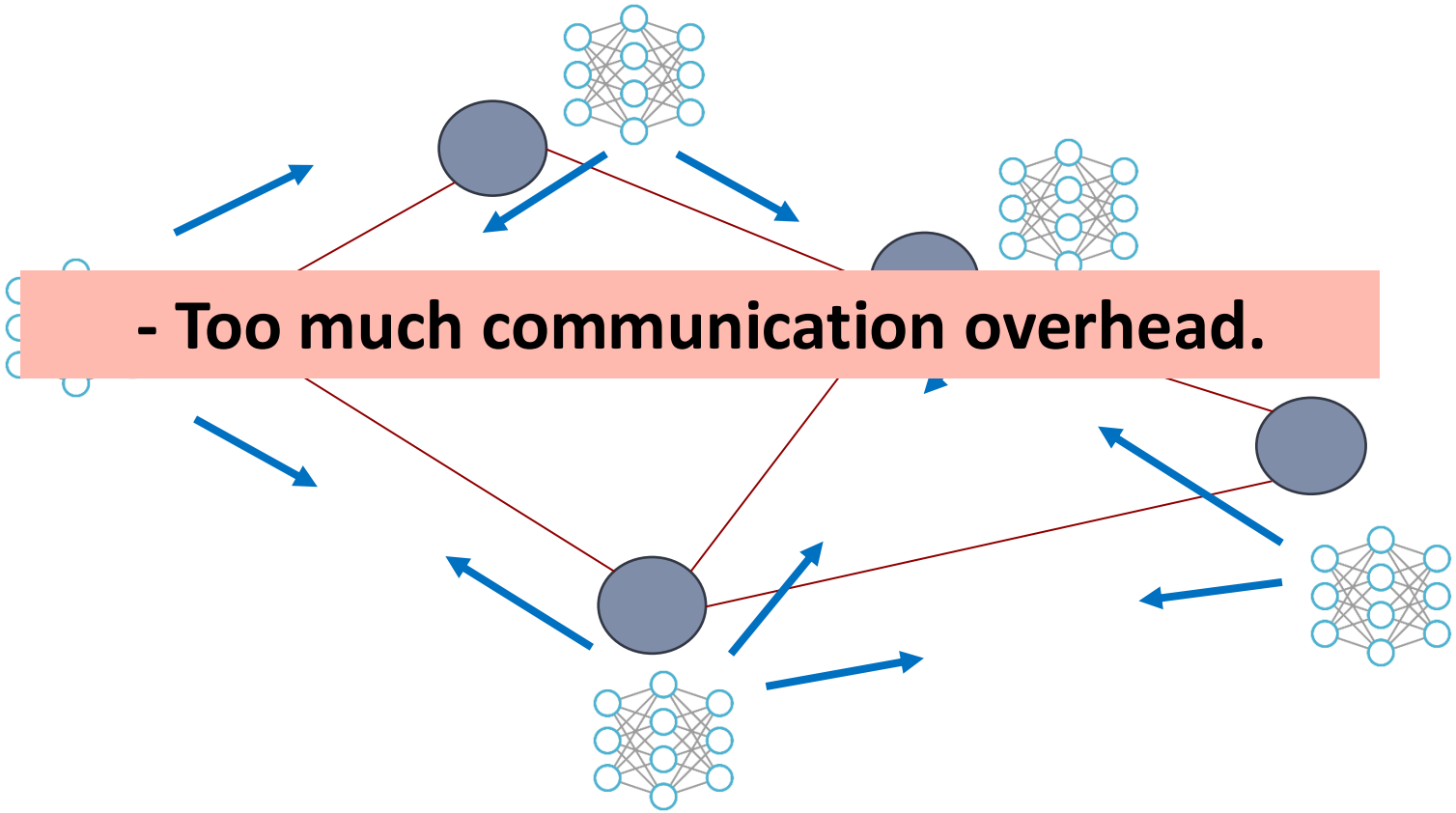
Gossip Algorithms

- **Decentralized algorithms have been extensively studied in the literature, with gossip algorithms receiving the lion's share of research attention.**



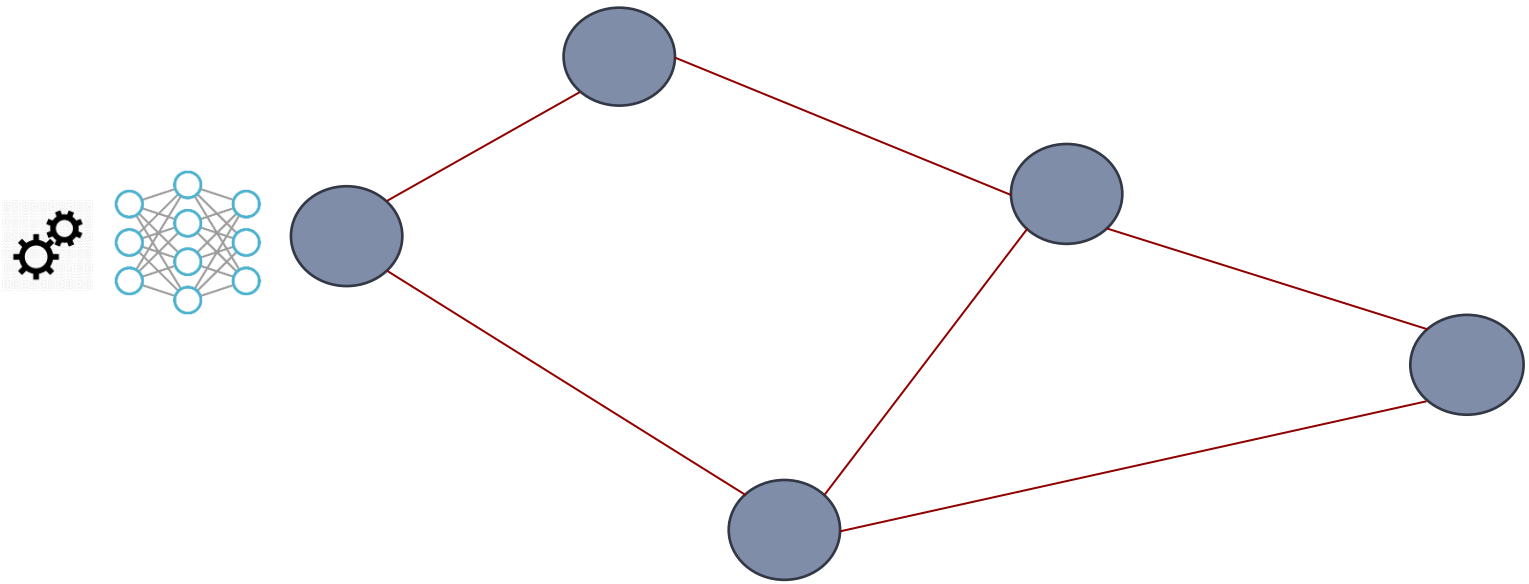
Gossip Algorithms

- **Decentralized algorithms have been extensively studied in the literature, with gossip algorithms receiving the lion's share of research attention.**



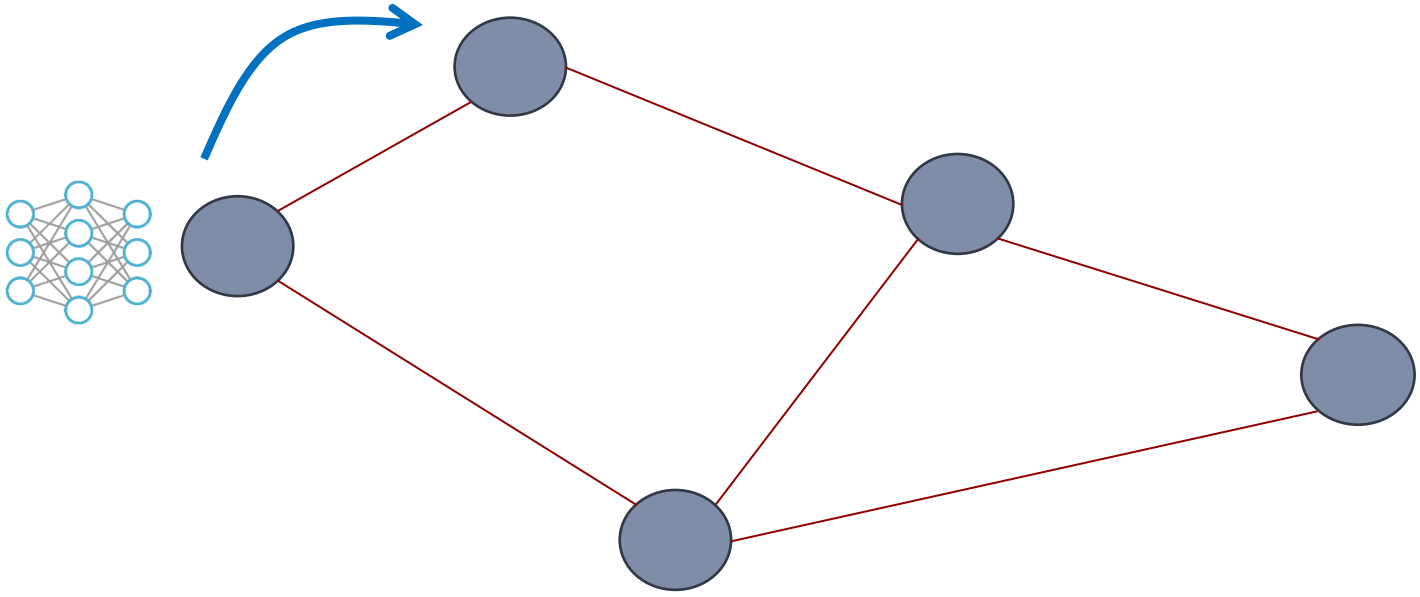
Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**



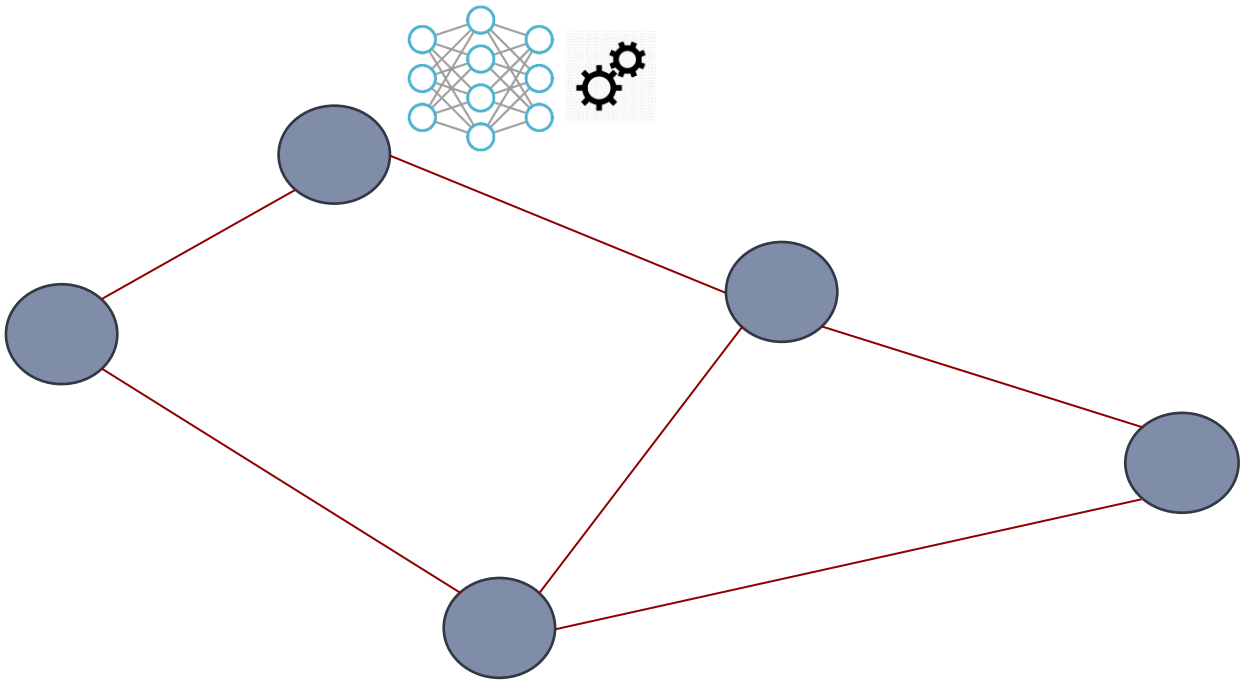
Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**



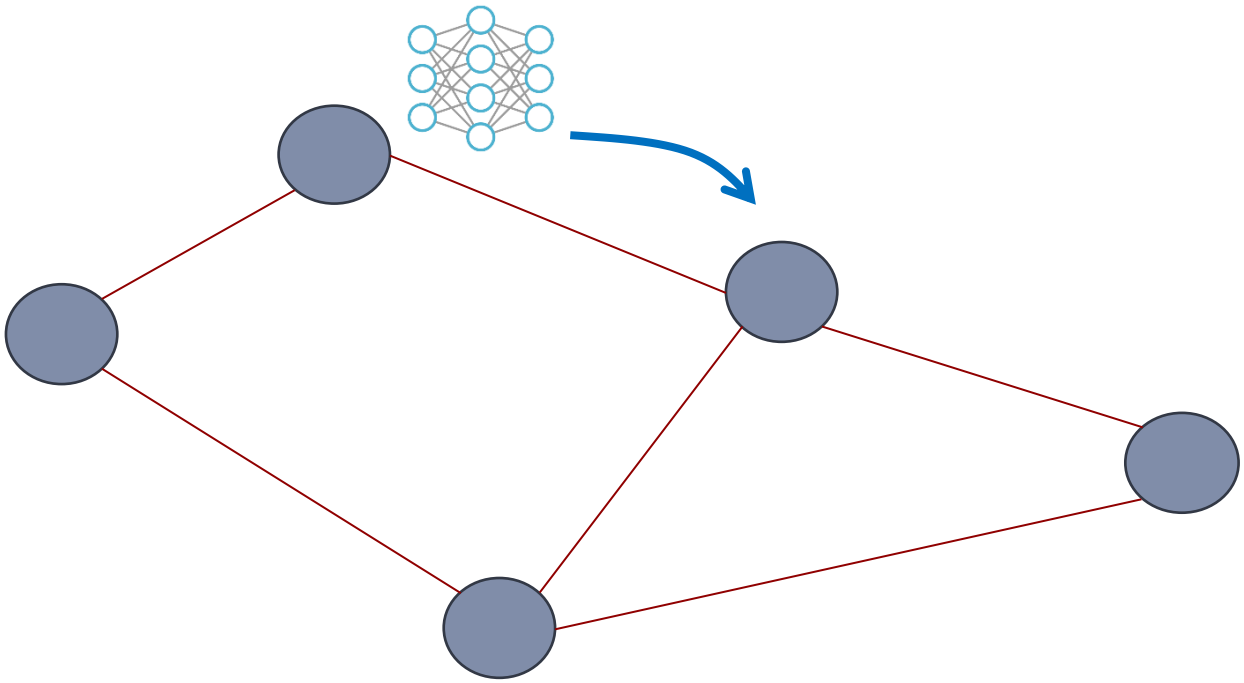
Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**



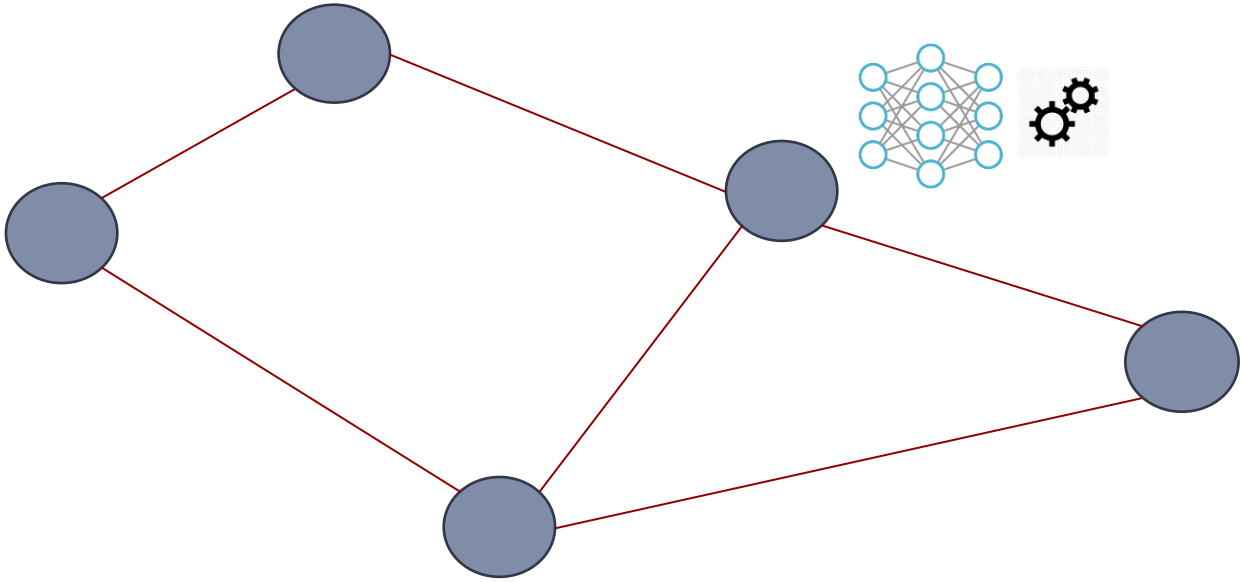
Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**



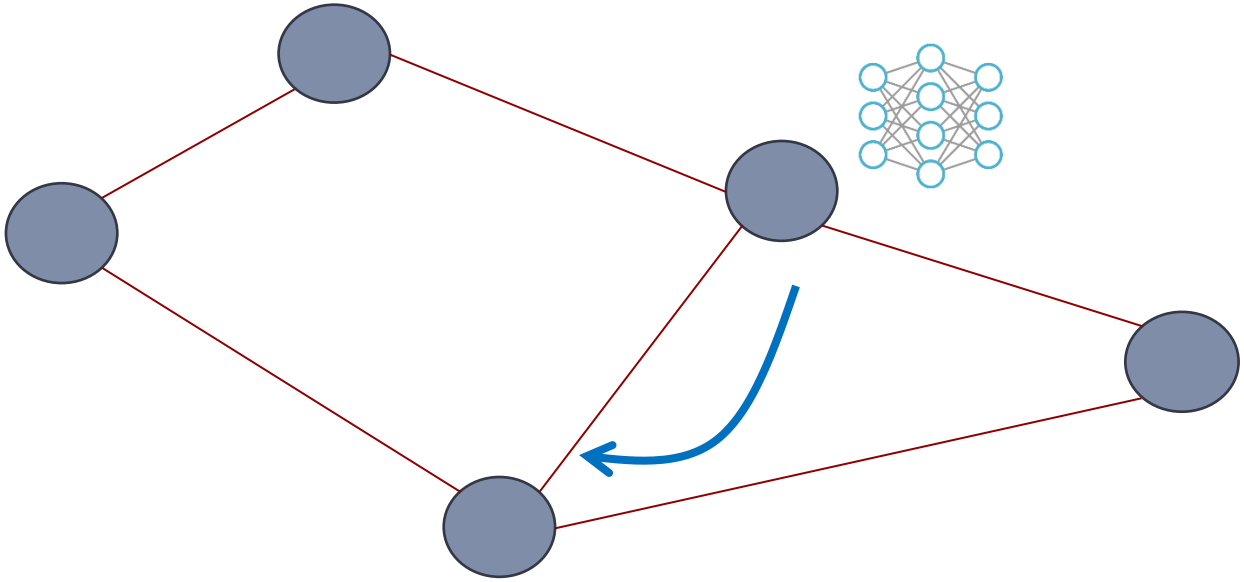
Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**



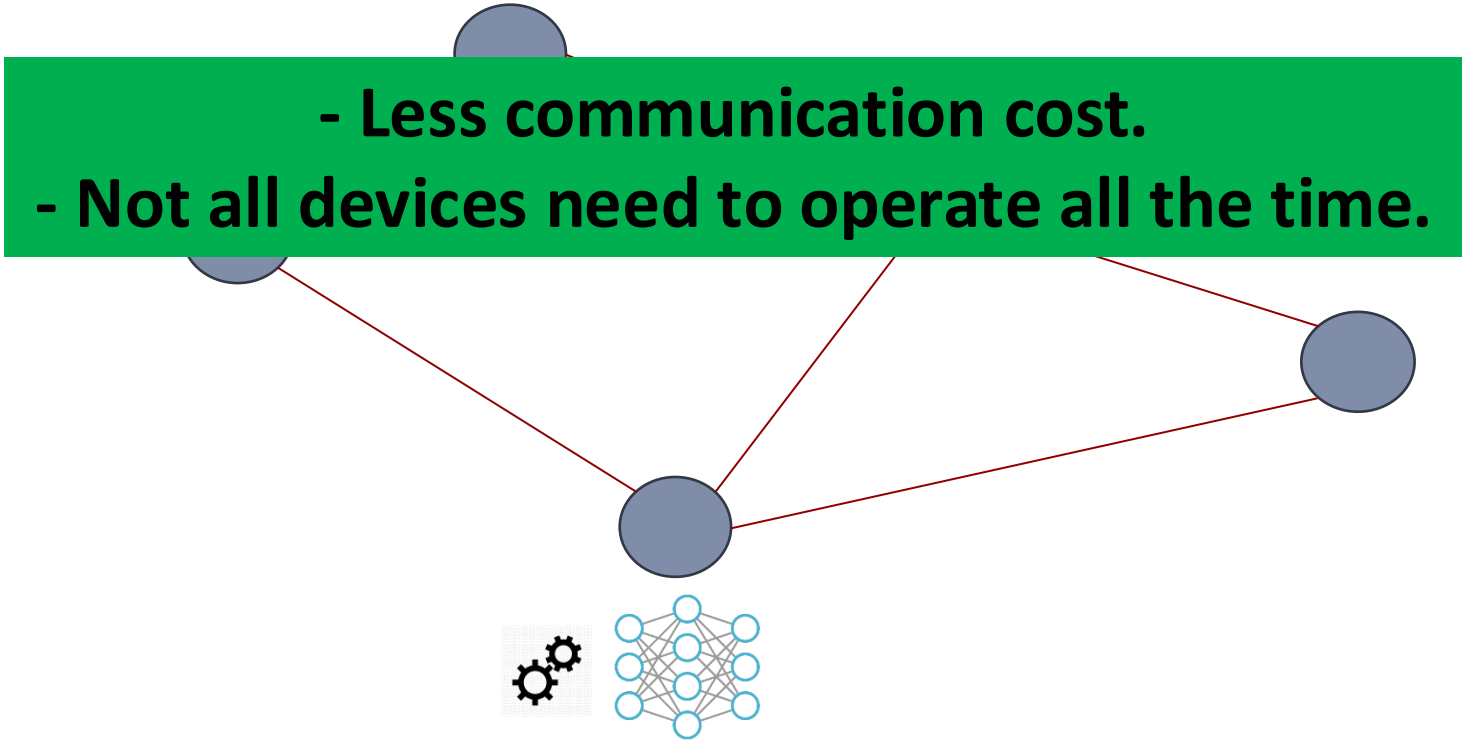
Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**



Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**

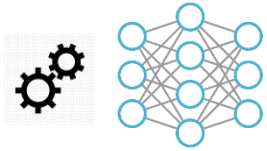


Random-Walk Based Learning

- **Random-walk based learning is emerging as an alternative.**

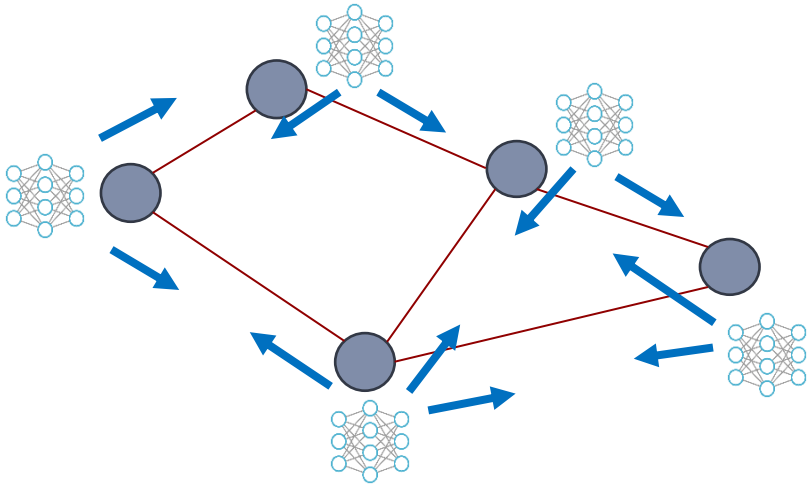
- Less communication cost.
- Not all devices need to operate all the time.

- **Convergence is slow.**
- **Model may not fit into a device.**
- **Can we further reduce the communication cost?**

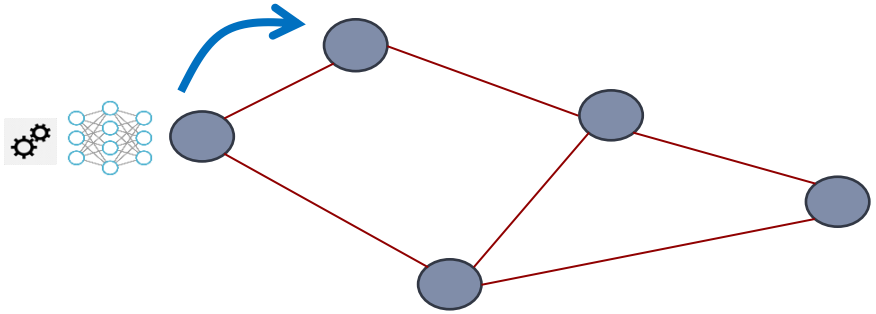


Serverless Federated Learning - DIGEST

Synchronous and Asynchronous Gossip



Random Walk



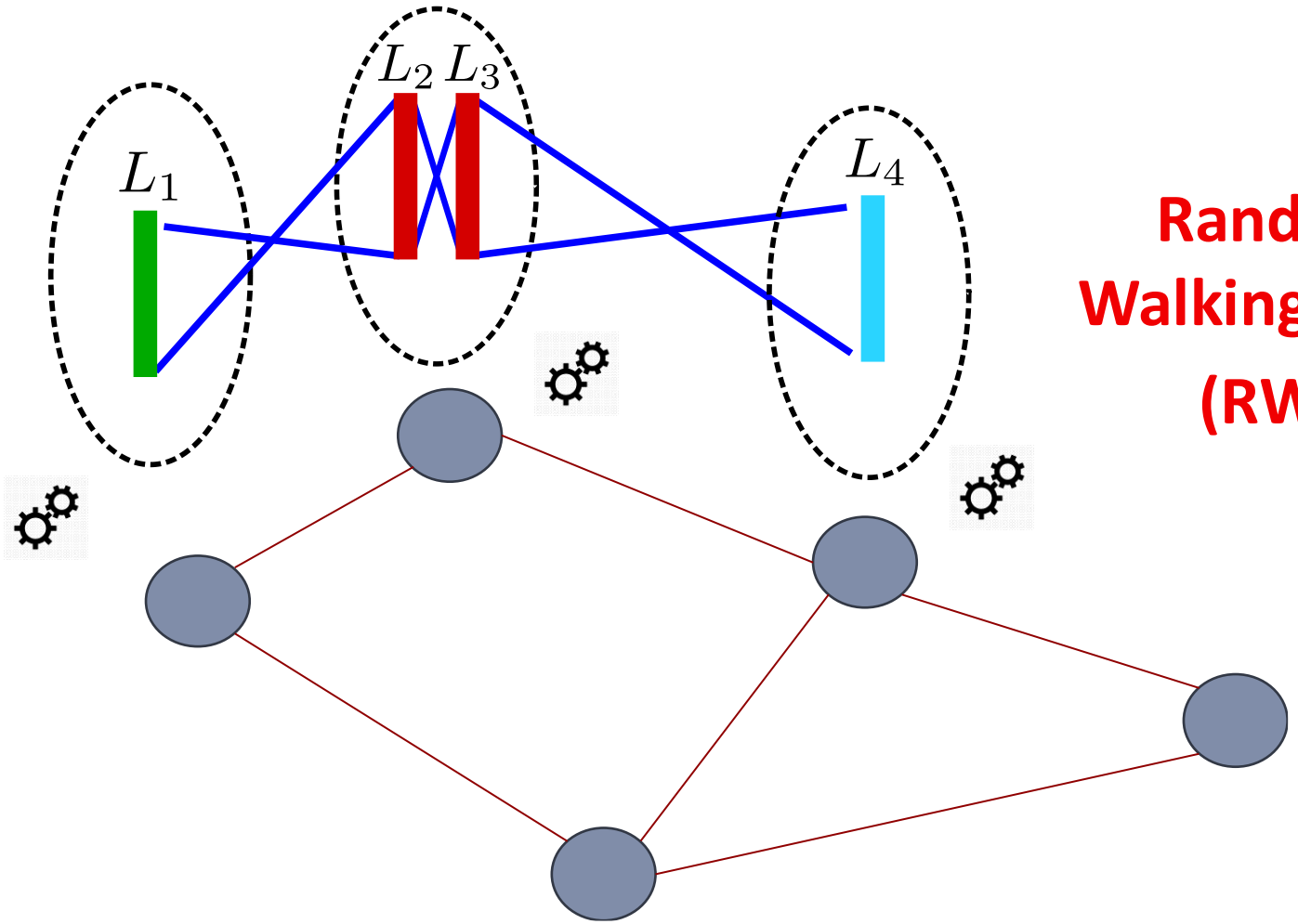
DIGEST

Idea: Take advantage of the best of Gossip and Random-Walk

- Update model all the time like Gossip, but do local updates
- Exchange models periodically using random walk idea.

P. Gholami, H. Seferoglu, "DIGEST: Fast and Communication Efficient Decentralized Learning with Local Updates," in IEEE Transactions on Machine Learning in Communications and Networking (TMLCN), Jan. 2024.

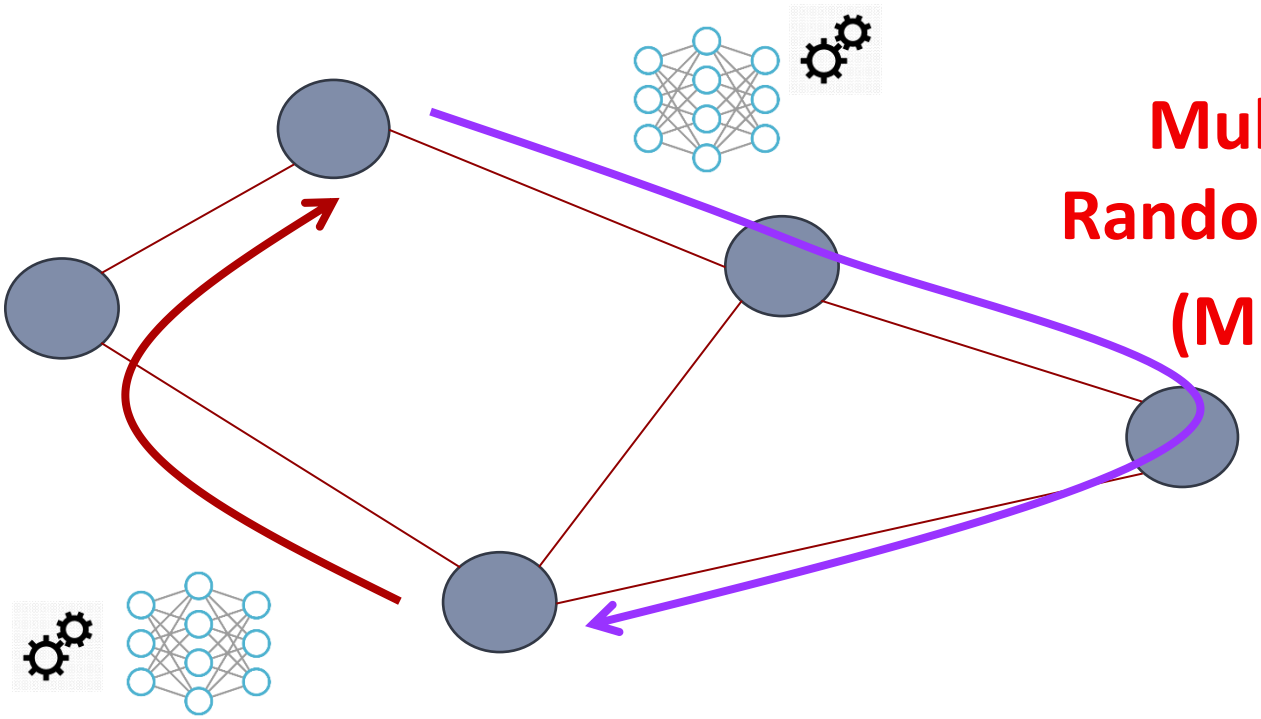
Serverless Federated Learning - RWS



Random-Walking Snake (RWS)

A. Ardic, H. Seferoglu, S. El Rouayheb, E. Koyuncu, “Random Walking Snakes for Decentralized Learning at Edge Networks,” in IEEE LANMAN, London, UK, July 2023. Major revision in IEEE TMC, 2026.

Serverless Federated Learning - MRW



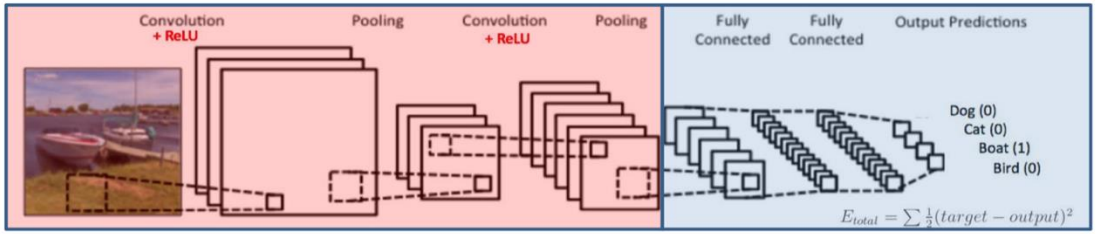
Multiple Random-Walk (MRW)

P. Gholami, H. Seferoglu, "A Tale of Two Learning Algorithms: Multiple Stream Random Walk and Asynchronous Gossip," in ACM MobiHoc, Tokyo, Japan, Nov. 2026.

Decentralized Learning - SRW

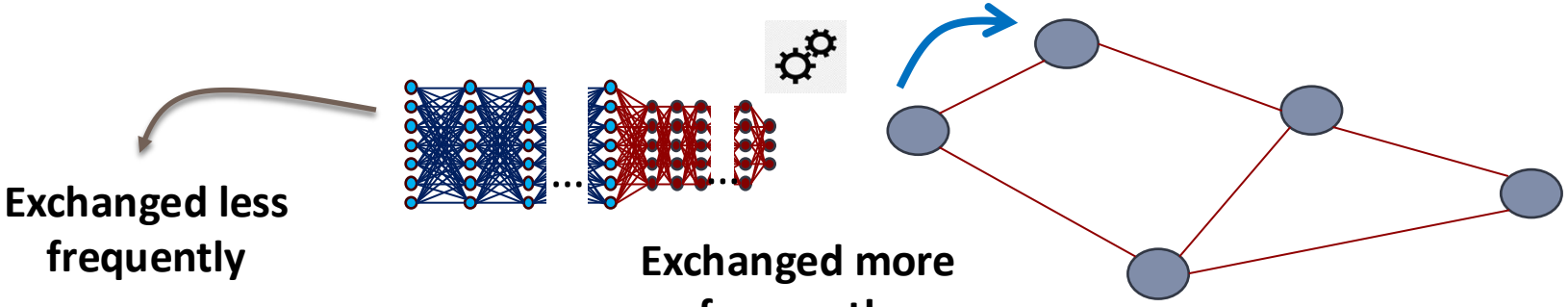
Split Random-Walk (SRW)

Idea: The initial layers (representation extractor) of the model tend to be more consistent across different clients in comparison to the final layers (head).



Learning the representation

Learning classification



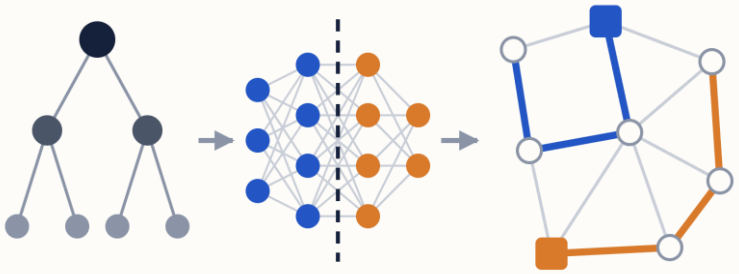
Exchanged less frequently

Exchanged more frequently

A. Feizbaksh, P. Gholami, H. Seferoglu, "Split Random Walk Learning", CODEC-FM, NeurIPS, 2026.

Ongoing/Future Directions

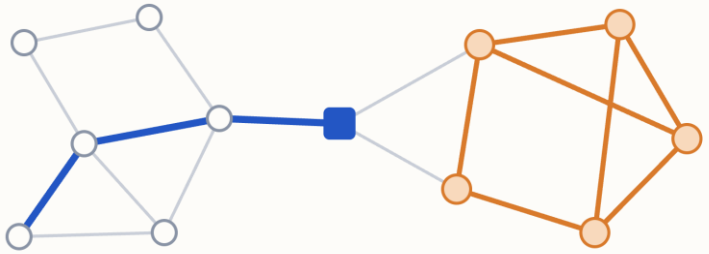
Ontology-Guided, Communication-Efficient Decentralized Learning



Ontology Split model Random walks

Split Random Walk

Splitting point of the model using ontology-based rules



Random walk Gossip

Hybrid Decentralized Algorithms

Recent Work

- Fast Inference with Model-Distribution and Parallelization
- Serverless Federated Learning
- Communication-Efficient and Heterogeneous Federated Learning
- Privacy-Preserving and Secure Distributed Learning

Our Prior Work:
Private & Secure Distributed AI/ML

Private coded computation and MPC

2021 IEEE ISIT
Adaptive coded matrix multiplication

2021 ICML Workshop
Coded privacy-preserving computation

2023 IEEE T-IFS
Efficient coded multi-party computation

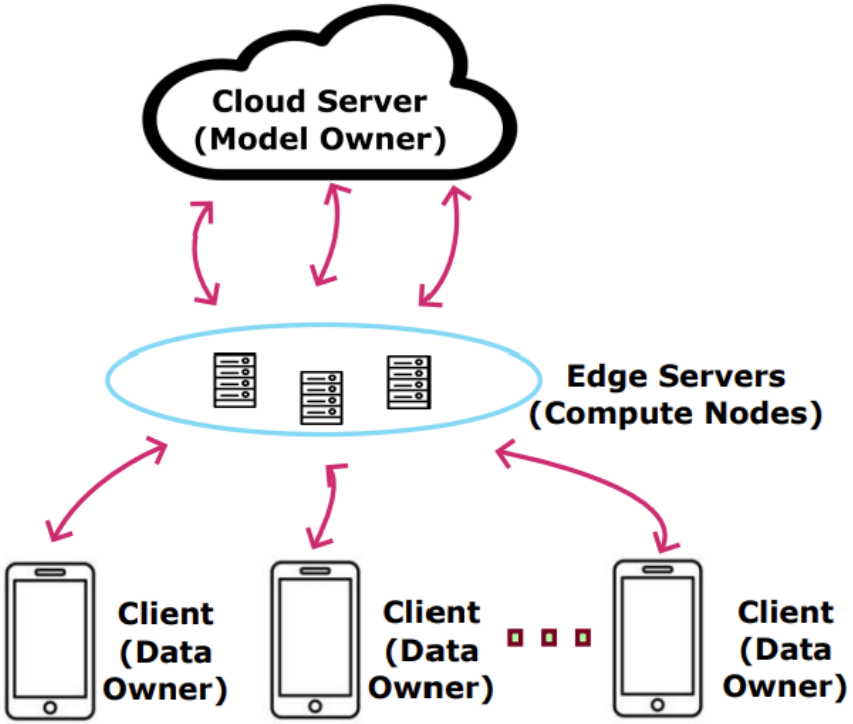
Private & verifiable AI/ML inference

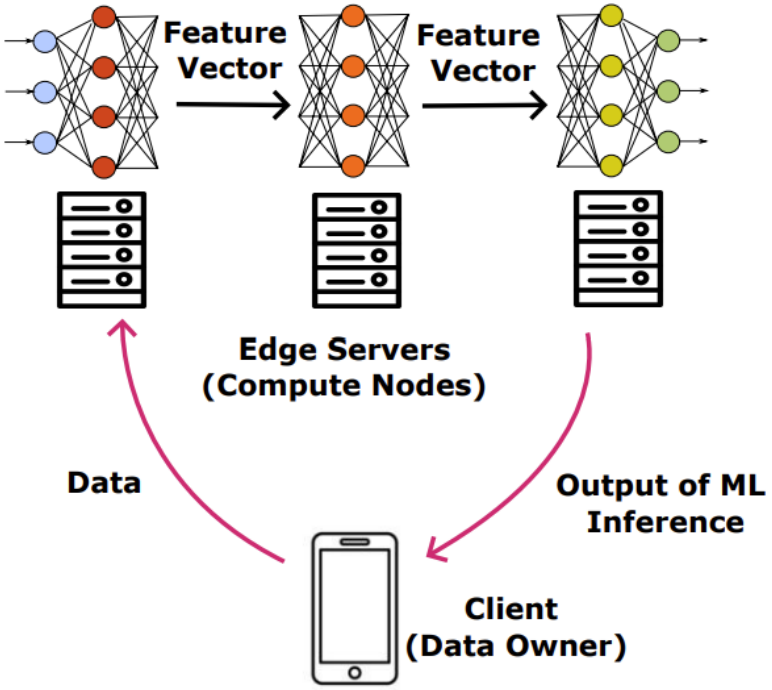
2024 IEEE CNS
Private hierarchical model-distributed inference

2025 Allerton
Private binary dot product via MPC

2026 ACM TAISAP
Private & verifiable matrix multiplication

Privacy-Preserving and Secure Distributed Learning

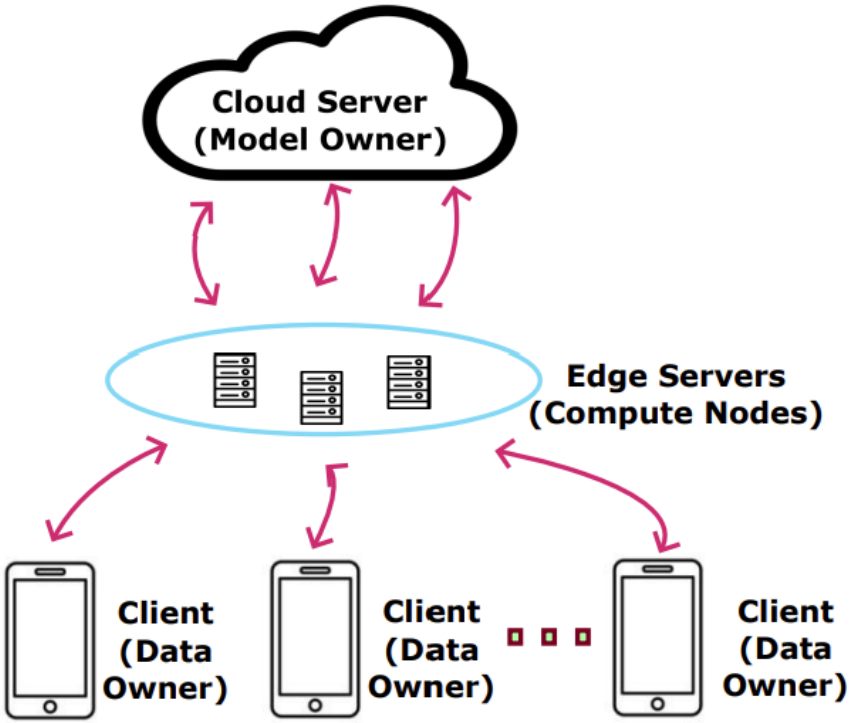




Q: How to provide privacy to both data and the model when model-distributed inference is used?

Our Solution: PrivateMDI

Privacy-Preserving Hierarchical MDI



Solution Overview

- Additive secret sharing, linear homomorphic encryption for linear operations and exchanging the keys.
- Garbled circuit and three-party oblivious transfer for non-linear operations.
- Offline (independent of the input data $\mathbf{x}$) and online phases.

F. Jafarian Dehkordi, Y. Keshtkarjahromi, H. Seferoglu, "Privacy-Preserving Model-Distributed Inference at the Edge," in IEEE CNS, Sep. 2024.

Questions?

https://sites.google.com/uic.edu/nrl/

hulya@uic.edu