Kang YAN, Nina SHU, Tao WU, Chunsheng LIU, Panlong YANG, 2024. A survey of energy-efficient strategies for federated learning in mobile edge computing. *Frontiers of Information Technology & Electronic Engineering*, 25(5):645-663. https://doi.org/10.1631/FITEE.2300181

# A survey of energy-efficient strategies for federated learning in mobile edge computing



**Key words:** Mobile edge computing; Federated learning; Energy-efficient

Corresponding author: Tao WU
E-mail: wutao20@nudt.edu.cn
ORCID: https://orcid.org/0000-0003-1344-835X

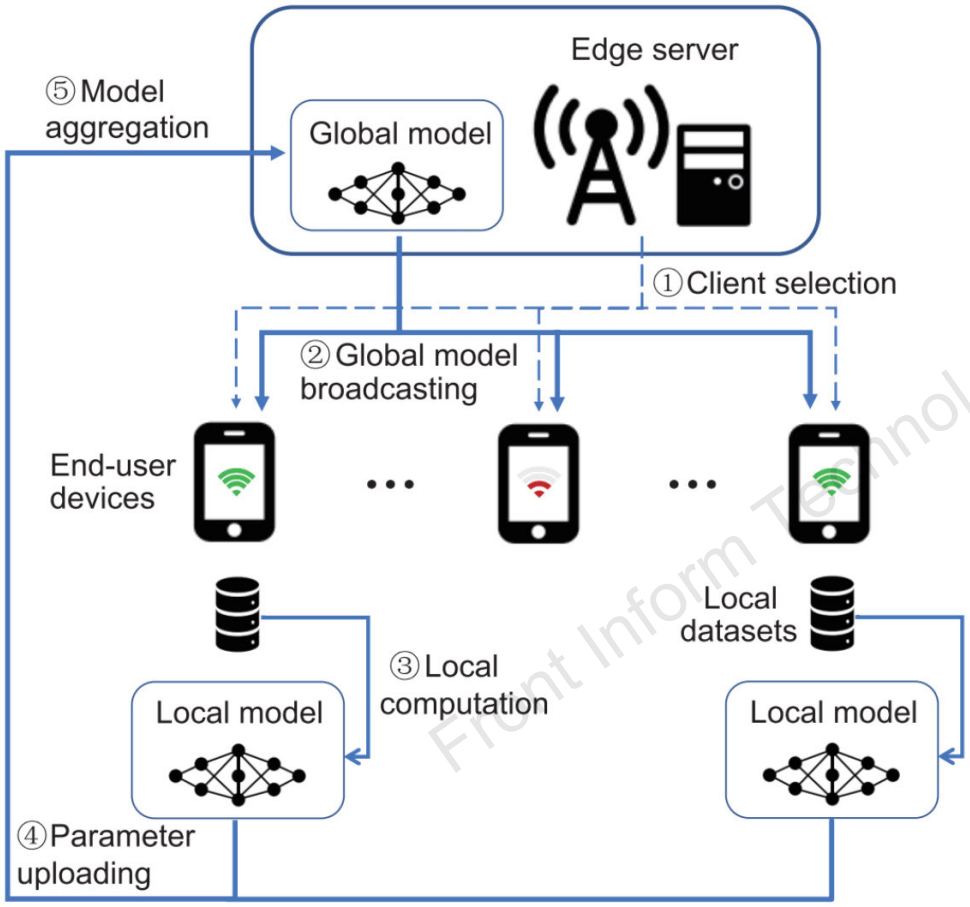
# Motivation

- Recently, federated learning (FL) has garnered attention for enabling edge devices to collaborate on training machine learning (ML) models without sharing raw data. With its efficiency, privacy preservation, and scalability advantages, FL is a promising ML training framework to realize edge artificial intelligence (AI).

- Given that end-user devices (EDs) are typically powered by batteries with limited capacity, it is challenging to execute energy-intensive FL tasks. To address this challenge, many strategies have been proposed to improve the energy efficiency of FL; however, there lacks a comprehensive survey and categorization of these strategies.

- We provide a comprehensive survey of recent advances in energy-efficient strategies for FL in mobile edge computing (MEC) including system models, challenges in improving energy efficiency, existing strategies, and future prospects.

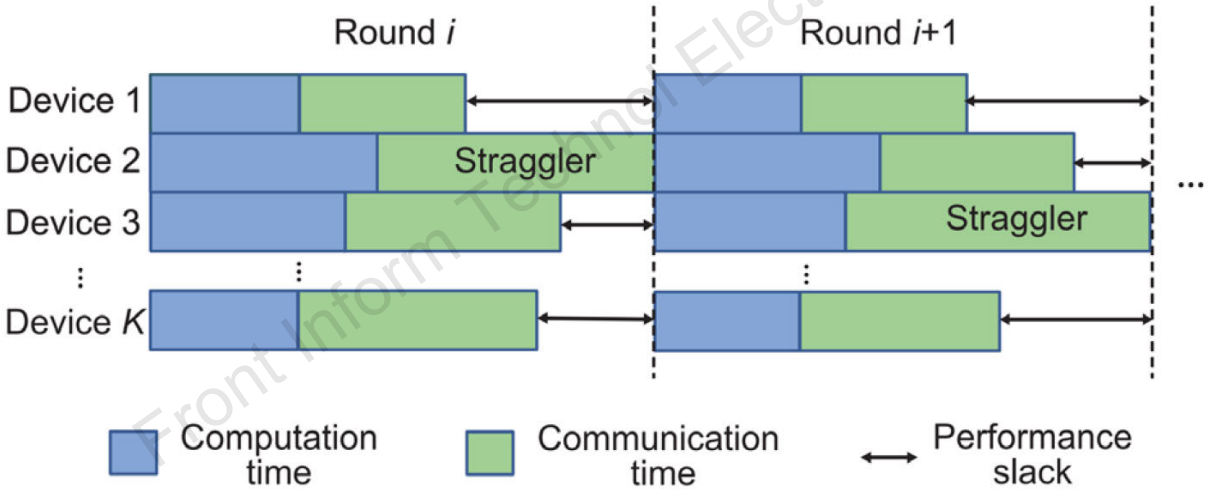
# Overview of the FL system




As shown in the figure, an FL system usually consists of a central unit (model owner), e.g., the edge server, and a group of EDs (data owners), e.g., mobile phones. The energy consumption at the client stems mainly from the computation cost of training the local model in step ③ and the communication cost of uploading parameters in step ④.

# Challenges in improving energy efficiency

- The heterogeneity in computing and communication capabilities among clients leads to the straggler problem. The presence of stragglers degrades the performance of FL because all devices have to wait for the slowest device, resulting in an inefficient use of time and energy resources.




- In FL, data on different devices may be non-independent and identically distributed (non-IID), and the quantity of data may vary. This data heterogeneity necessitates more rounds to achieve a satisfying accuracy, resulting in increased energy consumption.

# Challenges in improving energy efficiency

- ED computing and communication are highly dynamic due to the stochastic execution environment. Contention for computing resources or unstable network connections can reduce the efficiency of FL, leading to higher energy consumption.

- With the development of deep learning and the increasing precision requirements of intelligent applications, the complexity of ML models has significantly increased. On one hand, highly complex computational tasks lead to high local computational energy consumption on EDs. On the other hand, the continuously increasing number of model parameters significantly increases communication loads and energy consumption.

# Summary of energy-efficient strategies

**Table 2 Classification of energy efficiency optimization strategies**

| Strategy | Scheme | Reference | Advantage | Disadvantage |
|---|---|---|---|---|
| Learning-based strategies | Model compression | Abdelmoniem and Canini, 2021; Li et al., 2021; Prakash et al., 2022; Chen R et al., 2023 | Significantly reducing computation and communication costs by reducing the complexity of the model | May result in compression errors and impact performance accuracy |
| | Hyperparameter optimization | Luo et al., 2021; Prakash et al., 2022; Shi et al., 2022b; Sun et al., 2024 | Improving the model's accuracy, convergence speed, and generalization capability while enhancing resource utilization efficiency | Most of the existing works lack consideration for device heterogeneity |
| | Training algorithm improvement | Albaseer et al., 2021; Nguyen et al., 2021 | Improving the quality of model updates and accelerating model convergence | Introducing additional computation overhead to the training process |
| Resource allocation based strategies | Computation resource allocation | Li et al., 2019; Zhan et al., 2020; Kim J et al., 2022 | Increasing the utilization of computation resources and reducing training costs | 1. Most existing strategies are based on static models and assumptions, which may not effectively adapt to the dynamic energy consumption requirements in real-world environments.<br>2. Optimizing the allocation of resources requires additional data transmission and sharing for the participating parties, which may introduce potential risks of privacy breaches and security vulnerabilities.<br>3. Most resource allocation strategies assume an ideal interference-free environment, but in reality, interference is inevitable, which can undermine their effectiveness. |
| | Communication resource allocation | Zeng QS et al., 2020; Hu et al., 2022 | Increasing the utilization of communication resources and reducing communication costs | |
| | Joint $C^2$ resource allocation | Tran NH et al., 2019; Mo and Xu, 2021; Yang ZH et al., 2021; Zeng QS et al., 2021a; Battiloro et al., 2023 | Increasing the utilization of communication and computation resources and reducing overall costs | |
| Client selection strategies | Direct energy-efficient client selection | Li et al., 2020; Zeng QS et al., 2020; Kim YG and Wu, 2021; Zheng et al., 2021; Albelaihi et al., 2022; Arouj and Abdelmoniem, 2022; Peng et al., 2023; Wu T et al., 2023 | Reducing overall energy consumption by considering the heterogeneity among devices and deliberately selecting devices with better performance to participate in training | 1. Bias in device selection may limit training data diversity and hinder the model's ability to generalize to diverse scenarios.<br>2. The majority of client selection strategies have failed to consider the dynamic characteristics of clients, such as their battery status and concurrent applications.<br>3. Client combinations and their impact on federated learning performance are disregarded in most strategies. |
| | Indirect energy-efficient client selection | Cho et al., 2020; Xu and Wang, 2021; Perazzone et al., 2022; Tang et al., 2022; Zhao JX et al., 2022 | Accelerating model convergence and indirectly reducing overall energy consumption by selecting devices that contribute more significantly to model updates to participate in training | |

# Future directions

- Leveraging hardware advancements. Hardware research in FL can focus on reducing energy consumption by designing efficient EDs and optimizing computing architectures through low-power components and specialized accelerators for distributed learning tasks.

- Enhancing communication technologies. Enhancing communication technologies for FL involves developing efficient protocols tailored for EDs and leveraging advancements such as non-orthogonal multiple access (NOMA) and multiple-input multiple-output (MIMO) to improve spectrum efficiency and energy efficiency in wireless communication systems.

- Integration of AI methods. Future research in FL can integrate advanced AI methods like model compression, meta-learning, and reinforcement learning to tailor compressed models, facilitate knowledge transfer among nodes, and enable intelligent decision-making for optimized resource allocation, thus reducing energy consumption.

# Summary

In this paper, we concentrate on the energy consumption optimization problem of FL in MEC. First, we introduce the system model and the energy consumption models, encompassing both the computational and the communication energy consumption models. Subsequently, we provide a comprehensive overview of the primary challenges in terms of improving energy efficiency. We categorize the existing energy-efficient strategies for FL into three main categories: learning-based strategies, resource allocation strategies, and client selection strategies. For each category, we provide detailed introductions and summaries of the strategies. Finally, we discuss potential research directions for achieving energy-efficient FL.

Tao WU received the BS degree in automation from Hohai University, China in 2013, the MS degree in measurement and sensor technology from PLA University of Science and Technology, China in 2016, and the PhD degree in computer science and engineering from the Army Engineering University of PLA, China in 2019. He is currently an associate professor with the National University of Defense Technology and a postdoctoral researcher in the Hong Kong Polytechnic University, China. His research interests include wireless sensor networks, wireless charging, unmanned aerial vehicles, and edge computing.