



**SweWIN**  
Swedish Wireless Innovation Network

# Sustainable Semi-Decentralized Federated Learning with Stragglers

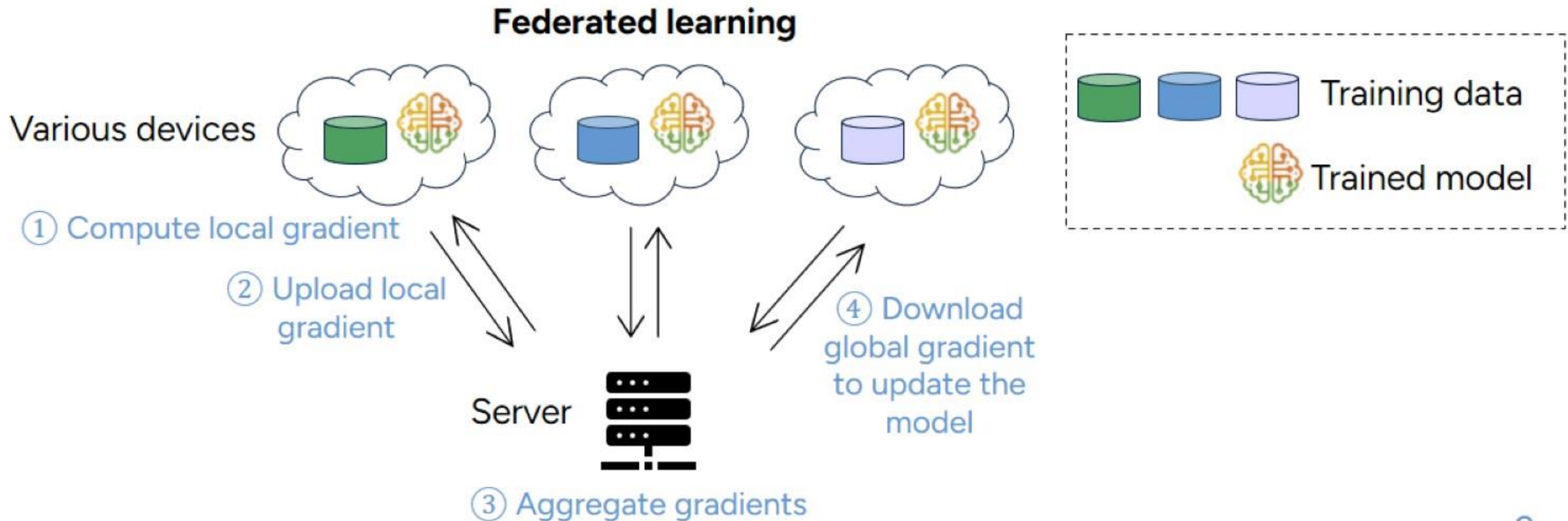
Chengxi Li, Postdoc researcher

Department of Information Science and Engineering (ISE), KTH

2026-04-14

# Background

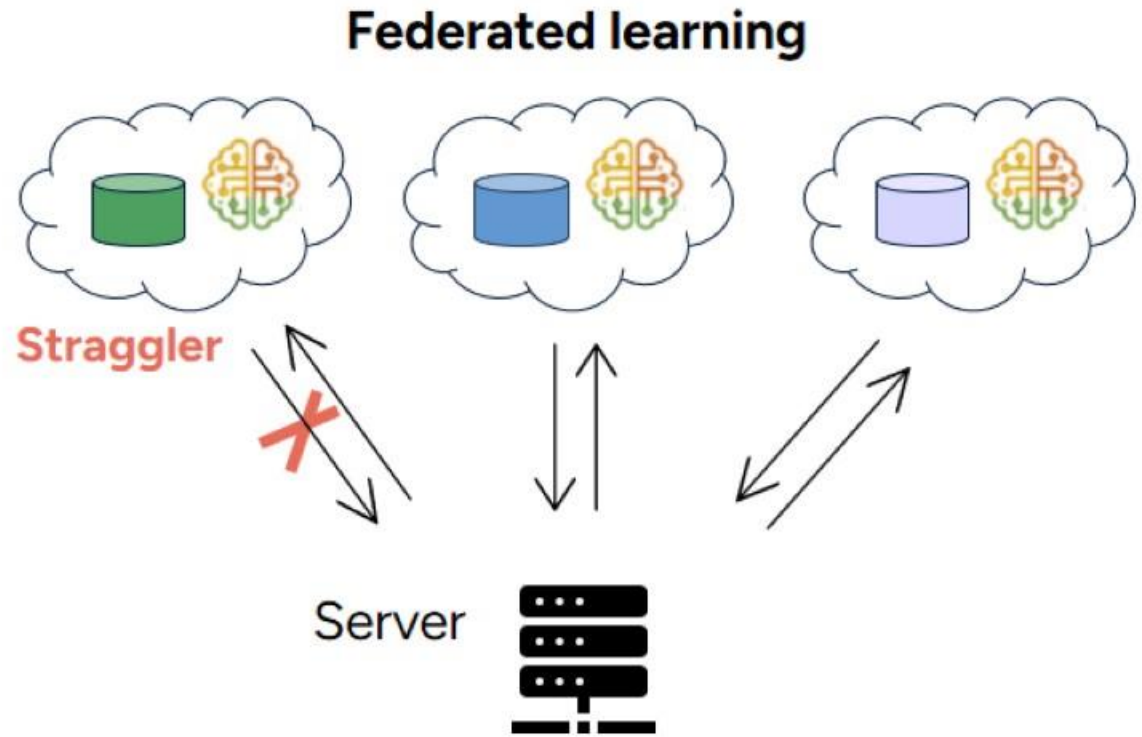
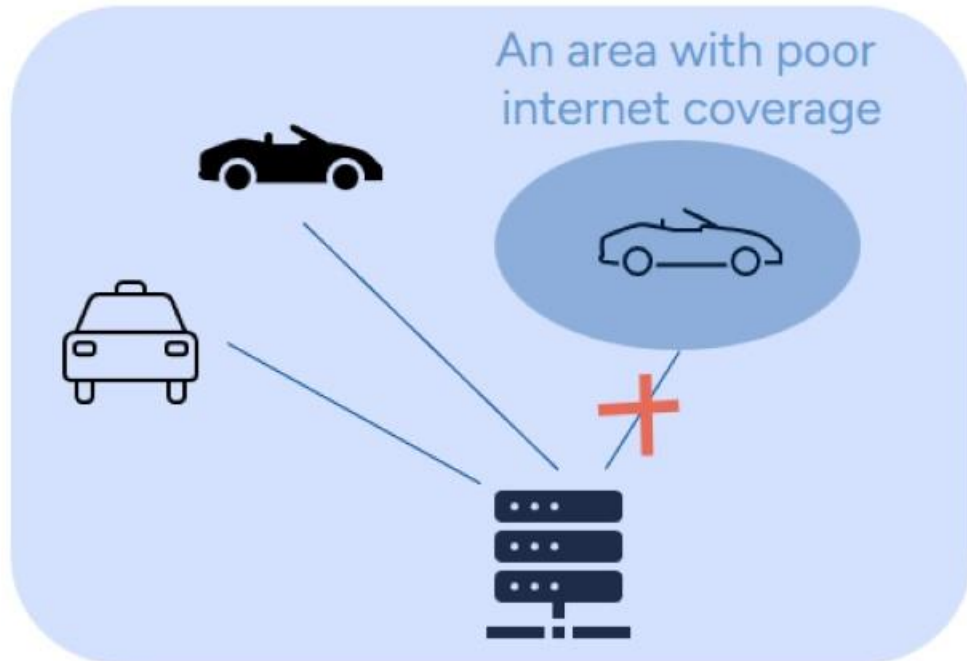
- Federated Learning (FL)
  - ✓ Training machine learning models using data owned by various devices
  - ✓ Privacy requirements: only share model parameters or model updates instead of raw data



# Background

- Challenges in FL
  - ✓ Communication stragglers: Devices may have intermittent connectivity to the server

A vehicle might traverse areas with poor internet coverage, leading to sporadic losses in server connectivity.

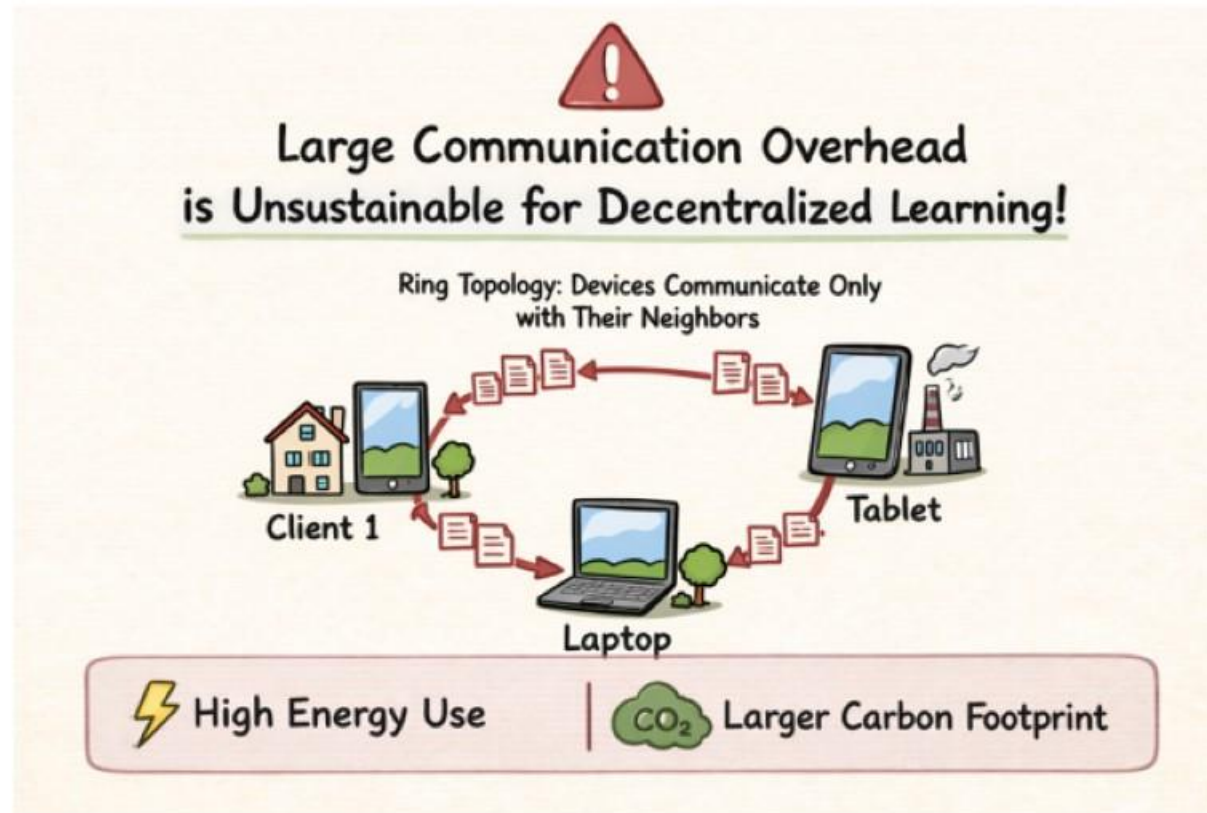


Missing information from the stragglers  
→ Degraded learning performance

# Background

Existing efforts to evade the stragglers in FL

- **Fully decentralized FL framework** [W. Liu, 2022; T. Vogels, 2022; Y. Lu, 2023]

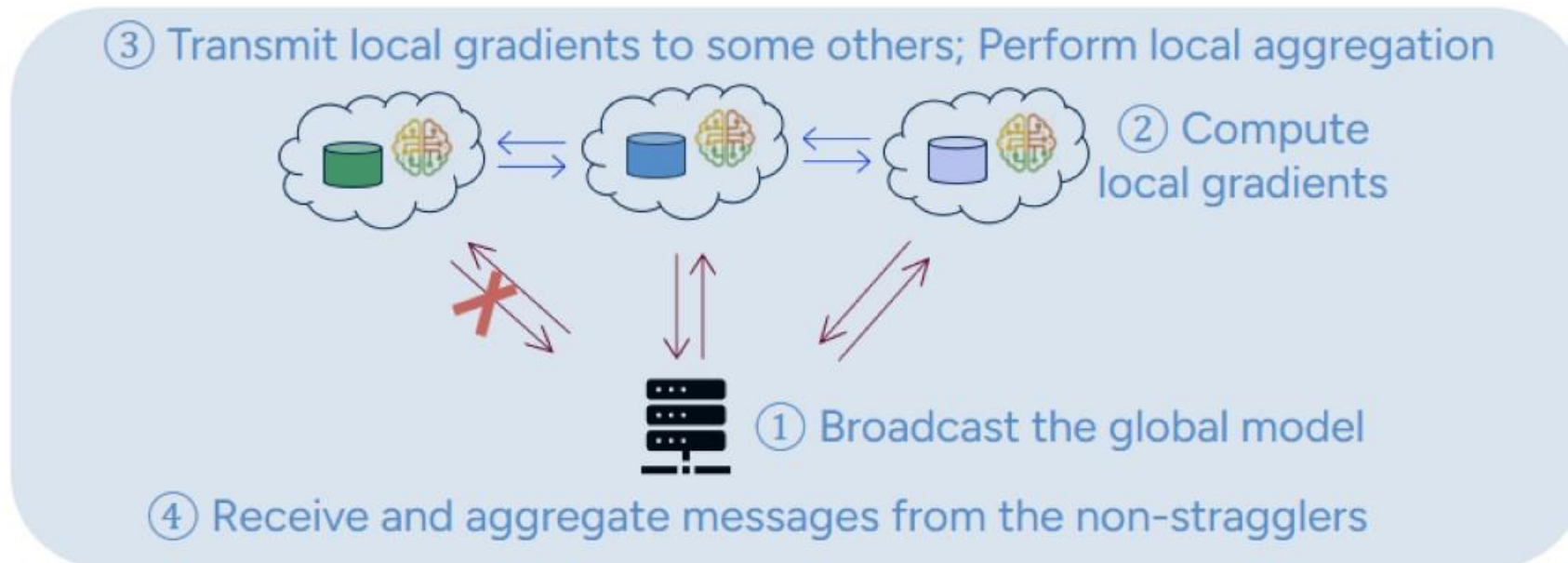


# Background

Existing efforts to evade the stragglers in FL

- **Semi-decentralized FL framework**

- A framework in between traditional FL and fully decentralized FL [Y. Guo, 2022; M. Yemini, 2023; H. Wang, 2024]
- It allows two types of communication {
  - Peer-to-peer communication among devices
  - Communication between devices and the server

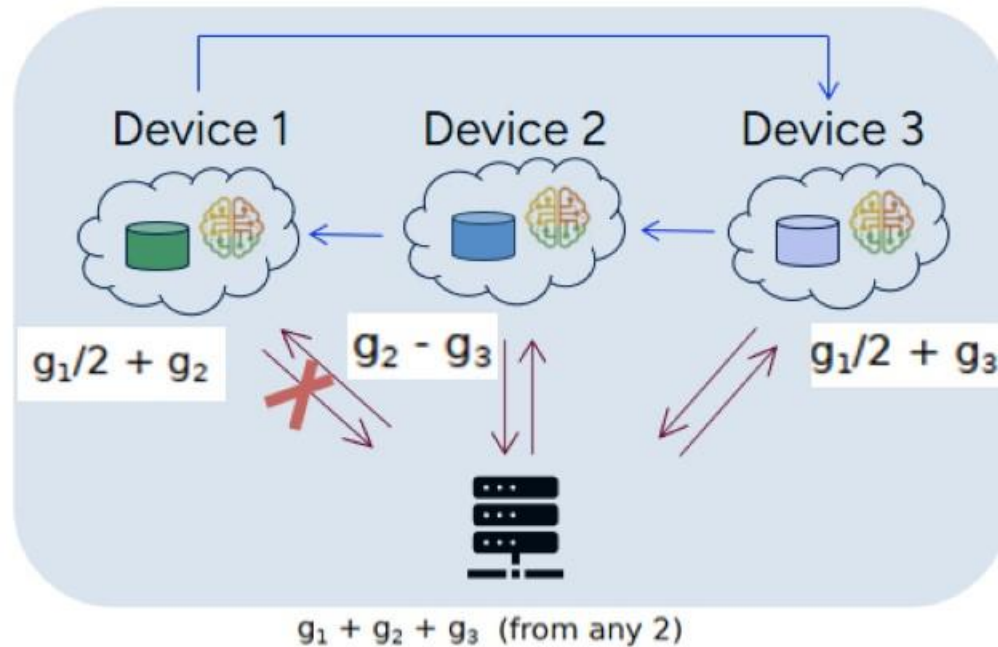


# Background

Existing efforts to evade the stragglers in FL

- **Semi-decentralized FL framework:**

To fully evade the negative impact of the stragglers--**Gradient coding**



The same performance can be attained as if there were no stragglers

s stragglers and  $N$  devices: Transmitting  $Ns$  vectors in each round → High communication overhead

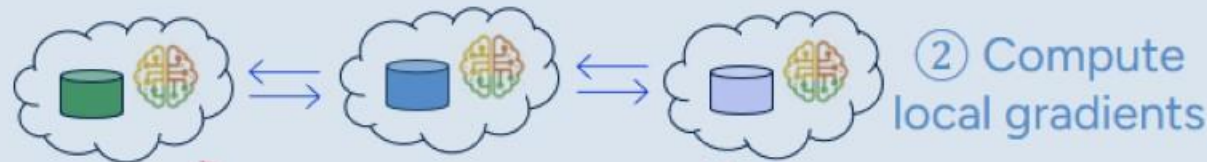
# Problem Model

- There are  $N$  devices and a central server. Each device owns a dataset  $\mathcal{D}_i, i = 1, \dots, N$ .
- Objective --- to learn a model parameter vector minimizing the total training loss

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \arg \min_{\mathbf{x} \in \mathbb{R}^d} \sum_{i=1}^N f_i(\mathbf{x})$$

- $f_i(\mathbf{x}): \mathbb{R}^d \rightarrow \mathbb{R}$  is the local loss function associated with  $\mathcal{D}_i: f_i(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}_i} F(\mathbf{x}, \xi)$

③ Transmit local gradients to some others; Perform local aggregation



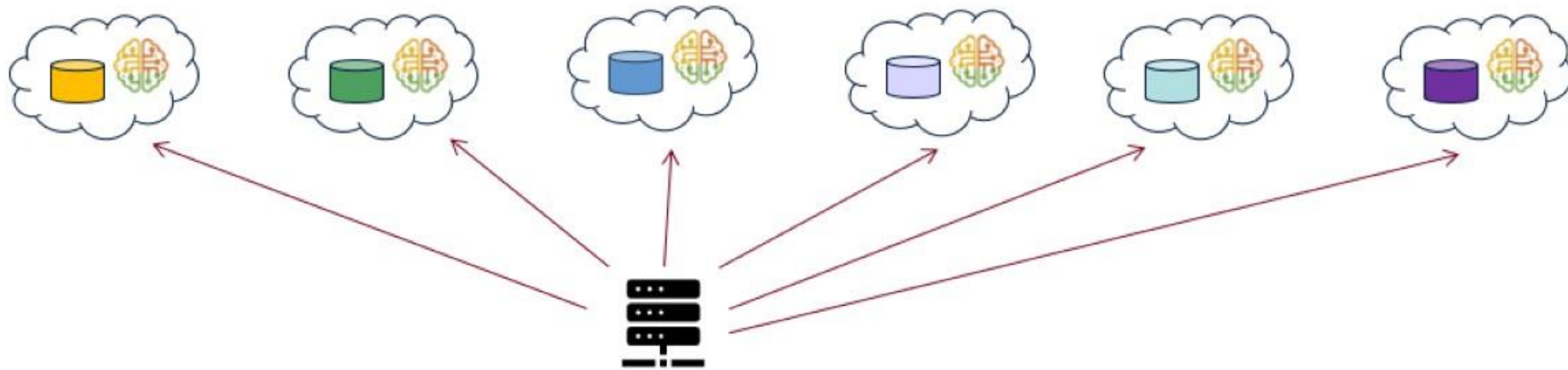
Intermittently available links -----  
 $I_i^t \sim \text{Bernoulli}(1 - p)$ , for device  $i$  in round  $t$

① Broadcast the global model

④ Receive and aggregate messages from the non-stragglers

# The proposed method: Communication-efficient semi-decentralized FL (COFFEE)

In each round  $t$ :



① Broadcast the global model  $\mathbf{x}^t$  to all devices

# The proposed method: COFFEE

② Each device  $i$  computes the stochastic gradient with the local batch  $\mathcal{D}_{i,t}$ :  $\mathbf{g}_i^t = \frac{1}{|\mathcal{D}_{i,t}|} \sum_{\zeta \in \mathcal{D}_{i,t}} \nabla F(\mathbf{x}^t, \zeta)$

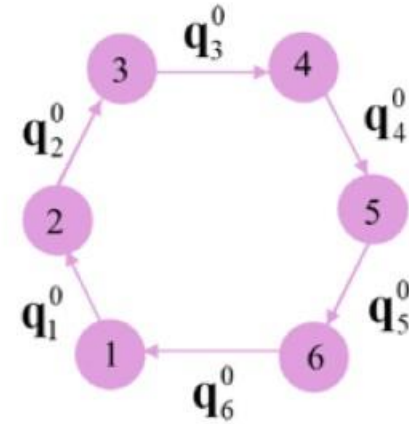


# The proposed method: COFFEE

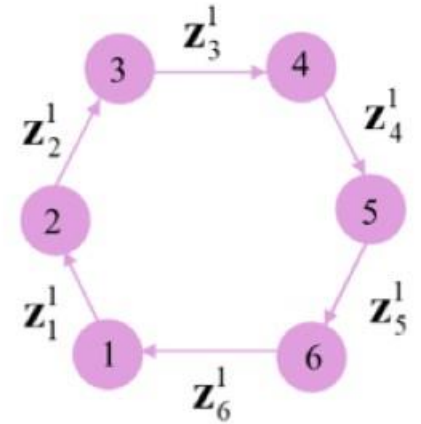
③ The devices exchange local gradients through a specified number of steps towards achieving **communication-optimal local consensus**, motivated by [L. Ding, 2023]

“**Communication-optimal**” means: Each device  $i$  obtains the average of the local gradients computed by itself and its  $n_R$  previous neighbors **under the minimal communication overhead**

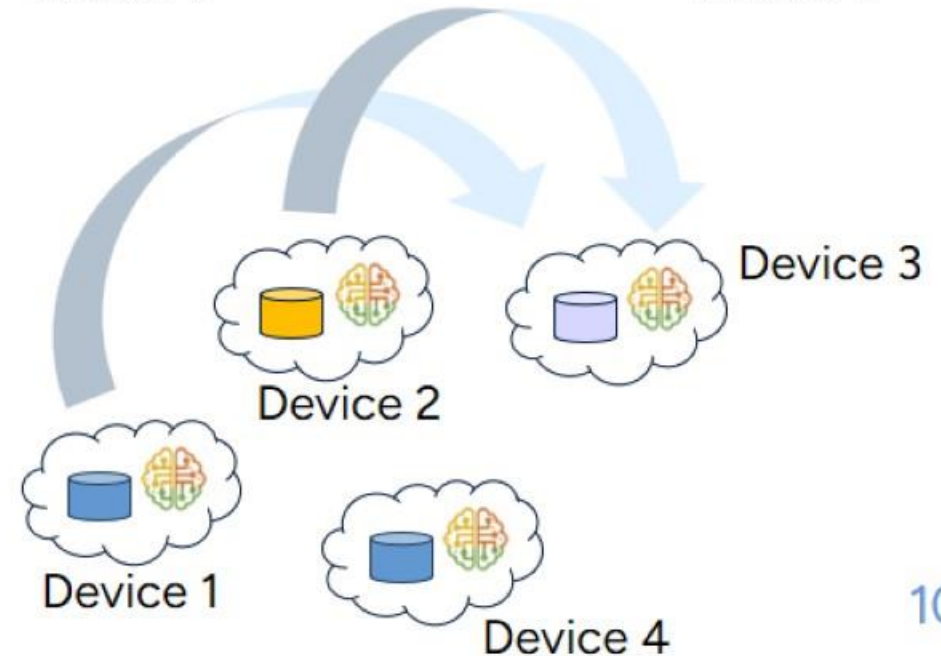
$$\mathbf{q}_i^R = \frac{1}{n_R + 1} \sum_{j=0}^{n_R} \mathbf{g}_{i-j}^t$$



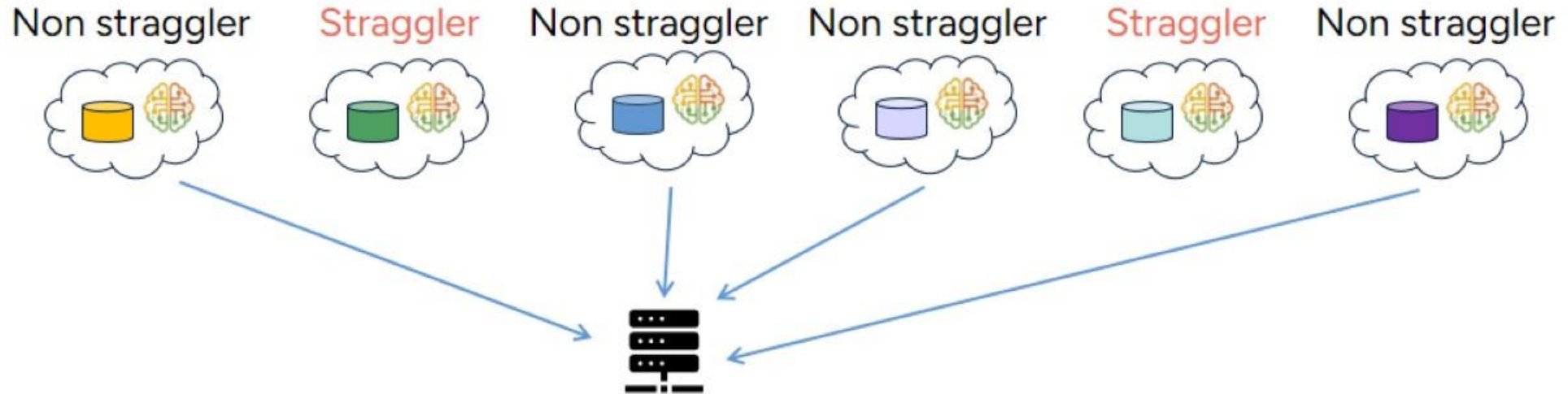
Iteration 0



Iteration 1

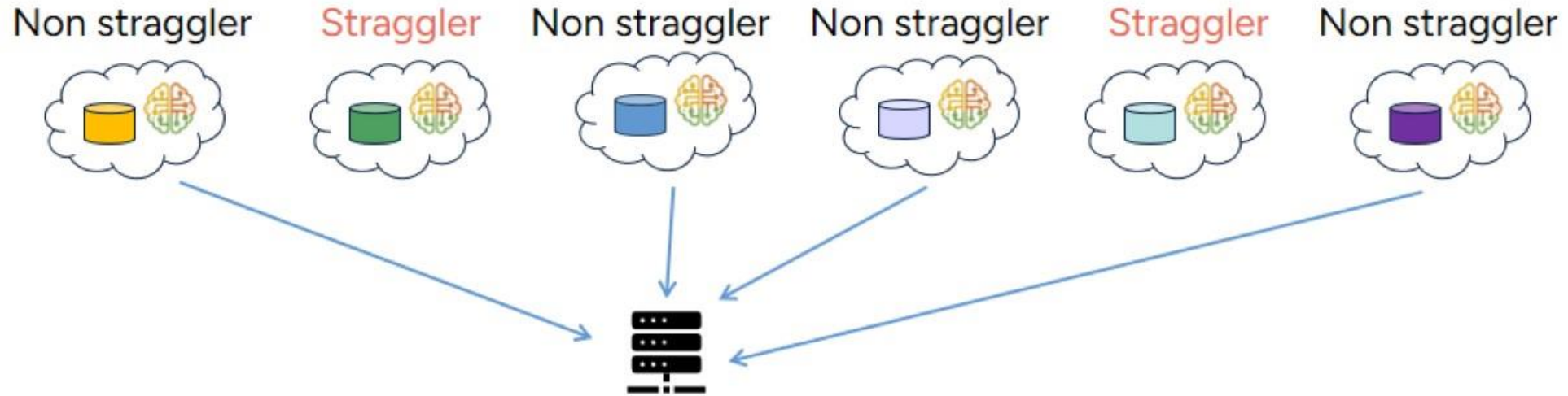


# The proposed method: COFFEE



④ Non-straggler device  $i$  transmits  $\mathbf{q}_i^R$  to the server. Stragglers send nothing.

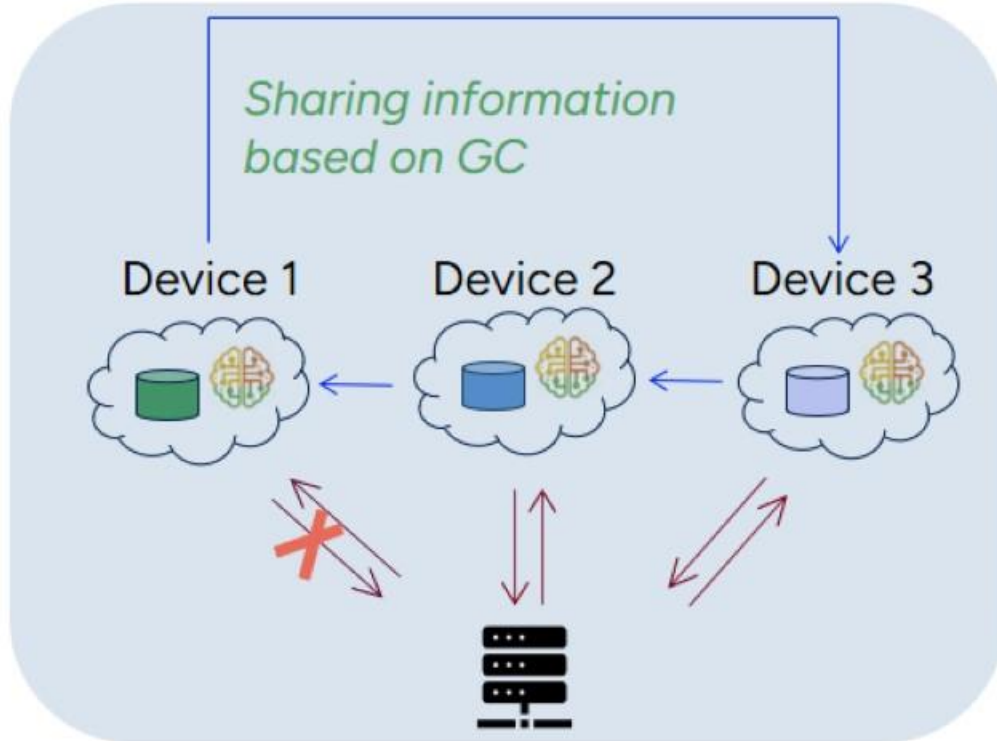
# The proposed method: COFFEE



⑤ The server aggregates the received messages and obtains the global gradient:  $\hat{\mathbf{g}}^t = \sum_{i=1}^N I_i^t \frac{1}{1-p} \mathbf{q}_i^R$   
 The server updates the global model:  $\mathbf{x}^{t+1} = \mathbf{x}^t - \gamma^t \hat{\mathbf{g}}^t$

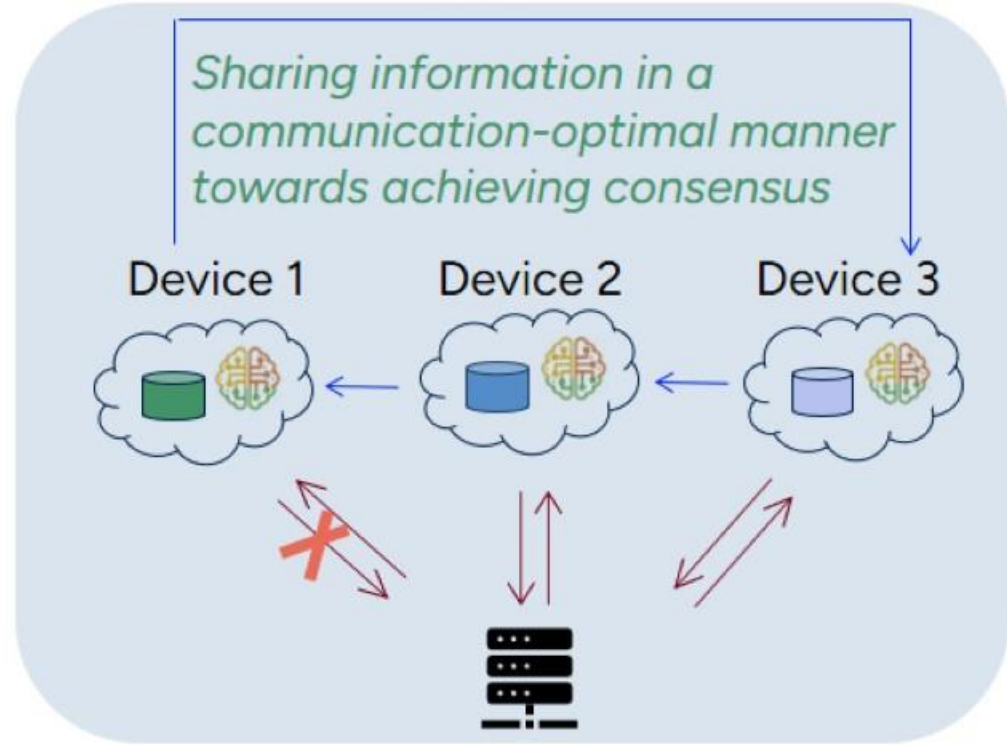
# Advantages of COFFEE: An intuitive comparison

## GC in Semi-decentralized FL



- The global gradient can be exactly recovered at the server.
- $s$  stragglers and  $N$  devices: Transmitting  $Ns$  vectors in each round → **High communication overhead**

## COFFEE



- The global gradient can be approximately recovered at the server.
- **Reduced communication overhead in each round**

# Theoretical analysis

**Assumption 1.** The stochastic gradients are unbiased with bounded variance:

$$\mathbb{E}_{\mathcal{D}_{i,t} \subset \mathcal{D}_i} (\mathbf{g}_i^t) = \nabla f_i(\mathbf{x}^t), \mathbb{E}_{\mathcal{D}_{i,t} \subset \mathcal{D}_i} \left[ \left\| \nabla f_i(\mathbf{x}^t) - \mathbf{g}_i^t \right\|_2^2 \right] \leq \sigma^2, \forall i$$

**Assumption 2.** The overall training loss is  $L$ -smooth:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2, \forall \mathbf{x}, \mathbf{y}$$

**Assumption 3.** The heterogeneity among the local training data is bounded:

$$\left\| \nabla f_i(\mathbf{x}) - \frac{1}{N} \nabla f(\mathbf{x}) \right\|_2^2 \leq \beta^2, \forall i, \forall \mathbf{x}$$

# Theoretical analysis

**Theorem 1** (Learning performance of COFFEE). *Based on Assumptions 1-3, by setting  $\gamma = \frac{\varepsilon}{\sqrt{T+1}}$  with  $\varepsilon > 0$ , for  $T > \left(\frac{pL}{(1-p)N} + \frac{L}{2}\right)^2 \varepsilon^2 - 1$ , COFFEE converges as*

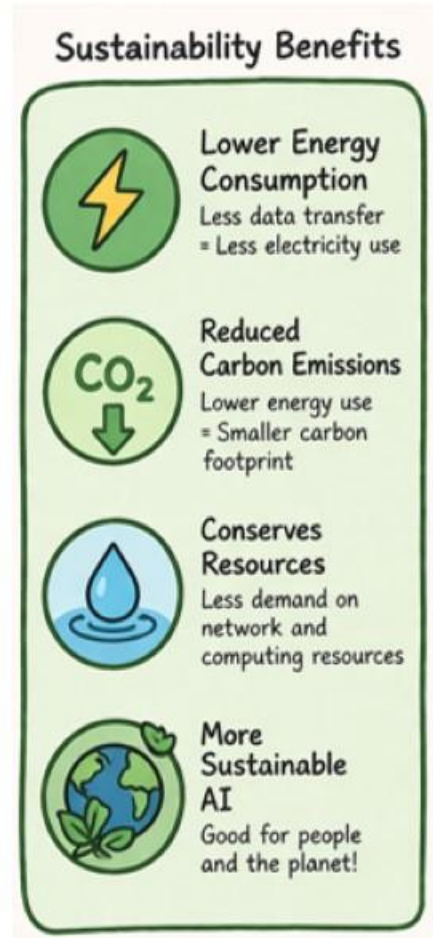
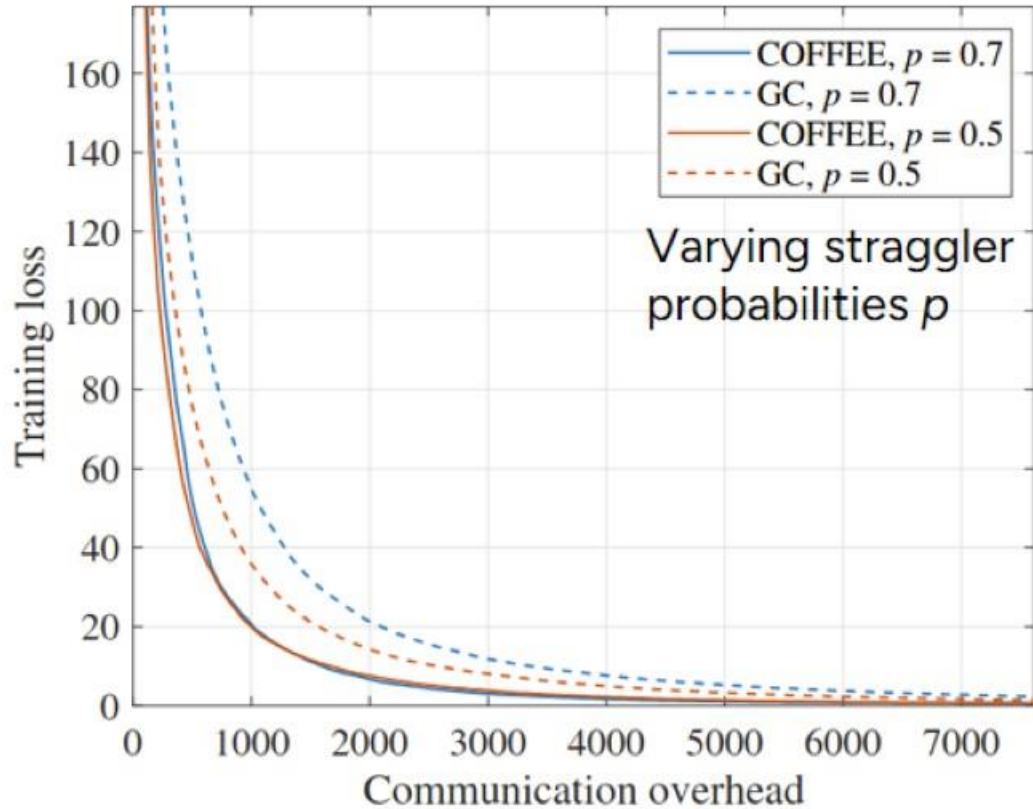
$$\begin{aligned} & \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left( \|\nabla f(\mathbf{x}^t)\|_2^2 \right) \\ & \leq \frac{f(\mathbf{x}^0) - f^*}{\varepsilon \sqrt{T+1} - \frac{\varepsilon^2 pL}{(1-p)N} - \frac{L}{2} \varepsilon^2} \\ & \quad + \frac{\varepsilon \left[ \frac{pLN\beta^2}{1-p} + \frac{pLN\sigma^2}{2(1-p)(n_R+1)} + \frac{L}{2} N\sigma^2 \right]}{\sqrt{T+1} - \frac{\varepsilon pL}{(1-p)N} - \frac{L}{2} \varepsilon}, \end{aligned}$$

where  $f^*$  is the minimum value of the overall training loss.

Decreasing value of  $p \rightarrow$   
enhanced learning performance

# Numerical results

- Logistic regression problem based on the MNIST dataset



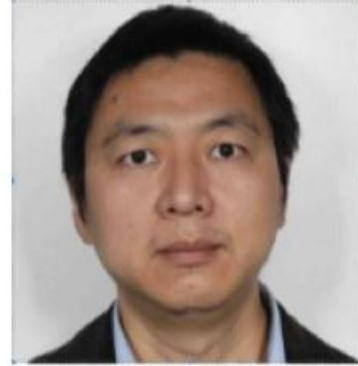
- COFFEE reduces the communication burden significantly compared with the baseline.
- Rationale: COFFEE adopts a **communication-efficient** way for information sharing among the devices

## Based on a joint work of



Chengxi Li

chengxli@kth.se



Ming Xiao



Mikael Skoglund



C. Li, M. Xiao and M. Skoglund, "Communication-Efficient Semi-Decentralized Federated Learning in the Presence of Stragglers," in *IEEE Transactions on Communications*, vol. 73, no. 12, pp. 13999-14013, Dec. 2025.



**Thank you for listening!**