







# ESR-MHFL: Edge Server Reallocation for Multi-Hierarchical Federated Learning

Tianao Xiang , Graduate Student Member, IEEE, Yuanguo Bi , Member, IEEE, Lin Cai , Fellow, IEEE, Chong Yu , Member, IEEE, Mingjian Zhi, Rongfei Zeng , Member, IEEE, and Tom H. Luan , Senior Member, IEEE

**Abstract**—Federated Learning (FL) enables efficient and privacy-preserving Edge Intelligence (EI) in Mobile Edge Computing (MEC). However, implementing FL-enabled EI services faces critical challenges, including data and device heterogeneity, limited network resources, uneven distribution of network infrastructure, etc., which may intensify with increasing system scale. These challenges are particularly acute in multi-provider environments where edge servers are suboptimally allocated across federations, leading to degraded convergence and increased training costs. In this article, we present a novel Multiple Hierarchical Federated Learning (MHFL) architecture for large-scale FL and design an Edge Server Reallocation scheme (ESR-MHFL) to enhance training efficiency by optimally redistributing edge servers among federations based on their contribution to model convergence. We first develop a closed-form analysis model for MHFL to quantify training time, computation, and communication costs. To improve training efficiency, we analyze the impacts of edge server allocation on convergence and formulate server reallocation as a multi-item auction problem with theoretical guarantees. We then propose ESR-MHFL, which leverages Coalition Structure Generation (CSG) and greedy matching methods to simplify the reallocation problem and enhance efficiency. Extensive numerical simulations demonstrate that ESR-MHFL not only improves model accuracy while reducing training cost but also exhibits strong compatibility with existing client selection methods, achieving improved training efficiency. The total economic expenditure combining all components.

**Index Terms**—Hierarchical federated learning (FL), edge server reallocation, mobile edge computing (MEC), training efficiency.

Received 20 February 2025; revised 27 August 2025; accepted 28 August 2025. Date of publication 4 September 2025; date of current version 9 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62471121, in part by the Fundamental Research Funds for the Central Universities of China under Grant N2424010-18, in part by Shenyang Science and Technology Plan Fund Project under Grant 23-503-6-17, in part by Liaoning Provincial Science and Technology Plan Project under Grant 2023JH2/101700370, and in part by the Fund of National Key Laboratory of Metallurgical Intelligent Manufacturing System. (Corresponding author: Yuanguo Bi.)

Tianao Xiang, Yuanguo Bi, and Mingjian Zhi are with the School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China, and also with the Engineering Research Center of Security Technology of Complex Network System, Ministry of Education, Shenyang 110169, China (e-mail: 2010652@stu.neu.edu.cn; biyuanguo@mail.neu.edu.cn; 2110657@stu.neu.edu.cn).

Lin Cai is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P 5C2, Canada (e-mail: cai@ece.uvic.ca).

Chong Yu is with the Department of Computer Science, University of Cincinnati, Cincinnati, OH 45221 USA (e-mail: yuc5@ucmail.uc.edu).

Rongfei Zeng is with the College of Software, Northeastern University, Shenyang 110819, China (e-mail: zengrf@swc.neu.edu.cn).

Tom H. Luan is with the School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: tom.luan@xjtu.edu.cn).

Digital Object Identifier 10.1109/TSC.2025.3606219

## I. INTRODUCTION

THE integration of Artificial Intelligence (AI) with Mobile Edge Computing (MEC) has spurred the development of Edge Intelligence (EI), where paradigms like Integrated Sensing, Computation, and Communication (ISCC) enable sophisticated on-device AI [1], [2]. Within this evolving landscape, where numerous devices need to learn collaboratively, Federated Learning (FL) has emerged as a critical distributed Machine Learning (ML) approach that fundamentally transforms how AI models are trained and deployed.

Unlike traditional centralized training methods, FL empowers geographically distributed User Equipment (UEs) to collaboratively train ML models using their local data [3], [4], [5]. This distributed approach has catalyzed innovations across diverse domains, including enhancing digital twins in the Internet of Things (IoT) [6], detecting trajectory anomalies [7], wireless video caching [8], and advanced object detection in autonomous driving systems [9].

While these FL applications demonstrate significant potential, their real-world deployment faces fundamental scalability challenges as participating devices grow into thousands or millions. Traditional single-federation approaches encounter bottlenecks in communication overhead, convergence speed, and resource utilization efficiency. To address this, hierarchical FL has been developed with cloud-level aggregators coordinating multiple edge-level aggregators [10]. However, when being scaled to massive multi-provider environments with heterogeneous data and limited resources, these hierarchical federations struggle with convergence due to communication failures, extended training periods, and reduced accuracy [11], [12], [13].

Though recent theoretical advances in FL have shown progress, a critical gap exists in practical multi-provider deployments. Real-world FL implementations in 5 G/6 G networks [14], [15], industrial IoT [16], and healthcare systems [17] reveal substantial challenges including cross-operator coordination complexity, resource heterogeneity, and economic inefficiency that existing theoretical frameworks cannot address. These deployment issues raise the urgent need for practical FL solutions that bridge academic advances with industry requirements, specifically addressing multi-provider resource optimization challenges.

In practice, diverse geographical distribution and varying infrastructure ownership create multiple coexisting federations within a single area, leading to two fundamental problems:

(1) edge servers allocated to federations where they contribute minimally to model convergence, and (2) resulting data heterogeneity that degrades overall system performance. This multi-provider environment exacerbates challenges through uneven UE distribution across networks, limited client pools within edge server coverage areas, increased client selection complexity, and privacy risks from client information collection.

To address these challenges, we design ESR-MHFL, an Edge Server Reallocation scheme in Multi-Hierarchical FL that enhances training efficiency for large-scale multi-provider services while preserve privacy of UEs by the architectural design. ESR-MHFL operates at the federation level, coordinating cloud and edge servers while remaining agnostic to specific edge-UE privacy mechanisms. ESR-MHFL enables efficient cross-federation collaboration through intelligent edge server reallocation while avoiding the leakage of UEs' privacy.

The main contributions of this paper are as follows.

- 1) A multi-hierarchical federation architecture is proposed, and an analytical model is developed to analyze training time, computation cost, and communication cost which quantifies the impacts of edge server assignments on FL convergence for further design of edge server reallocation scheme.
- 2) An edge server reallocation scheme ESR-MHFL is proposed, which formulates the reallocation problem as a multi-item auction with theoretical guarantees. Moreover, ESR-MHFL employs a value approximation framework and Coalition Structure Generation (CSG) to reduce complexity and improve decision-making efficiency.
- 3) A Vickrey-Clarke-Groves (VCG)-based payment rule is designed to ensure individual rationality and incentive compatibility. Furthermore, an efficient greedy matching algorithm is proposed to make ESR-MHFL scalable in large-scale edge server reallocation.
- 4) Extensive simulations and comprehensive numerical analysis demonstrate that ESR-MHFL enhances training efficiency while reducing reallocation complexity. In addition, the results show that ESR-MHFL achieves significant improvements in model accuracy and resource utilization across different client selection methods.

The remainder of this paper is organized as follows. Section II reviews the related literature. Section III introduces the system model and analyzes the limitations of the current single federation approaches. Section IV formulates the edge server reallocation problem as a multi-item auction, designs the corresponding VCG-based payment rule, and then presents a CSG-based edge server reallocation scheme. Section V demonstrates and analyzes the numerical results. Finally, the conclusion is drawn in Section VI.

## II. RELATED WORK

In FL, extensive research has been dedicated to developing resource management schemes, addressing practical challenges and economic incentives. We provide a comprehensive taxonomy of existing research by dividing them into two categories:

client selection with programming techniques and incentive mechanisms with game theory.

Most existing client selection solutions with optimization techniques allocate resources by collecting information from UEs and network performance, improving efficiency in FL and resource utilization. In [18], Yu et al. analyzed the client selection feature in the FL training process and proposed ELASTIC, an energy and latency-aware resource management and client selection algorithm, which effectively addresses the joint problem of client selection and resource management. In [19], Chen et al. analyzed communication resource utilization in wireless FL and presented a joint learning and communication framework aiming to maximize the utilization of limited communication resources. In [20], Xu et al. addressed client selection and bandwidth allocation problem in energy-constrained situation. They proposed a client selection algorithm that relies solely on in-time information to ensure long-term efficiency in FL. In [21], Huang et al. proposed a three-layer hierarchical incentive framework for FL, which addresses information asymmetry between workers and local model owners to improve the FL model performance.

In MEC networks, there are extensive solutions of incentive mechanisms for FL, particularly utilizing contract theory, auctions and other approaches. In [22], Ng et al. introduced a serverless hierarchical FL framework within a two-layer system architecture and presented a reputation-aware hedonic coalition formation game to enhance the sustainable FL efficiency. In [23], Wang et al. proposed a clustered VFL approach and designed optimal contract theory-based incentive mechanisms to motivate clusters for intra-cluster iterations while maximizing server utility. In [24], Tang et al. proposed CPMARL-AFL that groups model users into clusters with dedicated agents to improve efficiency and personalization in auction-based FL. In [25], Li et al. proposed a dual-identity double auction for personalized FL, which enables participants to act as both model trainers and users, using reinforcement learning for model selection to optimize personal model performance.

In addition, there are also existing works focusing on reducing training cost and improve FL process training efficiency. In [26], Cui et al. proposed an adaptive compression framework for FL that dynamically adjusts compression rates to optimize the tradeoff between communication overhead and model accuracy under non-convex loss conditions. In [27], Oh et al. proposed a communication-efficient FL framework, and it uses vector quantized compressed sensing that combines dimensionality reduction and vector quantization, which reduces transmission overhead while maintaining model accuracy. In [28], Zhang et al. proposed SparsiFL, a graph sparsification-based secure aggregation protocol for FL, which reduces communication and computational overhead while maintaining correctness and privacy through uncertain graph sparsification.

However, these theoretical advances face critical limitations, including a lack of comprehensive economic optimization frameworks, the absence of multi-provider resource allocation mechanisms, and insufficient consideration of real-world deployment constraints. Previous research manages UEs within single federations, overlooking suboptimal edge server allocation across multiple federations, which degrades convergence

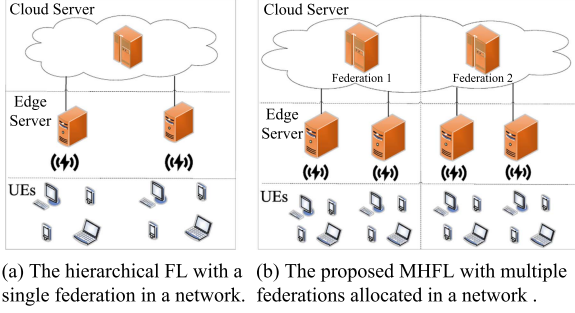


Fig. 1. The architectures of hierarchical FL and MHFL.

and increases training costs in large-scale multi-provider environments.

Unlike previous theoretical-focused works, ESR-MHFL bridges FL advances with practical deployment needs. ESR-MHFL uniquely addresses real-world industry challenges in multi-operator networks, smart city consortiums, and healthcare alliances through model similarity-based prediction and lightweight pre-training, providing the first comprehensive solution that combines theoretical rigor with practical applicability for real-world multi-provider FL deployments.

### III. PRELIMINARIES

In this section, we introduce the proposed MHFL, a novel architecture that effectively models the complex MHFL scenarios in MEC. Subsequently, we establish a cost model to analyze the training cost, and quantify the impact of irrational edge server assignments on hierarchical FL convergence.

#### A. System Model

The single hierarchical FL and MHFL architectures are demonstrated in Fig. 1. MHFL consists of multiple hierarchical federations, each with cloud-level aggregators, edge-level aggregators, and corresponding UEs as shown in Fig. 1(b), alleviating data heterogeneity impact from massive participating UEs.

The MEC network is represented as  $(\mathcal{ES} \cup \mathcal{CS})$ , where  $\mathcal{ES}$  and  $\mathcal{CS}$  are the sets of edge and cloud servers, respectively. Each edge server provides services to UEs within its coverage and participates in a designated federation managed by a cloud server. The edge server  $es_i$  establishes wireless connections with UEs and wired connections with the cloud server  $cs_i$ . Let  $ue_k$  represent UE  $k$  with local data  $due_k$ , where  $k \in [1, K]$  and  $K$  denotes the total number of UEs. The data volume in  $ue_k$  is  $|due_k|$ .

#### B. Federated Learning in MEC Networks

In MHFL, the network facility owner organizes cloud and edge servers to collaboratively train a global model  $w$  with UEs. Let  $w^*$  denote the optimal global parameter. The training process minimizes the loss function  $f(\cdot)$  as

$$w^* = \arg \min_w \sum_{k \in K} p_k f(w, due_k), \quad (1)$$

where  $p_k = \frac{|due_k|}{\sum_{k' \in K} |due_{k'}|}$  is the weight of UE  $k$  determined by its local data volume.

In synchronous FL training, each UE updates its local model to the edge server after  $I$  local iterations. According to vanilla SGD, the local update of  $ue_k$  is

$$w_i^k(t+1) = w_i^k(t) - \eta_t \nabla f(w_i^k(t), \xi_k), \quad (2)$$

where  $t$  is the iteration number,  $\eta_t$  is the learning rate, and  $\xi_k$  represents a sample of  $due_k$ . When  $t \bmod I = 0$ ,  $ue_k$  updates to edge server  $es_i$ .

The edge-level aggregation in  $es_i$  after  $I$  iterations is

$$w_i(t+I) = w_i(t) - \sum_{ue_k \in \mathcal{UE}_i} \sum_{t'=t}^{t+I} p_k \eta_{t'} \nabla f(w_i^k(t'), \xi_k), \quad (3)$$

where  $w_i(\cdot)$  denotes the edge-level model of  $es_i$ . This constitutes an **edge communication round**. After  $R$  edge communication rounds, the cloud server aggregates edge-level models in a **cloud communication round**

$$w(t+R \cdot I) = \sum_{i \in \mathcal{ES}} p_i w_i(t+R \cdot I), \quad (4)$$

where  $p_i = \sum_{k \in \mathcal{UE}_i} p_k^k$ . After  $G$  cloud communication rounds, the cloud server obtains a sub-optimal global model satisfying

$$f(w(t)) - f(w^*) \leq \epsilon, \quad (5)$$

where  $\epsilon$  is an arbitrary small constant and  $t = G \cdot R \cdot I$  is the total training iterations.

#### C. Cost Models

To enable the optimal edge server reallocation, we develop comprehensive cost models quantifying how different allocation strategies impact system performance. A conventional FL process includes communication, model aggregation, and local training costs.

a) *Communication cost model*: Communication cost includes two parts: UE-edge server and edge server-to-cloud server transmission. Let  $c_{k,i,r}^t$  denote the cost of transmitting the model from UE  $ue_k$  to edge server  $es_i$  in the  $r$ th iteration, and  $c_i^t$  represent the cost from edge server  $es_i$  to cloud server. Due to unstable network environments, both costs vary dynamically with training iteration  $t$ . The total communication cost during FL is

$$c_{ttl}^t = \sum_{es_i \in \mathcal{ES}} \left\{ \sum_{r=1}^{G \cdot R} \sum_{k \in \mathcal{UE}_i} c_{k,i,r}^t + G \cdot c_i^t \right\}, \quad (6)$$

where  $G$  and  $R$  denote cloud and edge communication rounds, respectively. Communication cost is determined by UE and edge server communication capability, network condition, the number of UEs and edge servers, and the required number of communication rounds. Since the number of UEs and communication capability and the required number of communication rounds are pre-determined, edge server allocation can impact the communication cost.



b) *Aggregation cost model*: Model aggregation cost occurs in edge servers and cloud servers. Each edge server's cost depends on its registered UEs, while cloud server cost depends on the number of edge servers. The total aggregation cost is

$$c_{ttl}^a = \sum_{r=1}^{G \cdot R} \sum_{i \in \mathcal{ES}} c_{i,r}^a + \sum_{r'=1}^G c_{r'}^a, \quad (7)$$

where  $c_{i,r}^a$  denotes aggregation cost in edge server  $es_i$  at edge-level communication round  $r$  and  $c_{r'}^a$  denotes cloud server aggregation cost at cloud-level communication round  $r'$ . Aggregation cost depends on the number of participating edge servers and UEs. More participants improve model performance through data diversity but increase training costs, making optimal edge server allocation crucial for cost-effective training

c) *Training time model*: Local training cost is proportional to local dataset size. Let  $c_{k,r}^l$  and  $c_{k,r}^u$  represent training cost in  $ue_k$  and unit data cost in iteration  $r$ , respectively

$$c_{k,r}^l = |due_k| \cdot c_{k,r}^u. \quad (8)$$

Since the FL process is usually synchronous, the training time is determined by the longest training time among all edge servers

$$rt_{ttl} = \max_{i \in \mathcal{ES}} \sum_{t=1}^{G \cdot R} T_i^{\max}(t), \quad (9)$$

where  $T_i^{\max}(t)$  is the longest time of  $t$  edge-level communication round, obtained directly in cloud-level aggregation.

d) *Economic expenditure*: Due to different units across communication, aggregation, and training costs, we introduce economic expenditure using monetary units to unify all cost components into a comparable framework. Federation deployment incurs three cost components dependent on training duration characterized by  $T_i^{\max}(t)$  and  $rt_{ttl}$  in (9): maintenance (energy, infrastructure upkeep), computation (model aggregation processing), and transmission (network traffic for model updates).

$$ec_{mc} = rt_{ttl} \cdot \left( \sum_{es_i \in |\mathcal{ES}|} \iota_i + \iota_{cs} \right), \quad (10)$$

where  $\iota_i$  and  $\iota_{cs}$  denote economic expenditure per unit time for edge server  $es_i$  and cloud server, respectively.

Transmission cost depends on wireless and wired communication from (6), and we have

$$ec_{tc} = \sum_{es_i \in \mathcal{ES}} \left\{ \sum_{r=1}^{G \cdot R} \sum_{k \in \mathcal{UE}_i} c_{k,i,r}^t \cdot \vartheta_{wl} + G \cdot c_i^t \cdot \vartheta_{wd} \right\}, \quad (11)$$

where  $\vartheta_{wl}$  and  $\vartheta_{wd}$  represent economic expenditure per unit of wireless and wired communication costs, respectively.

Computation cost for data aggregation is

$$ec_{cs} = R \cdot G \cdot \sum_{es_i \in \mathcal{ES}} c_i^a \cdot \varsigma_i + G \cdot c^a \cdot \varsigma_c, \quad (12)$$

where  $\varsigma_i$  and  $\varsigma_c$  denote economic expenditure for aggregating one model at edge server  $es_i$  and cloud server, respectively. Therefore, the total economic expenditure combining all

components is

$$ec_{ttl} = ec_{mc} + ec_{tc} + ec_{cs}. \quad (13)$$

#### D. Problem Formulation and Analysis

1) *Problem Formulation*: We first establish the optimization framework for a single federation to understand fundamental constraints. FL aims to optimize model performance while achieving a sub-optimal solution satisfying (5). Practical constraints require balancing model performance and training efficiency, minimizing training time while maintaining model quality and optimizing resource utilization. FL client selection can be formally expressed as

$$\min \quad ec_{ttl} + \kappa \cdot rt_{ttl} \quad (14a)$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{ES}} x_{i,r} \leq NC_i, \quad x_{i,r} \in \{0, 1\}, \quad \forall es_i \in \mathcal{ES}, \quad (14b)$$

$$T_i^{\max}(t) \leq T_{syn}^{\max}, \quad \forall es_i \in \mathcal{ES}, \quad (14c)$$

$$f(w(t)) - f(w^*) \leq \epsilon, \quad (14d)$$

$$ec_{ttl} \leq ec_{bd}, \quad (14e)$$

where  $\kappa$  trades off running time,  $NC_i$  represents the maximum number of edge servers that one single federation can access, and training cost and  $t = R \cdot G \cdot I$ . The constraints are shown as follows: (14b) limits edge servers per federation, (14c) bounds synchronous time per edge communication round, (14d) defines accuracy requirements, and (14e) enforces economic budget  $ec_{bd}$ .

Equation (14) captures the trade-off between economic efficiency and training time for a single federation. However, diverse geographical distribution and varying infrastructure ownership create multiple coexisting federations within a single area, resulting in challenges for uniform UE sampling and heightened data heterogeneity. This single-federation perspective fails to address suboptimal edge server allocation across multiple federations.

2) *Problem Analysis*: In MHFL, each federation operates with a three-tier hierarchical structure where edge servers independently select participants and cloud servers aggregate edge-level models into a global model. Edge-level model quality serves as the primary metric for evaluating edge server contributions. In this process, UE-sensitive information remains confined to edge servers, ensuring no data leakage to other entities. Meanwhile, this structure introduces increased data heterogeneity across federations. To quantify this heterogeneity, we introduce Model Difference as our evaluation metric.

*Definition 1: (Model Difference)* For any edge server  $es_i$  at iteration  $t$  with its model parameter  $w_i(t)$ , the cloud-level model difference of  $es_i$  is denoted by  $\Phi_i(t)$ , and we have

$$\|w_i(t) - \bar{w}(t)\|_2^2 \leq \Phi_i(t), \quad (15)$$

where  $\bar{w}(t)$  denotes the global model aggregated with the UE models at iteration  $t$  and can be formulated as

$$\bar{w}(t) = \sum_{i \in \mathcal{ES}} \sum_{k \in \mathcal{UE}_i} p_i^k w_i^k(t), \quad (16)$$

where  $\mathbf{w}_i^k(t)$  is the model of  $ue_k \in \mathcal{UE}_i$  at the  $t$ th iteration and  $p_i^k$  is the corresponding weight. Similar to (15), the edge-level model difference of  $ue_k$  in  $es_i$  is denoted by  $\Phi_i^k(t)$ , and we have

$$\|\mathbf{w}_i^k(t) - \mathbf{w}_i(t)\|_2^2 \leq \Phi_i^k(t). \quad (17)$$

In general, as training iterations increase, all UE models tend to achieve the optimal model  $\mathbf{w}^*$ , and thus the value of Model Difference tends to decrease.

According to Definition 1, the convergence of the hierarchical FL can be derived as follows.

*Theorem 1:* Assume  $f(\cdot)$  is  $L$ -smooth and  $\mu$ -strong convex, with bounded gradients  $\|\nabla f(\mathbf{w}_i^k(t), \xi_k)\|^2 \leq \varpi^2$  [29]. Then, we have the following results.

1) The convergence of a hierarchical FL satisfies

$$\mathbb{E}[f(\bar{\mathbf{w}}(t)) - f(\mathbf{w}^*)] \leq \frac{L}{2(\gamma+t)} \left( \frac{\beta^2 B}{\beta\mu - 1} + (\gamma+1)\Delta_1 \right), \quad (18)$$

where  $B = 8R^2(I-1)^2\varpi^2 + 6L\Gamma + \sum_{i \in \mathcal{ES}} p_i \delta_i^2$ ,  $\beta$  is a constant,  $\Delta_1 = \mathbb{E}\|\bar{\mathbf{w}}(1) - \mathbf{w}^*\|$ , and the step size  $\eta_t$  satisfies  $\eta_t = \frac{\beta}{\gamma+t}$ ,  $\eta_1 \leq \min\{\frac{1}{\mu}, \frac{1}{4L}\}$ , and  $2\eta_{t+1} \leq \eta_t$  if  $\beta > \mu$ .

2) Additionally, the total model difference  $\Phi(t)$ , measuring the FL convergence speed, is computed as

$$\Phi(t) = \mathbb{E} \sum_{i \in \mathcal{ES}} p_i \Phi_i(t) + \mathbb{E} \sum_{i \in \mathcal{ES}} \sum_{k \in \mathcal{UE}_i} p_i^k \Phi_i^k(t). \quad (19)$$

*Proof:* The proof of Theorem 1 is provided in Appendix A, available online.  $\square$

According to Theorem 1, data heterogeneity among UEs significantly impacts hierarchical FL convergence. This challenge is particularly severe in MHFL, where federations connecting with UE subsets experience heightened cloud-level data heterogeneity and slower convergence.

Since minimal Model Difference among edge-level models correlates with faster convergence, MHFL federations can reduce this difference by eliminating disparate edge-level models or incorporating models from other federations through edge server reallocation. This approach operates independently of UE operations, maintaining compatibility with various client selection methods and enabling further optimization.

While Theorem 1 employs convex assumptions for mathematical tractability, it establishes the theoretical foundation for ESR-MHFL's design with practical non-convex models such as CNNs and ResNets. The analysis demonstrates that minimizing Model Difference serves as an effective heuristic for enhancing training efficiency across federated learning systems. Our comprehensive experiments with non-convex architectures provide crucial empirical validation, confirming that ESR-MHFL's theoretically motivated design achieves robust performance in realistic deployment scenarios. The following section details ESR-MHFL's implementation framework.

#### IV. DESIGN OF EDGE SEVER REALLOCATION SCHEME IN MULTI-HIERARCHICAL FEDERATED LEARNING

This section presents ESR-MHFL, a novel edge server reallocation scheme that enhances training efficiency and optimizes resource utilization in MHFL. Our approach addresses three critical issues: identifying edge servers for reallocation, grouping them efficiently, and ensuring fair allocation among federations. ESR-MHFL operates at the federation level without accessing UE information, while maintaining compatibility with various client selection methods. As shown in Fig. 2, ESR-MHFL follows six systematic steps: (1) identify unqualified edge servers based on efficiency analysis, (2) group them into coalitions using Coalition Structure Generation with efficiency-based preferences, (3) estimate reservation values for potential accuracy improvements, (4-6) implement multi-item auctions with bidding, winner determination, and payment processing, and (7) complete reallocation to optimize performance and reduce costs.

##### A. Edge Server Quality Assessment

Theorem 1 establishes that Model Difference directly impacts convergence speed, providing the theoretical foundation for ESR-MHFL. We use this insight to design a practical edge server quality assessment framework. Since federation convergence depends critically on edge-level model quality, low-quality edge models degrade global performance and extend training time. Our framework assesses edge model quality by combining training cost metrics and training loss reductions. Similar to (13), the training cost of edge server  $es_i$  in the total FL process is computed as

$$ec_{i,tll} = \sum_{r=1}^{G \cdot R} \sum_{k \in \mathcal{UE}_i} \cdot c_{k,i,r}^t + G \cdot c_i^t + \sum_{r=1}^{G \cdot R} c_{i,r}^a. \quad (20)$$

Meanwhile, Theorem 1 indicates that a reduction in global training loss can be utilized to evaluate the quality of an edge-level model. Let  $\tau$  denote the reduction in global training loss, which can be calculated as

$$\begin{aligned} \tau &= \mathbb{E}[f(\bar{\mathbf{w}}(t)) - f(\bar{\mathbf{w}}(t + R \cdot I))] \\ &\leq \frac{L \cdot R \cdot I}{2(\gamma+t)(\gamma+t+R \cdot I)} \left( \frac{\beta^2 B}{\beta\mu - 1} + (\gamma+1)\Delta_1 \right). \end{aligned} \quad (21)$$

Combining (20) and (21), the quality evaluation of edge level models can be performed in two steps. First, based on the convergence analysis of Theorem 1, if the reduction in training loss of an edge server between two cloud-level rounds does not exceed  $\tau$ , it indicates an inefficient training process. Hence, the edge server can be categorized as unqualified. Second, the edge servers with high training cost can be categorized as unqualified. Let  $l_i$  denote the training loss reduction of  $es_i$ , i.e.,  $l_i = \mathbb{E}[f(\bar{\mathbf{w}}_i(t)) - f(\bar{\mathbf{w}}_i(t + R \cdot I))]$ . The edge servers can be classified by

$$\begin{cases} l_i < \tau \text{ or } ec_{i,tll} \geq ec_{bd}, & \text{unqualified,} \\ \text{otherwise,} & \text{qualified,} \end{cases} \quad (22)$$

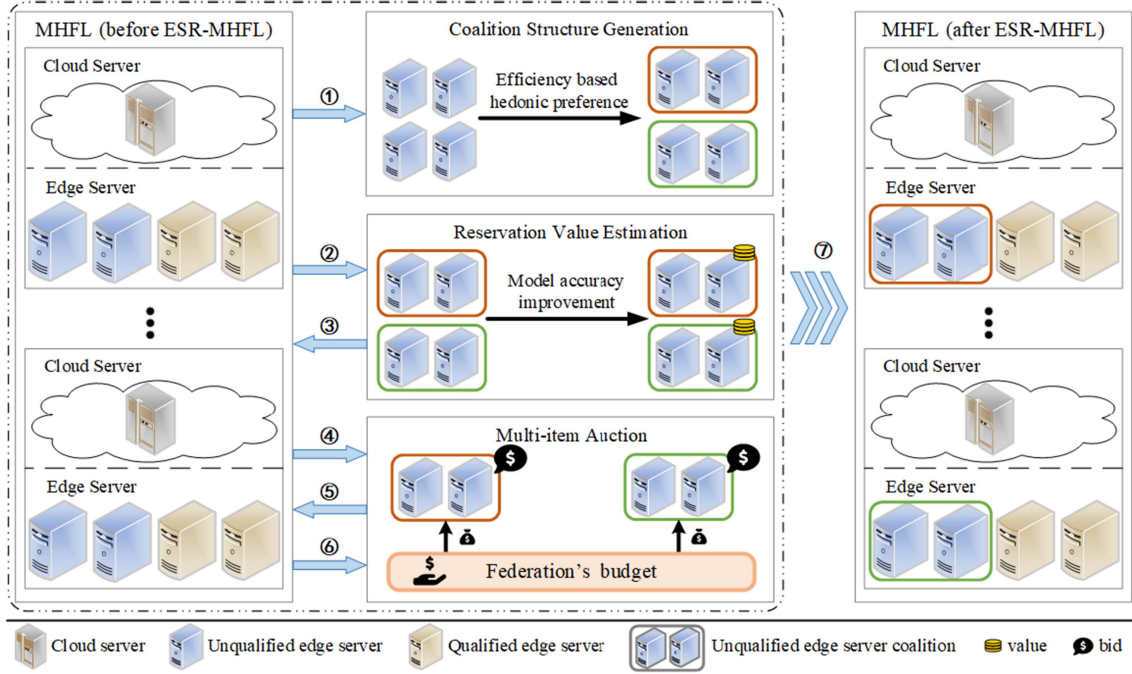


Fig. 2. ESR-MHFL process overview: (1) Select unqualified edge servers based on efficiency analysis; (2) Coalition Structure Generation groups servers using hedonic preferences; (3) Reservation value estimation evaluates potential accuracy improvements; (4-6) Multi-item auction mechanism with bidding, winner determination, and payment processing; (7) Edge server reallocation optimizes federation performance.

where  $ec_{bd}$  is a pre-defined training cost threshold.

### B. Coalition Structure Generation: Reducing Complexity

The direct optimization of edge server reallocation requires evaluating  $2^{|\mathcal{ES}_{uq}|}$  possibilities per federation, which is computationally intractable. While unqualified edge servers may impair their current federation's training efficiency, they can potentially benefit other federations through strategic reallocation. To address this complex redistribution challenge, we propose a Coalition Structure Generation (CSG)-based formation scheme that optimizes edge server reallocation by reducing computational complexity through systematic partitioning.

In MHFL, cloud server  $cl_i \in \mathcal{CL}$  represents a federation with edge server set  $\mathcal{ES}_i$ , where  $\bigcup_{cl_i \in \mathcal{CL}} \mathcal{ES}_i = \mathcal{ES}$ . Each  $\mathcal{ES}_i$  contains qualified servers  $\mathcal{ES}_{i,q}$  and unqualified servers  $\mathcal{ES}_{i,uq}$ , with  $\mathcal{ES}_{uq} = \bigcup_{cl_i \in \mathcal{CL}} \mathcal{ES}_{i,uq}$ . Our goal is improving FL performance and reducing training cost by reallocating unqualified edge servers in  $\mathcal{ES}_{uq}$ .

Maximum efficiency matching requires traversing  $2^{|\mathcal{ES}_{uq}|}$  possibilities per federation, becoming complex due to non-overlapping reallocation constraints. Therefore, we utilize Top-Responsive CSG to reduce decision-making complexity.

**Definition 2 (Top-Responsive Coalition Structure Generation):** [30] Let  $\mathcal{C}$  be a finite set of agents. A coalition structure  $\mathcal{CS} = C_1, \dots, C_k$  over  $\mathcal{C}$  satisfies  $\bigcup_{i=1}^k C_i = \mathcal{C}$  and  $C_i \cap C_j = \emptyset$  for distinct  $i, j$ . Given characteristic function  $v: 2^{\mathcal{C}} \rightarrow \mathbb{R}$ , the optimal coalition structure  $\mathcal{CS}^*$  maximizes  $\sum_{C \in \mathcal{CS}} v(C)$ . Coalition formation is top-responsive if for each agent  $i \in \mathcal{C}$ : (1) choice set  $Ch(i, X)$  is singleton for each  $X \in \mathcal{C}_i$ , (2)  $X \succ_i Y$

### Algorithm 1: Unqualified Edge Servers Formation Algorithm.

**Input:** Unqualified edge server set  $\mathcal{ES}_{uq}$ , preference  $\succsim$ , qualified accuracy  $\tau$   
**Output:** The optimal coalition structure  $\mathcal{CS}_{uq}^*$

- 1: Initialize  $R^1 \leftarrow \mathcal{ES}_{uq}$ ;  $k = 1$ ;  $\mathcal{CS}_{uq}^* \leftarrow \emptyset$ .
- 2: **repeat**
- 3:   **for all**  $i \in R^k$  **do**
- 4:     **for all**  $j \in R^k$  **do**
- 5:       **if**  $|CC(i, R^k)| \geq |CC(j, R^k)|$  **then**
- 6:          $S_{ra}^k \leftarrow CC(i, R^k)$ .
- 7:        $\mathcal{CS}_{uq}^* \leftarrow \mathcal{CS}_{uq}^* \cup S_{ra}^k$ .
- 8:        $R^{k+1} \leftarrow R^k \setminus S_{ra}^k$ .
- 9:        $k = k + 1$ .
- 10: **until**  $k = |\mathcal{CS}_{uq}^*|$  **or**  $R^{k+1} = \emptyset$
- 11: **return**  $\mathcal{CS}_{uq}^*$ .

if  $ch(i, X) > ch(i, Y)$ , and (3)  $X \succ_i Y$  if  $ch(i, X) = ch(i, Y)$  and  $X \subset Y$ .

*Proof:* The detailed definition of Definition 2 is provided in Appendix E, available online.  $\square$

According to Definition 2, we find the optimal partition  $\mathcal{CS}_{uq}^*$  of  $\mathcal{ES}_{uq}$  to reduce complexity, enabling each federation to consider only  $2^{|\mathcal{CS}_{uq}^*|}$  possibilities with stable partitions driven by agent preferences.

Algorithm 1 outlines the CSG process for the set of unqualified edge servers to obtain the optimal coalition structure  $\mathcal{CS}_{uq}^*$ . Lines 2 to 10 illustrate the process of utilizing connected

**Algorithm 2: Optimal Auction-Based ESR Scheme.**


---

**Input:** Matrix  $\mathbf{v}$ ,  $\mathcal{CL}$ ,  $\mathcal{CS}_{uq}^*$ ,  $\{ec_{i,bd}, rt_{ttl}^q, NC_i\}_{cl_i \in \mathcal{CL}}$   
**Output:** Matrix  $\mathbf{x}^*$ ,  $\mathbf{p}^v$   
 1: Initialize  $\mathbf{x}^*$ ,  $\mathbf{p}^v$ ;  
 2:  $\mathbf{x}^* \leftarrow$  Solve the ILP model(32);  
 3: **for all**  $i \in \mathcal{N}$  **do**  
 4:   Compute  $p_i$  according to (33);  
 5:    $\mathbf{p}^v \leftarrow$  append ( $p_i$ ).  
 6: **return**  $\mathbf{x}^*$ ,  $\mathbf{p}^v$ .

---

component result  $S_{ra}^k$  to find the members in the optimal CS. The optimal coalitions are stored in  $\mathcal{CS}_{uq}^*$ , which can be expressed as

$$\mathcal{CS}_{uq}^* = \{CS_1^*, CS_2^*, \dots, CS_k^*\}, \quad (23)$$

where  $k = |\mathcal{CS}_{uq}^*|$  and  $\bigcup_{CS \in \mathcal{CS}_{uq}^*} CS = \mathcal{ES}_{uq}$ .

Based on  $\mathcal{CS}_{uq}^*$ , the reallocation process can be considered as each federation finding an optimal edge server set and combining it with its qualified edge server set  $\mathcal{ES}_{i,q}$  to improve the FL training efficiency. The process can be formulated as

$$\mathcal{ES}_i^* = \mathcal{ES}_{i,q} \cup \left( \bigcup_{CS \in \mathcal{CS}_{i,uq}^*} CS \right), \quad (24)$$

where  $\bigcup \mathcal{CS}_{i,uq}^* = \mathcal{CS}_{uq}^*$  and  $i$  denotes the  $i$ th federation.

### C. Auction Mechanism: Ensuring Fairness and Efficiency

While CSG reduces computational complexity, we need a mechanism ensuring fair allocation among competing federations. We formulate edge server reallocation as a multi-item auction problem that maximizes total ESR-MHFL profit with theoretical guarantees. To guarantee auction efficiency, we introduce allocation and payment rule properties.

1) *Allocation and Payment Rules Properties:* Assuming each federation cannot learn about other federations' optimal CS, we model the allocation problem of  $\mathcal{CS}_{uq}^*$  as a multi-item auction with MHFL constraints: (1) each federation contains qualified edge server set  $\mathcal{ES}_{i,q}$ , (2) edge servers in  $\mathcal{CS}_{uq}^*$  participate only by forming coalition  $CS_i \in \mathcal{CS}_{uq}^*$ , and (3) edge servers in  $CS_i$  cannot be allocated to different federations simultaneously.

Let  $\mathbf{v}$  denote the valuation matrix with  $\mathbf{v}_i \in \mathbf{v}$  representing the real value vector of  $\mathcal{CS}_{uq}^*$  for  $cl_i$ , where element  $v_{ij} \in \mathbf{v}_i$  denotes the real value of  $CS_i$  for  $cl_i$ . To obtain  $\mathbf{v}_i$  practically, federation  $cl_i$  selects edge server coalitions in  $\mathcal{CS}_{uq}^*$  that maximize accuracy improvement while satisfying training cost budget. These edge server coalitions selected by  $cl_i$  can be represented as  $\mathcal{CS}_i^c \subseteq \mathcal{CS}_{uq}^*$ . Let  $\theta_i$  represent the estimated accuracy improvement of  $cl_i$ , the valuation estimation process can be formulated as

$$\mathbf{v}_i = \begin{cases} 0, & \text{if } ec_c > ec_{bd}, \\ \theta_i, & \text{if } ec_c \leq ec_{bd}, \end{cases} \quad (25)$$

where  $ec_c$  is the average training cost of edge servers in  $\mathcal{CS}_i^c$ . Building on Theorem 1,  $\theta_i$  represents the practical approximation of convergence bound improvement achievable by incorporating specific coalitions, where only economically feasible coalitions ( $ec_c \leq ec_{bd}$ ) are considered. Then, the value  $v_{ij} \in \mathbf{v}_i$

can be estimated as the marginal contribution to the accuracy improvement. The value  $v_{ij}$  can be formulated as

$$v_{ij} = \theta_i - \theta_{i,-j}, \quad (26)$$

where  $\theta_{i,-j}$  represents the accuracy improvement of  $\{\mathcal{CS}_i^c \setminus \mathcal{CS}_j^*\}$  participating in  $cl_i$ . This marginal estimation provides a practical implementation of theoretical Model Difference reduction, transforming abstract convergence improvements into quantifiable auction values through lightweight pre-training estimates. We use allocation variable  $\mathbf{x}^v = \{x_{i,j}^v\}_{cl_i \in \mathcal{CL}, \mathcal{CS}_j^* \in \mathcal{CS}_{uq}^*}$  to denote if coalition structure  $\mathcal{CS}_j^* \in \mathcal{CS}_{uq}^*$  is reallocated to federations  $cl_i \in \mathcal{CL}$  under valuation matrix  $\mathbf{v}$ . Specifically,  $x_{i,j}^v = 1$  means  $\mathcal{CS}_i^c$  is allocated to  $cl_j$  with the valuation matrix  $\mathbf{v}$ . The decision variable  $\mathbf{p}^v = \{p_i^v\}_{\mathcal{CS}_i^* \in \mathcal{CS}_{uq}^*}$  denotes the payment matrix of federations according to the current valuation matrix  $\mathbf{v}$ , where  $p_i^v \geq 0$  is the total payment of federation  $cl_i$  for using selected coalition structures. Given the allocation variable  $\mathbf{x}^v$  and payment variable  $\mathbf{p}^v$ , we can derive the utility (i.e., the profit of total valuation and payment) of federation  $cl_i$  as

$$U_i^v = \sum_{\mathcal{CS}_j^* \in \mathcal{CS}_{uq}^*} v_{ij} \cdot x_{ij}^v - p_{ij}^v, i \in |\mathcal{CL}|, \mathbf{v} \in \mathcal{U}, \quad (27)$$

where  $\mathcal{U}$  is the union of all valuation matrices.

The allocation and payment variables should satisfy the following properties to implement an efficient multi-item auction:

*Individual Rationality (IR):* Ensures non-negative utilities for truthful bidders:

$$p_i^v \leq \sum_{\mathcal{CS}_j^* \in \mathcal{CS}_{uq}^*} v_{ij} \cdot x_{ij}^v, \forall i \in |\mathcal{CL}|, \forall \mathbf{v} \in \mathcal{U}. \quad (28)$$

*Budget Feasibility (BF):* Ensures payments remain within budget constraints:

$$p_i^v \leq \mathcal{B}_i, \forall i \in |\mathcal{CL}|, \forall \mathbf{v} \in \mathcal{U}. \quad (29)$$

*Weak Budget Balance (WBB):* Requires non-negative total payments, ensuring the mechanism operates without subsidizing federations.

*Incentive Compatibility (IC):* Ensures truthful bidding is optimal:

$$U_i^{\mathbf{v}_i, \mathbf{v}_{-i}} \geq U_i^{\mathbf{u}_i, \mathbf{v}_{-i}}, \forall (\mathbf{v}_i, \mathbf{v}_{-i}), (\mathbf{u}_i, \mathbf{v}_{-i}) \in \mathcal{U}, \quad (30)$$

where  $\mathbf{v}_i = \{v_{ij}\}_{j \in \mathcal{CS}_{uq}^*}$  is the truthful valuation,  $\mathbf{u}_i = \{u_{ij}\}_{j \in \mathcal{CS}_{uq}^*}$  is a random valuation, and  $\mathbf{v}_{-i} = \{v_{kj}\}_{k \in \mathcal{CL} \setminus \{cl_i\}, j \in \mathcal{CS}_{uq}^*}$  excludes  $cl_i$ . The IC constraint expands to:

$$\sum_{\mathcal{CS}_j^* \in \mathcal{CS}_{uq}^*} v_{ij} \left( x_{ij}^{(v_i, \mathbf{v}_{-i})} - x_{ij}^{(u_i, \mathbf{v}_{-i})} \right) + p_{ij}^{(u_i, \mathbf{v}_{-i})} - p_{ij}^{(v_i, \mathbf{v}_{-i})} \geq 0, (u_i, \mathbf{v}_{-i}), (v_i, \mathbf{v}_{-i}) \in \mathcal{U}. \quad (31)$$

*Proof:* The proofs of IR, IC, WBB are provided in Appendix F, available online.  $\square$

2) *Problem Statement:* Based on the above constraints, the optimal auction design can be formulated as an Integer Linear Programming (ILP) problem, with the objective of maximizing



the revenue of all federations for all the valuations in set  $\mathcal{U}$ .

$$\max_{\mathbf{x}^v} \sum_{cl_i \in \mathcal{CL}} \sum_{CS_j^* \in \mathcal{CS}_{uq}^*} x_{ij}^v v_{ij} \quad (32a)$$

$$s.t. \ p_i^v \leq \sum_{j \in \mathcal{CS}_{uq}^*} v_{ij} \cdot x_{ij}^v, \forall cl_i \in \mathcal{CL}, \forall v \in \mathcal{U}, \quad (32b)$$

$$U_i^{\mathbf{v}_i, \mathbf{v}_{-i}} \geq U_i^{\mathbf{u}_i, \mathbf{v}_{-i}}, \forall (\mathbf{v}_i, \mathbf{v}_{-i}), (\mathbf{u}_i, \mathbf{v}_{-i}) \in \mathcal{U}, \quad (32c)$$

$$\sum_{cl_i \in \mathcal{CL}} \sum_{CS_j^* \in \mathcal{CS}_{uq}^*} x_{ij}^v \leq |\mathcal{CS}_{uq}^*|, \forall v \in \mathcal{U}, \quad (32d)$$

$$\bigcup_{CS_j^* \in \mathcal{CS}_{uq}^*} CS_j \cup \mathcal{ES}_{i,q} \leq NC_i, \forall cl_i \in \mathcal{CL}, \quad (32e)$$

$$ec_{i,tot} \leq ec_{i,bd}, \quad \forall cl_i \in \mathcal{CL}, \quad (32f)$$

$$x_{ij}^v \in \{0, 1\}, \quad \forall cl_i \in \mathcal{CL}, \forall v \in \mathcal{U}, \quad (32g)$$

$$p_i^v \geq 0, \quad \forall cl_i \in \mathcal{CL}, \forall v \in \mathcal{U}. \quad (32h)$$

Constraints (32b), (32c), and (32f) ensure the allocation result of each federation satisfies IR, IC, and BF properties, respectively. Constraint (32d) represents the amounts of allocated CS cannot be more than the size of  $\mathcal{CS}_{uq}^*$ . Constraint (32e) demonstrates the resource constraint of  $cl_i$ .  $NC_i$  represents the maximum number of edge servers that  $cl_i$  can access. Finally, the constraints (32g) and (32h) present allocation indicators and non-negative payments, respectively. Additionally,  $\mathbf{v} \in \mathcal{U}$  is defined as  $\{\mathbf{v} = (v_1, v_2, \dots, v_m) | v_j \in \mathcal{U}_j, \forall j \in \mathcal{CS}_{uq}^*\}$ .

3) *Payment Rule Design*: Having defined the ILP model for the multi-item auction problem, we design payment rules that ensure truthful and fair bidding by each federation  $cl_i$ . Following the VCG mechanism [31], our payment rules define the price paid to winner  $cl_i$  as the damage it causes to other participants—specifically, its marginal utility calculated as the gap between total profit with and without  $cl_i$ 's participation. Thus, the payment rule guaranteed IR and IC properties is

$$p_k^v = \sum_{CS_j^* \in \mathcal{CS}_{uq}^*} v_{kj} x_{kj}^* + \max_{cl_i \in \mathcal{CL} \setminus \{cl_k\}} \sum_{CS_j^* \in \mathcal{CS}_{uq}^*} x_{ij}^v v_{ij} - \sum_{cl_i \in \mathcal{CL}} \sum_{CS_j^* \in \mathcal{CS}_{uq}^*} x_{ij}^* v_{ij}, \quad (33)$$

where  $x_{ij}^* \in \mathbf{x}^*$  is the allocation result of  $CS_j^*$  to  $cl_i$ .

4) *Optimal Auction-Based ESR-MHFL*: The optimal auction-based ESR-MHFL is designed based on the unqualified edge servers formation algorithm and theoretical analysis. The scheme is demonstrated in Algorithm 2. This algorithm takes the optimal coalition structure of unqualified edge servers  $\mathcal{CS}_{uq}^*$ , the set of federations  $\mathcal{CL}$ , the bid matrix  $\mathbf{v}$ , and the individual properties corresponding to each federation  $cl_i$  as input parameters. The outputs include the reallocation matrix  $\mathbf{x}^*$  and the payment vector  $p^v$ . The steps of Algorithm 2 are illustrated as follows. After initializing  $\mathbf{x}^*$  and  $p^v$  in Line 1, Line 2 involves solving the ILP problem to determine the allocation that minimizes the objective function given by (32). In Lines 3–5, each federation computes the price that maximizes their

---

**Algorithm 3: Reservation Value Estimation.**


---

**Input:** the edge server coalition structure  $\mathcal{CS}_{uq}^*$ , the qualified edge server set  $\mathcal{CS}_{i,q}$

**Output:** the reservation value vector  $\mathbf{rv}_i$

```

1: Initialize  $SU \leftarrow \emptyset, \mathbf{rv}_i \leftarrow \emptyset$ 
2: for all  $CS_j^* \in \mathcal{CS}_{uq}^*$  do
3:   if  $v(\mathcal{CS}_{i,q} \cup \{CS_j^*\}) \geq v(\mathcal{CS}_{i,q})$  then
4:      $SU \leftarrow SU \cup \{CS_j^*\}$ 
5:   for  $CS_j^* \in SU$  do
6:      $ctb_i \leftarrow v(\mathcal{CS}_{i,q} \cup \{CS_j^*\}) - v(\mathcal{CS}_{i,q})$ 
7:      $SU \leftarrow SU \setminus \{CS_j^*\}, \mathbf{rv}_i \leftarrow ctb_i$ 
8: return  $\mathbf{rv}_i$ 

```

---

valuations using (33). Finally, in Line 6, the algorithm outputs the reallocation matrix  $\mathbf{x}^*$  and the payment vector  $p^v$ .

#### D. Greedy Matching-Based ESR-MHFL

Algorithm 2 obtains the theoretically optimal solution but requires solving an NP-hard ILP problem (32). With increasing dimensions of  $\mathcal{CS}_{uq}^*$  and  $\mathcal{CL}$  in massive access networks, problem size grows exponentially, creating significant computational challenges. To simplify problem scale, we propose a greedy matching-based edge server reallocation scheme.

1) *Reservation Value Estimation*: Algorithm 2's complexity depends on value matrix  $\mathbf{v}$  with dimension  $d_v = 2^{|\mathcal{CS}_{uq}^*|}$ , creating exponential search cost. To reduce complexity, we use reservation value matrix  $\mathbf{rv}$  instead of  $\mathbf{v}$ . Unlike searching the power set of  $\mathcal{CS}_{uq}^*$ , each federation evaluates single coalition contributions, yielding  $d_{rv} = |\mathcal{CS}_{uq}^*|$  and making  $\mathbf{rv}$  a significantly smaller  $d_{rv} \times d_{rv}$  matrix.

Algorithm 3 contains greedy search and estimation phases. Lines 2–4 show the greedy search phase where federation  $cl_i$  finds singleton sets of  $\mathcal{CS}_{uq}^*$  to improve performance, adding corresponding coalitions to supplement set  $SU$ . Lines 5–7 determine coalition values through marginal contribution  $ctb_i$  ((25) and (26)), obtaining reservation value vector  $\mathbf{rv}_i$  and combining all federation vectors into matrix  $\mathbf{rv}$ .

2) *Greedy Matching Algorithm*: Using the reservation value matrix  $\mathbf{rv}$ , Algorithm 4 details our greedy matching-based ESR scheme, which consists of allocation and payment phases. In the allocation phase, for each coalition  $CS_j^* \in \mathcal{CS}_{uq}^*$ , federations are sorted by their bids. The coalition is assigned to the highest-bidding federation that satisfies its economic budget ( $ec_{i,bd}$ ) and connection capacity ( $NC_i$ ). The payment phase follows second-price auction principles [32], where the price for each coalition is determined by the highest unsuccessful bid. A winning federation's final payment is the sum of these market clearing prices for all coalitions it acquires, as determined by the allocation matrix  $\mathbf{x}$ .

#### E. Computational Complexity Analysis

This section analyzes the computational complexity of ESR-MHFL to establish scalability. Let  $N_{uq}$  denote the number of



**Algorithm 4:** Greedy Matching-Based ESR Scheme.

---

**Input:** reservation value matrix  $\mathbf{rv}$ , cloud server set  $\mathcal{CL}$ , optimal CS  $\mathcal{CS}_{uq}^*$

**Output:**  $\mathbf{x}^*, \mathbf{p}$

- 1:  $\mathcal{M} \leftarrow \text{Sort}(\mathcal{CS}_j^* \in \mathcal{CS}_{uq}^*, v(\cdot)); \mathcal{N} \leftarrow \mathcal{CL}$
- 2: **while**  $\mathcal{M} \neq \emptyset$  **do**
- 3:    $j \leftarrow \text{Next}(\mathcal{M}); \mathcal{M} \leftarrow \mathcal{M} \setminus \{j\}$
- 4:    $\mathcal{N} \leftarrow \text{Sort}(i \in \mathcal{CL}, b_{ij})$
- 5:   **while**  $\sum_{i \in \mathcal{N}} x_{ij} \leq 1$  **do**
- 6:      $i \in \text{Next}(\mathcal{N})$
- 7:     **if**  $\mathbf{x}_i \cdot \mathbf{v}_i \leq e_{ci, bd}$  and  $\sum_{j \in \mathcal{M}} x_{ij} \leq NC_i$  **then**
- 8:        $x_{ij} \leftarrow 1$
- 9:   **for**  $j \in \mathcal{M}$  **do**
- 10:      $p_j \leftarrow \max_{\{x_{ij}=0\}} b_i$
- 11:   **for**  $i \in \mathcal{N} \wedge x_{ij} = 1$  **do**
- 12:      $p_i \leftarrow \sum_{j \in \mathcal{M}} p_j \cdot x_{ij}$

---

unqualified servers,  $n = |\mathcal{CS}_{uq}^*| \leq N_{uq}$  denote coalitions formed via CSG, and  $m = |\mathcal{CL}|$  denote federations.

An optimal auction without CSG requires solving an NP-hard ILP problem with worst-case complexity  $O(2^{N_{uq} \cdot m})$ . In contrast, ESR-MHFL operates as a two-stage polynomial-time process: CSG partitions  $N_{uq}$  servers into  $n$  coalitions with complexity  $O(N_{uq}^3)$ , followed by greedy matching auction allocating coalitions in  $O(n \cdot m \log m)$  time. The total computational complexity is  $O(N_{uq}^3 + n \cdot m \log m)$ .

This analysis demonstrates that ESR-MHFL reduces complexity from exponential to polynomial time, enabling viability in large-scale service computing environments.

#### F. Discussion on Formal Privacy Guarantees

While ESR-MHFL provides architectural privacy preservation by operating at the federation level without user data access, integration with formal privacy frameworks like Differential Privacy (DP) requires additional considerations. In DP-enabled systems, server reallocation may cause privacy budget leakage as reallocated models implicitly encode information about their original user populations. A privacy-aware extension would incorporate privacy costs into our auction mechanism, where server valuation balances utility contribution against consumed privacy budgets. Quantitative analysis of this privacy-utility trade-off represents a significant direction for future research.

### V. NUMERICAL RESULTS

#### A. Simulations Settings

In this section, we evaluate the training efficiency and scalability of our proposed ESR-MHFL framework through comprehensive simulations. Our experimental setup comprises 1000 clients with Dirichlet-distributed data ( $\alpha = 0.1$ ) randomly assigned across 10 edge servers organized into 3 federations, serving as the baseline MHFL configuration. Cost parameters follow established frameworks with wireless transmission cost of  $[0.1, 0.3]$  units/MB, wired cost of  $[0.05, 0.2]$  units/MB, cloud computation cost of  $[0, 0.1]$  units, and edge cost of  $[0, 0.05]$

TABLE I  
SIMULATION PARAMETERS

Parameter	Variable
batch size	32
number of UE in total	1000
client num per round	20
client sampling method	random
edge server participation	random
learning rate	0.05
group number	10
federation number	3
client optimizer	SGD
edge aggregation frequency	10
cloud aggregation frequency	100

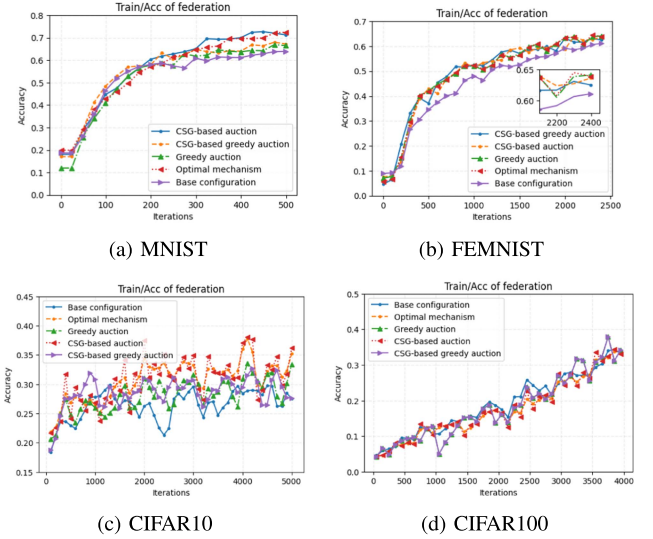


Fig. 3. Average accuracy versus training iterations of each federation with different reallocation schemes under various datasets: (a) MNIST, (b) FEMNIST, (c) CIFAR10, (d) CIFAR100.

units. We evaluate performance across four datasets (MNIST, FEMNIST, CIFAR10, CIFAR100) with corresponding model architectures (Logistic Regression, CNN, LeNet, ResNet-34) using the FedML framework on a server with Intel Xeon Platinum 8474 C processor, RTX 4090D GPU (24 GB), and 80 GB memory. Based on (22), 6 edge servers are identified as unqualified for reallocation, with performance measured through model accuracy, training cost, and training efficiency metrics. The detailed simulation settings are demonstrated in Table I.

#### B. Benchmarks

To validate our method, we employ several existing resource allocation schemes in FL and MEC as benchmarks:

*Base configuration:* 10 edge servers randomly assigned to 3 federations as the MHFL baseline, widely used in existing works [22], [33].

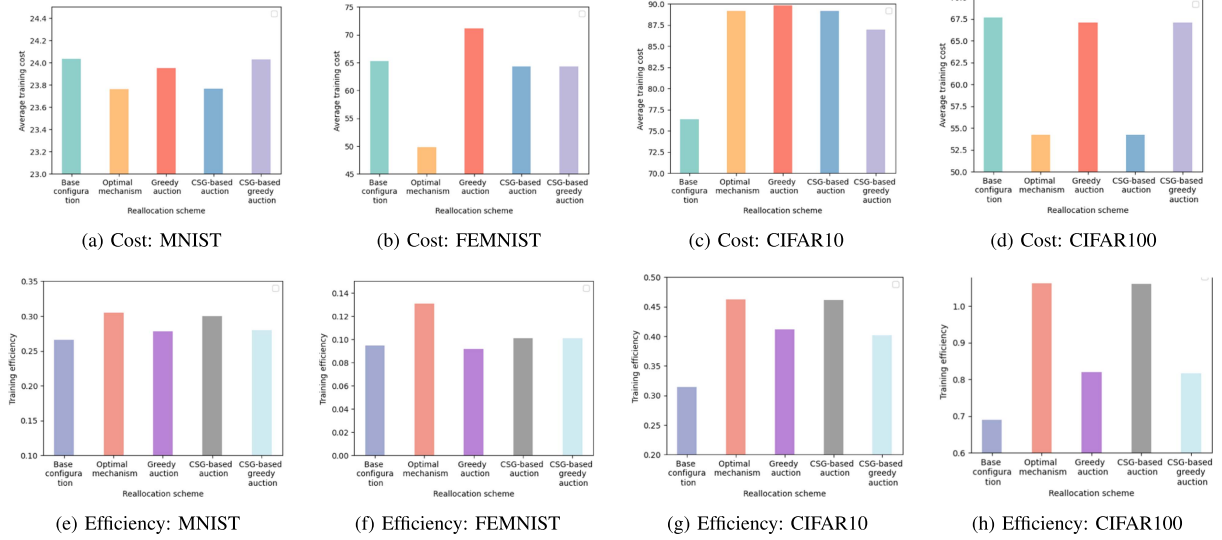


Fig. 4. Average training cost and efficiency with different reallocation schemes: (a)–(d) training cost under MNIST, FEMNIST, CIFAR10, CIFAR100; (e)–(h) training efficiency under the same datasets.

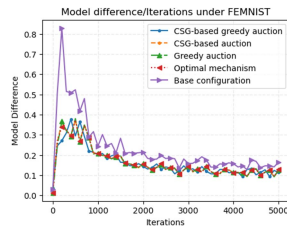


Fig. 5. The simulation results of model difference versus training iterations.

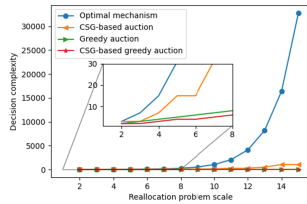


Fig. 6. The trend of the edge server reallocation complexity with the increasing number of edge servers.

**Optimal mechanism:** The optimal solution obtained by solving the ILP problem in (32) without the unqualified edge servers formation algorithm. Despite the highest algorithmic complexity, it serves as the upper bound of performance [34], [35].

**Greedy auction:** Utilizes greedy matching to solve the ILP problem in (32) without the formation algorithm, reallocating each unqualified edge server to the highest bidding federation. This approach provides an approximate solution with low complexity and is widely used for budgeted incentive mechanisms [36], [37].

### C. Performance Comparison

ESR-MHFL performance is compared with the above benchmarks using key indicators: training accuracy, edge server

reallocation complexity, model difference trend, average training cost, and training efficiency. We categorize ESR-MHFL into two versions:

**CSG-based auction:** The process of Algorithm 2, serving as base ESR-MHFL to enhance edge server reallocation efficiency.

**CSG-based greedy auction:** Contains processes in Algorithms 3 and 4. This improved ESR-MHFL handles larger-scale edge server reallocation through a greedy matching strategy, decreasing complexity and improving efficiency compared to CSG-based auction.

1) **Training Accuracy:** Training accuracy serves as a crucial FL performance indicator. As shown in Fig. 3, all reallocation schemes outperform the base configuration. Our CSG-based auction achieves comparable or superior accuracy to the optimal mechanism with significantly reduced computational overhead. In Fig. 3(c)–(d), CSG-based auction produces identical solutions to the optimal mechanism at substantially lower complexity and consistently outperforms standard greedy auction.

While accuracy improvements appear modest, ESR-MHFL achieves substantial training efficiency improvements (up to 12%) while maintaining comparable accuracy, as demonstrated in Figs. 3 and 4. This optimal balance prioritizes economic efficiency over marginal accuracy gains, establishing ESR-MHFL's superiority for cost-effective, large-scale FL systems.

2) **Training Cost and Efficiency:** Our evaluation examines training cost and efficiency across different reallocation schemes in Fig. 4. The training cost is shown in Figs. 4(a)–(d), while training efficiency (ratio of accuracy to cost) is presented in Fig. 4(e)–(h). While reallocation schemes incur higher costs than the base configuration, they consistently achieve higher training efficiency, indicating that additional resources yield proportionally greater accuracy improvements, particularly for complex models and larger datasets. Our CSG-based auction achieves comparable training efficiency to the Optimal mechanism (Fig. 4(b)–(d)), demonstrating that ESR-MHFL maintains performance quality while substantially reducing decision-making complexity.

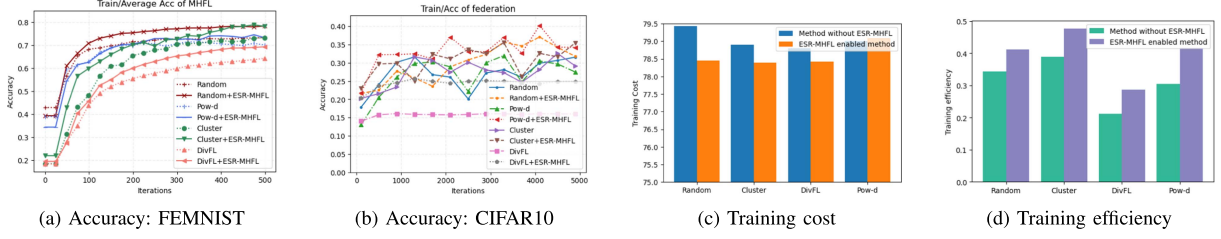


Fig. 7. Performance of ESR-MHFL with different client selection methods: (a)-(b) accuracy versus training iterations under FEMNIST and CIFAR10; (c) training cost comparison with/without ESR-MHFL; (d) training efficiency comparison with/without ESR-MHFL.

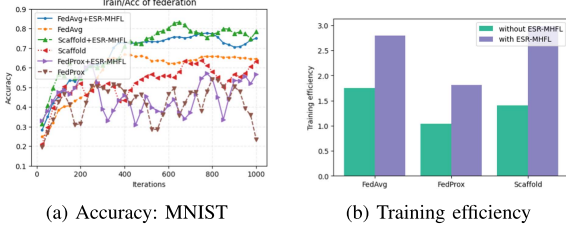


Fig. 8. Performance of ESR-MHFL with different training methods: (a) accuracy versus training iterations; (b) training efficiency comparison with/without ESR-MHFL.

CSG-based greedy auction outperforms standard Greedy auction by reducing cost and increasing training efficiency (Fig. 4(b), (c), (f), (g)). These results validate the effectiveness of ESR-MHFL in reducing edge server reallocation complexity while improving model performance and resource utilization.

3) *Trend of Model Difference and Complexity of Edge Server Reallocation*: To demonstrate Model Difference as an effective convergence indicator in FL training, we analyze its correlation with training iterations. Fig. 5 shows Model Difference decreases monotonically as training progresses, converging to a lower bound after sufficient iterations. Compared to the Base configuration, other schemes show only slight fluctuations due to edge server reallocation. Rational server allocation reduces data heterogeneity within federations, leading to consistent update directions and improved FL performance. Notably, reallocation schemes achieving higher accuracy also exhibit faster Model Difference convergence, consistent with performance trends in Fig. 3.

Fig. 6 illustrates the escalating complexity of edge server reallocation with increasing edge servers. The optimal solution grows exponentially, while ESR-MHFL significantly reduces complexity to a tolerable range. The CSG-based auction method nearly matches the Optimal mechanism's training efficiency with substantial complexity reduction. Although both Greedy auction and CSG-based greedy auction achieve linear complexity growth, CSG-based greedy auction outperforms standard Greedy auction. The CSG approach mitigates complexity increases by reducing the decision space through coalition-based edge server allocation.

#### D. Scalability of ESR-MHFL

To verify ESR-MHFL's scalability, we combine it with three client selection methods (Pow-d [38], Cluster [39], and

DivFL [40]) using random selection as baseline. As shown in Fig. 7, all methods are compatible with ESR-MHFL while enhancing performance. For example, ESR-MHFL improves average accuracy by nearly 4% for Pow-d and 5% for the Cluster method. Additionally, all methods show no cost increase, with most achieving cost reductions: DivFL and Cluster reduce cost by 0.5 units, while random selection reduces by 1 unit.

We further evaluate training efficiency impact across different client selection methods. As illustrated in Fig. 7(d), ESR-MHFL consistently yields substantial improvements in training efficiency across all methods, with gains ranging from 20.3% to 36.9%. The Pow-d selection method achieves the most significant 36.9% improvement, while even random selection achieves a 20.3% improvement. Combined with the accuracy, training cost, and efficiency results, ESR-MHFL demonstrates compatibility and scalability for different client selection methods.

To validate ESR-MHFL's compatibility, we integrated it with three foundational training algorithms, (e.g., FedAvg [41], FedProx [42], and Scaffold [43]). As shown in Fig. 8, ESR-MHFL provides significant performance enhancements across all frameworks. For example, ESR-MHFL improved the average accuracy for both FedProx and Scaffold by approximately 16% and increased training efficiency by over 60%. These results demonstrate ESR-MHFL's strong compatibility and its ability to serve as a scalable enhancement to state-of-the-art FL algorithms.

The isolation between cloud servers and UEs in ESR-MHFL enables seamless integration with existing client selection methods while avoiding privacy leakage issues in practice. As demonstrated, ESR-MHFL consistently enhances the performance of various client selection approaches while reducing overall training cost in MHFL deployments, confirming our theoretical analysis in Section III-D.

#### E. Parameter Sensitivity Analysis

Fig. 9 presents parameter sensitivity analysis demonstrating ESR-MHFL's robustness across various system configurations, examining data heterogeneity, UE populations, edge server counts, and network dynamics.

Dirichlet parameter analysis ( $\alpha = 0.05, 0.1, 0.5$ ) shows ESR-MHFL provides maximum benefits in highly heterogeneous scenarios, achieving up to 310% training efficiency improvement for high heterogeneity ( $\alpha = 0.1$ ) versus 60% for moderate heterogeneity ( $\alpha = 0.5$ ), validating effectiveness for multi-provider environments with diverse data distributions. UE



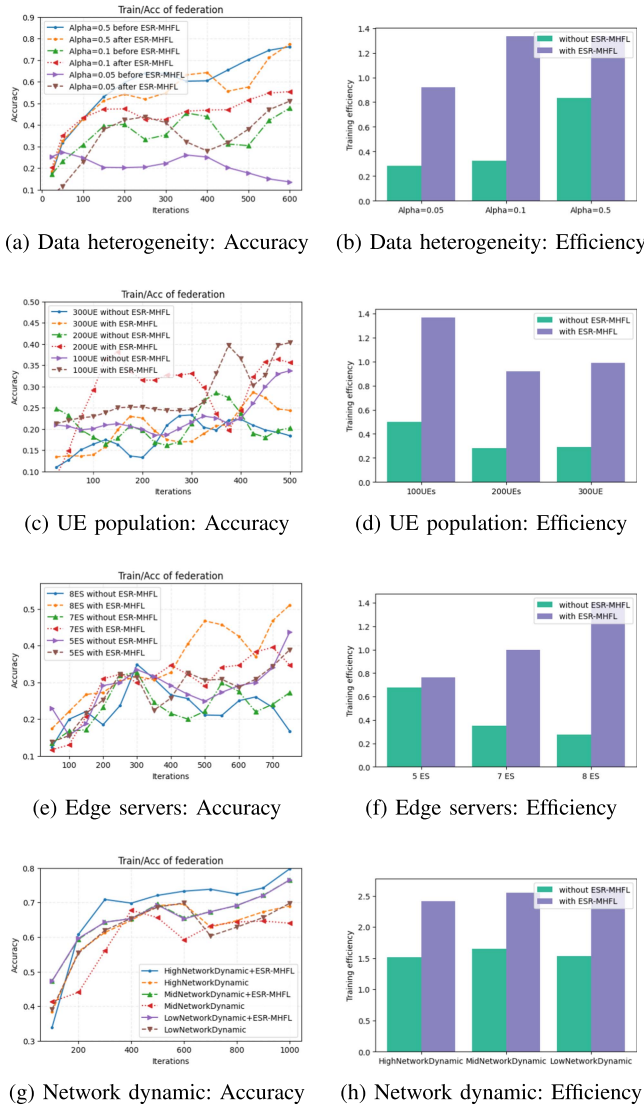


Fig. 9. Parameter sensitivity analysis of ESR-MHFL: (a)-(b) data heterogeneity impact on accuracy convergence and efficiency; (c)-(d) UE population scalability; (e)-(f) edge server configuration sensitivity; (g)-(h) network dynamic sensitivity.

scaling analysis (100–300 UEs) confirms consistent scalability with 174–241% efficiency improvements across all population scales. Edge server scaling (5–8 servers) reveals increasing benefits with infrastructure scale, from 12.7% improvement (5 servers) to 413% (8 servers), indicating greater reallocation opportunities in complex infrastructures. Finally, we analyze the impact of the network dynamic, where a higher level corresponds to a greater probability of UE disconnection. As shown in Fig. 9(g) and (h), ESR-MHFL proves robust across all tested levels, improving model performance by at least 10% and training efficiency by over 56%.

The analysis establishes deployment guidelines: highly heterogeneous data ( $\alpha \leq 0.1$ ) and substantial edge infrastructure ( $\geq 7$  servers) yield maximum improvements, while moderate scenarios still provide measurable gains, demonstrating broad practical applicability.

## VI. CONCLUSION

In this paper, we have proposed a novel MHFL architecture and ESR-MHFL scheme that enhance model performance and reduce training cost for large-scale FL. Our work introduces a comprehensive cost analysis model that quantifies how edge server assignments impact FL convergence. We formulated edge server reallocation as a multi-item auction problem, implementing a VCG-based payment rule to ensure individual rationality and incentive compatibility. To address scalability challenges, we developed an efficient greedy algorithm that effectively solves the profit maximization problem even for large-scale deployments. Extensive simulations demonstrate that ESR-MHFL significantly improves training efficiency while reducing reallocation complexity. Future research will focus on three key directions: (1) developing efficient federated architectures for massive-scale networks, (2) optimizing FL training efficiency by incorporating diverse network resource factors in MEC environments, and (3) quantitatively analyzing the privacy-utility trade-off under formal privacy guarantees. Additionally, we will explore integrating ESR-MHFL with complementary paradigms like ISCC to create a two-level hierarchy where our scheme manages macro-level server allocation while ISCC handles micro-level task optimization, with ISCC performance metrics informing auction valuation functions.

## REFERENCES

- [1] D. Wen et al., “Task-oriented sensing, computation, and communication integration for multi-device edge AI,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2486–2502, Mar. 2024.
- [2] M. Yang et al., “An improved federated learning algorithm for privacy preserving in cyberwin-driven 6G system,” *IEEE Trans. Ind. Inform.*, vol. 18, no. 10, pp. 6733–6742, Oct. 2022.
- [3] P. Wang, W. Sun, H. Zhang, W. Ma, and Y. Zhang, “Distributed and secure federated learning for wireless computing power networks,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9381–9393, Jul. 2023.
- [4] W. Sun, Y. Zhao, W. Ma, B. Guo, L. Xu, and T. Q. Duong, “Accelerating convergence of federated learning in MEC with dynamic community,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1769–1784, Feb. 2024.
- [5] X. Huang et al., “Federated learning-empowered AI-generated content in wireless networks,” *IEEE Netw.*, vol. 38, no. 5, pp. 304–313, Sep. 2024.
- [6] J. Yang, W. Jiang, and L. Nie, “Hypernetworks-based hierarchical federated learning on hybrid non-IID datasets for digital twin in industrial IoT,” *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 2, pp. 1413–1423, Mar./Apr. 2024.
- [7] X. Wang, W. Liu, H. Lin, J. Hu, K. Kaur, and M. S. Hossain, “AI-empowered trajectory anomaly detection for intelligent transportation systems: A hierarchical federated learning approach,” *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4631–4640, Apr. 2023.
- [8] M. F. Pervej and A. F. Molisch, “Resource-aware hierarchical federated learning in wireless video caching networks,” 2024, *arXiv:2402.04216*.
- [9] S. Behera, M. Adhikari, V. G. Menon, and M. A. Khan, “Large model-assisted federated learning for object detection of autonomous vehicles in edge,” *IEEE Trans. Veh. Technol.*, vol. 74, no. 2, pp. 1839–1848, Feb. 2025.
- [10] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin, “Hierarchical federated learning ACROSS heterogeneous cellular networks,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2020, pp. 8866–8870.
- [11] W. Wu, L. He, W. Lin, and R. Mao, “Accelerating federated learning over reliability-agnostic clients in mobile edge computing systems,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 7, pp. 1539–1551, Jul. 2021.
- [12] Z. Zhang, B. Guo, W. Sun, Y. Liu, and Z. Yu, “Cross-FCL: Toward a cross-edge federated continual learning framework in mobile edge computing systems,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 1, pp. 313–326, Jan. 2024.
- [13] S. Lian, H. Zhang, W. Sun, and Y. Zhang, “Lightweight digital twin and federated learning with distributed incentive in air-ground 6G networks,” in *Proc. IEEE Veh. Technol. Conf.*, Helsinki, Finland, 2022, pp. 1–5.

- [14] S. Paul, "Investigating federated learning implementation challenges in 6G network," in *Proc. 4th Int. Conf. Adv. Electr. Comput. Commun. Sustain. Technol.*, 2024, pp. 1–6.
- [15] C. Zhang et al., "Privacy-preserving federated learning for data heterogeneity in 6G mobile networks," *IEEE Netw.*, vol. 39, no. 2, pp. 134–141, Mar. 2025.
- [16] X. Hu, H. Cai, M. Alazab, W. Zhou, M. S. Haghighi, and S. Wen, "Federated learning in industrial IoT: A privacy-preserving solution that enables sharing of data in hydrocarbon explorations," *IEEE Trans. Ind. Inform.*, vol. 20, no. 3, pp. 4337–4346, Mar. 2024.
- [17] Y. Tian, S. Wang, J. Xiong, R. Bi, Z. Zhou, and M. Z. A. Bhuiyan, "Robust and privacy-preserving decentralized deep federated learning training: Focusing on digital healthcare applications," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 21, no. 4, pp. 890–901, Jul./Aug. 2024.
- [18] L. Yu, R. Albelaihi, X. Sun, N. Ansari, and M. Devetsikiotis, "Jointly optimizing client selection and resource management in wireless federated learning for Internet of Things," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4385–4395, Mar. 2022.
- [19] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.
- [20] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, Feb. 2021.
- [21] J. Huang, B. Ma, Y. Wu, Y. Chen, and X. Shen, "A hierarchical incentive mechanism for federated learning," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 12731–12747, Dec. 2024.
- [22] J. S. Ng et al., "Reputation-aware hedonic coalition formation for efficient serverless hierarchical federated learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 11, pp. 2675–2686, Nov. 2022.
- [23] S. Wang, H. Zhao, W. Wen, W. Xia, B. Wang, and H. Zhu, "Contract theory based incentive mechanism for clustered vehicular federated learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 8134–8147, Jul. 2024.
- [24] X. Tang and H. Yu, "Efficient large-scale personalizable bidding for multiagent auction-based federated learning," *IEEE Internet Things J.*, vol. 11, no. 15, pp. 26518–26530, Aug. 2024.
- [25] J. Li, Z. Chen, T. Zang, T. Liu, J. Wu, and Y. Zhu, "Reinforcement learning-based dual-identity double auction in personalized federated learning," *IEEE Trans. Mobile Comput.*, vol. 24, no. 5, pp. 4086–4103, May 2025.
- [26] L. Cui, X. Su, Y. Zhou, J. Liu, and S. Wen, "Toward optimized federated learning with compressed communications by rate adaption," *IEEE/ACM Trans. Netw.*, vol. 33, no. 3, pp. 1025–1040, Jun. 2025.
- [27] Y. Oh, Y.-S. Jeon, M. Chen, and W. Saad, "FedVQCS: Federated learning via vector quantized compressed sensing," *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 1755–1770, Mar. 2024.
- [28] J. Zhang, X. Li, P. Vijayakumar, W. Liang, V. Chang, and B. B. Gupta, "Graph sparsification-based secure federated learning for consumer-driven Internet of Things," *IEEE Trans. Consum. Electron.*, vol. 70, no. 3, pp. 5188–5200, Aug. 2024.
- [29] X. Li et al., "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Representations*, Addis Ababa, Ethiopia, 2020, pp. 1–26.
- [30] D. Dimitrov and S. C. Sung, "On top responsiveness and strict core stability," *J. Math. Econ.*, vol. 43, no. 2, pp. 130–134, 2007.
- [31] L. Makowski et al., "Vickrey-Clarke-Groves mechanisms and perfect competition," *J. Econ. Theory*, vol. 42, no. 2, pp. 244–261, Aug. 1987.
- [32] I. Brocas et al., "Second-price common value auctions with uncertainty, private and public information: Experimental evidence," *J. Behav. Exp. Econ.*, vol. 67, pp. 28–40, Apr. 2017.
- [33] L. Liu, J. Zhang, S. H. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *Proc. IEEE Int. Conf. Commun.*, Dublin, Ireland, 2020, pp. 1–6.
- [34] T. H. Thi Le et al., "An incentive mechanism for federated learning in wireless cellular networks: An auction approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 4874–4887, Aug. 2021.
- [35] S. Paris, F. Martignon, I. Filippini, and L. Chen, "An efficient auction-based mechanism for mobile data offloading," *IEEE Trans. Mobile Comput.*, vol. 14, no. 8, pp. 1573–1586, Aug. 2015.
- [36] Y. Deng et al., "AUCTION: Automated and quality-aware client selection framework for efficient federated learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 8, pp. 1996–2009, Aug. 2022.
- [37] H. Wang, S. Guo, J. Cao, and M. Guo, "MeLoDy: A long-term dynamic quality-aware incentive mechanism for crowdsourcing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 4, pp. 901–914, Apr. 2018.
- [38] Y. J. Cho et al., "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," 2020. [Online]. Available: <https://arxiv.org/abs/2010.01243>
- [39] Y. Fraboni et al., "Clustered sampling: Low-variance and improved representativity for clients selection in federated learning," 2021. [Online]. Available: <https://arxiv.org/abs/2105.05883>
- [40] R. Balakrishnan et al., "Diverse client selection for federated learning via submodular maximization," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–18.
- [41] H. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," 2023. [Online]. Available: <https://arxiv.org/abs/1602.05629>
- [42] T. Li et al., "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, 2020, pp. 429–450.
- [43] S. P. Karimreddy et al., "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 5132–5143.
- [44] T. Rahwan et al., "Coalition structure generation: A survey," *Artif. Intell.*, vol. 229, pp. 139–174, 2015.
- [45] J. Alcalde and P. Revilla, "Researching with whom? Stability and manipulation," *J. Math. Econ.*, vol. 40, no. 8, pp. 869–887, 2004.
- [46] B. Wu et al., "A truthful auction mechanism for resource allocation in mobile edge computing," in *Proc. IEEE 22nd Int. Symp. World Wireless Mobile Multimedia Netw.*, 2021, pp. 21–30.



**Tianao Xiang** (Graduate Student Member, IEEE) received the BE degree in software engineering from Northeastern University, Shenyang, China, in 2018. He is currently working toward the PhD degree in computer science with Northeastern University, Shenyang, China. His research interests include mobile edge computing, Internet of vehicles and federated learning, etc.



**Yuanguo Bi** (Member, IEEE) received the PhD degree in computer science from Northeastern University, Shenyang, China, in 2010. He was a visiting PhD student with the BroadBand Communications Research (BBRC) Lab, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada from 2007 to 2009. He is currently a professor with the School of Computer Science and Engineering, Northeastern University. He has authored/coauthored more than 50 journal/conference papers, including high quality journal papers, such

as *IEEE Journal on Selected Areas in Communications*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Intelligent Transportation Systems*, *IEEE Transactions on Vehicular Technology*, *IEEE IoT Journal*, *IEEE Communications Magazine*, *IEEE Wireless Communications*, *IEEE Network*, and mainstream conferences, such as IEEE Global Communications Conference, IEEE International Conference on Communications. His research interests include medium access control, QoS routing, multihop broadcast, and mobility management in vehicular networks, software-defined networking, and mobile edge computing. He has served as an editor/guest editor for *IEEE Communications Magazine*, *IEEE Wireless Communications*, *IEEE Access*. He has also served as the Technical Program Committee member for many IEEE conferences.



**Lin Cai** (Fellow, IEEE) has been with the Department of Electrical & Computer Engineering, University of Victoria since 2005 and is currently a professor. She is a Royal Society of Canada (RSC) fellow, an NSERC E.W.R. Steacie Memorial fellow, a Canadian Academy of Engineering (CAE) fellow, an Engineering Institute of Canada (EIC) fellow. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting ubiquitous intelligence.



**Rongfei Zeng** (Member, IEEE) received the BSc degree in computer science from Northeastern University, Shenyang, China, in 2006, and the PhD degree in computer science from Tsinghua University, Beijing, China, in 2012. He is an associate professor with the College of Software, Northeastern University. His research interests include federated learning, distributed machine learning, mobile edge computing, network security, and privacy.



**Chong Yu** (Member, IEEE) received the BSc degree in communication engineering, the MSc degree in communication and information systems from Northeastern University, Shenyang, China, in 2015 and 2017, respectively, and the PhD degree from the Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, in 2023. She is working as an assistant professor with the Department of Computer Science, University of Cincinnati. Her research interests include artificial intelligence (AI), machine learning, and cybersecurity with a broad

range of applications including data analytics, edge-based AI, and the Internet of Things.



**Mingjian Zhi** received the BS degree in network engineering from the Hebei University of Technology, Tianjin, China, in 2018, and the MS degree in computer application technology from Northeastern University, Shenyang, China, in 2021, where she is currently working toward the PhD degree in computer science and technology with the School of Computer Science and Engineering. Her current research interests include personalized federated learning and applications in mobile edge computing.



**Tom H. Luan** (Senior Member, IEEE) received the BE degree in electrical and computer engineering from Xi'an Jiaotong University, Xi'an, Shaanxi, China, the MS degree in electrical and computer engineering from The Hong Kong University of Science and Technology, Hong Kong, and the PhD degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada. He is currently with the School of Cyber Science and Engineering, Xi'an Jiaotong University. He has authored or coauthored more than 150 peer-reviewed papers in journals and conferences. His research interests include content distribution in vehicular networks, the performance evaluation of digital networks, and edge computing. He was a recipient of the 2017 IEEE VTS Best Land Transportation Best Paper Award and the IEEE ICCS 2018 Best Paper Award.