Multi-Server FL with Overlapping Clients: A Latency-Aware Relay Framework

Yun Ji[1], Zeyu Chen[1], Xiaoxiong Zhong[2], Yanan Ma[3], Sheng Zhang[1,*], and Yuguang Fang[3]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[2]Department of New Networks, Peng Cheng Laboratory, Shenzhen 518000, China
[3]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

[*]*Corresponding author: Sheng Zhang, email: zhangsh@sz.tsinghua.edu.cn*

***Abstract*—Multi-server Federated Learning (FL) has emerged as a promising solution to mitigate communication bottlenecks of single-server FL. In a typical multi-server FL architecture, the regions covered by different edge servers (ESs) may overlap. Under this architecture, clients located in the overlapping areas can access edge models from multiple ESs. Building on this observation, we propose a cloud-free multi-server FL framework that leverages Overlapping Clients (OCs) as relays for inter-server model exchange while uploading the local updated model to ESs. This enables ES models to be relayed across multiple hops through neighboring ESs by OCs without introducing new communication links. We derive a new convergence upper bound for non-convex objectives under non-IID data and an arbitrary number of cells, which explicitly quantifies the impact of inter-server propagation depth on convergence error. Guided by this theoretical result, we formulate an optimization problem that aims to maximize dissemination range of each ES model among all ESs within a limited latency. To solve this problem, we develop a conflict-graph-based local search algorithm optimizing the routing strategy and scheduling the transmission times of individual ESs to its neighboring ESs. This enables ES models to be relayed across multiple hops through neighboring ESs by OCs, achieving the widest possible transmission coverage for each model without introducing new communication links. Extensive experimental results show remarkable performance gains of our scheme compared to existing methods.**

***Index Terms*—Multi-server federated learning, Edge computing, Overlapping clients.**

I. Introduction

Federated Learning (FL) has gained considerable attention due to its ability to enable collaborative model training while preserving data privacy [1], [2]. A typical FL system operates in rounds, where clients perform local training based on their own datasets and then upload model updates to a cloud server for aggregation and broadcasting in the next round of training. However, the limitation of this single-server architecture is the long-range communication latency between clients and the cloud server (CS), especially in large-scale FL systems. To address this issue, multi-server FL architectures have been proposed [3], [4]. A prominent example is hierarchical federated learning (HFL) [3], where ESs are responsible for aggregating local model updates from their associated clients and subsequently uploading their aggregated models to the CS. However, frequent communications between ESs and the CS can still lead to high latency and resource consumption.

In [5] and [6], a new FL architecture utilizing multiple servers is studied, which exploits the realistic deployment of 5G-and-beyond networks where a client can be located in the overlapping coverage areas of multiple servers. The key idea is that clients download multiple models from all the edge servers they can access and train their local models based on the average of these models. However, OCs must wait for all relevant ESs to deliver their models before training, leading to considerable delay when inter-edge heterogeneity is pronounced. Moreover, when the number of OCs is limited, their contributions to global model convergence is marginal.

To solve these challenges, we have proposed FedOC [7], where OCs can act as relay nodes that forward edge models between adjacent cells upon reception, thereby enabling distributed aggregation across all ESs. However, this work proved convergence only for three cells, and each ES model in the current round could reach only its neighboring ESs. In this paper, we extend the convergence proof to an arbitrary number of cells and optimize ES transmission timing to maximize the propagation range of each ES model among all ESs, thereby accelerating convergence within a finite time horizon.

In summary, we highlight our main contributions as follows:

- We propose a distributed aggregated multi-server FL framework, where we exploit the OCs for model exchange between neighboring cells and facilitate cross-server collaboration. We derive the upper bound of the gap between the expected and optimal global loss values with respect to the propagation range of each ES model among all ESs. Based on this, we merge model relaying with local update uploading into a single transmission, and further optimize the model relay strategy and transmission scheduling to neighboring ESs for each ES model, so as to maximize the propagation range of each ES model within a finite time without introducing new communication links.
- To solve the optimization problem, we design a conflict-graph-based local search algorithm that jointly considers path dependency and communication constraints. For each ES l, the algorithm identifies both the farthest reachable path and the farthest received model path, constructing a conflict graph to capture mutual transmission conflicts. By combining greedy initialization with

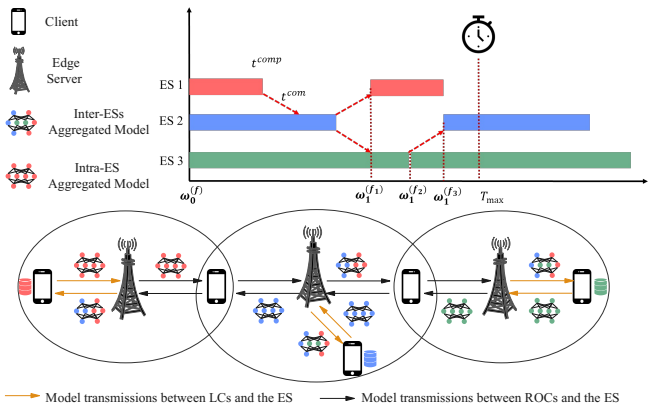


Fig. 1: Illustration of our proposed algorithm.

iterative local search, it maximizes the path weight, i.e., the total aggregated data volume across all ESs, thereby enhancing overall model propagation depth.

- We conduct extensive experiments on multiple datasets under varying numbers of cells and heterogeneous data distribution. The results demonstrate that our algorithm consistently outperforms the compared benchmarks in terms of both convergence accuracy and speed.

II. System Model

A. Network Architecture

The system consists of a set of L ESs indexed by $\mathcal{L} = \{1, 2, \ldots, L\}$, where each ES serves a group of clients within its coverage area, and some clients fall within the overlapping regions of multiple ESs. We refer to the local coverage area of each ES as a *cell*. Clients are categorized into two types: local clients (LCs), which are located in non-overlapping regions and communicate with only one ES, and overlapping clients (OCs), which are located in overlapping regions and can communicate with multiple ESs.

As shown in Fig. 1, we model overlaps using a chain topology, where each ES l overlaps with its immediate neighbors, ES $l-1$ and $l+1$, except for boundary ESs (ES 1 and ES L), which overlap with only one neighbor. While not necessarily implying a physical line, this structure captures typical cases of localized overlaps, e.g., due to limited signal range or linear deployment patterns like roads or urban blocks. It simplifies analysis while preserving the key challenges posed by overlapping coverage and multi-server coordination [7]. A single client in each overlap region, denoted $b_{l,l+1}$, serves as the Relay Overlapping Client (ROC) to forward models between ESs, while the rest are Normal OCs (NOCs). The commonly used notations are summarized in Table I.

B. Algorithm Description

Now we describe our FL algorithm tailored to the above setup. As shown in Fig. 1, the training process can be summarized as following stages, i.e., 1) local model update and upload. 2) intra-ES model aggregation. 3) inter-cell model exchange through ROC relays. 4) ES model aggregation and broadcasting.

TABLE I: Frequently Used Nomenclature and Notations

Notation	Description
$n^{(k)}$	The number of data samples of client k
$b_{l,l+1}$	The relay overlapping client (ROC) in the overlapped area between cell l and $l+1$
R	The total number of rounds
$\mathcal{S}_l$	The set of clients transmitting local models to ES l
$\tilde{N}_r^{(f_l)}$	The total number of data volume clients in $\mathcal{S}_l$
N	The total number of data samples of all clients
$\tilde{\boldsymbol{w}}_{r,E}^{(f_l)}$	The initial ES model of cell l is obtained by aggregating the models of clients in $\mathcal{S}_l$.
$\boldsymbol{w}_{r,E}^{(l+1,l)}$	The model transmitted from ES $l+1$ to ES l by ROC $b_{l+1,l}$
$\boldsymbol{w}_{r+1}^{(f_l)}$	The ES model of cell l for broadcasting to clients at the start of round $r+1$

At the beginning of each round, clients download the edge models from ESs and update their local models, with operations varying depending on client roles:

For each LC k, local training is performed for E iterations using stochastic gradient descent (SGD) on the received edge model. The update in the e-th iteration is computed as

$$\boldsymbol{w}_{r,e+1}^{(k)} = \boldsymbol{w}_{r,e}^{(k)} - \eta_{r,e} \nabla \ell_k(\boldsymbol{w}_{r,e}^{(k)}), \tag{1}$$

where $e = 0, 1, \ldots, E-1$, and $\nabla \ell_k(\boldsymbol{w}_{r,e}^{(k)})$ is the gradient of the local loss function. After local updating, each LC k transmits the local model $\boldsymbol{w}_{r,E}^{(k)}$ to corresponding ES.

For OCs, an overlapping client $k \in \mathcal{V}_{l,l+1}$ could receive two models: $\boldsymbol{w}_r^{(f_l)}$ and $\boldsymbol{w}_r^{(f_{l+1})}$. To reduce the wait latency, we assign each OC to the nearest ES as its participating server for training. Once it receives the model disseminated by that ES, it performs local training as the same as LCs based on the received model. Finally, each NOC transmits $\boldsymbol{w}_{r,E}^{(k)}$ to selected ES i, ROCs keep their updated models locally and upload them during the subsequent phase. Each ES l aggregates the received local models as

$$\tilde{\boldsymbol{w}}_{r,E}^{(f_l)} = \frac{\sum_{k \in \mathcal{S}_r^{(l)}} n^{(k)} \boldsymbol{w}_{r,E}^{(k)}}{\tilde{N}_r^{(f_l)}}, \tag{2}$$

where $\tilde{N}_r^{(f_l)} = \sum_{k \in \mathcal{S}_r^{(l)}} n^{(k)}$, and $\mathcal{S}_r^{(l)}$ consists of LCs and the NOCs that upload local models to ES l.

In the stage of inter-cell model exchange, since no additional communication links are introduced, each ES l can communicate with its ROCs at most once in each direction to relay the model to other ESs. Each ES first decides whether to transmit its model to neighboring ESs, and then determines whether to wait for the models received from its neighbors to be aggregated before forwarding them. In this way, the transmission timing to different neighbors is optimized to achieve the widest possible model propagation within a finite time horizon. The above routing strategy will be elaborated in the section IV. The relay rule of each ROC is: once ROC $b_{l,l+1}$ receives an edge model from either ES l or ES $l+1$, it will first aggregate this model with its own local model $\boldsymbol{w}_{r,E}^{(b_{l,l+1})}$

trained during the model update stage, and then forward it to the other ES, and we denote the model that ROC $b_{l,l+1}$ transmits to ES l as $\boldsymbol{w}_{r,E}^{(l+1,l)}$

For each ES l, after receiving the models and corresponding data volume from ROCs $b_{l,l+1}$ and $b_{l-1,l}$, ES l will aggregate these models with its intra-cell model to update edge model $\boldsymbol{w}_{r+1}^{(f_l)}$ for broadcast in $r+1$-th iteration, which can be written as

$$\boldsymbol{w}_{r+1}^{(f_l)} = \frac{N_r^{(l+1,l)}\boldsymbol{w}_{r,E}^{(l+1,l)} + N_r^{(l-1,l)}\boldsymbol{w}_{r,E}^{(l-1,l)} + \tilde{N}_r^{(f_l)}\tilde{\boldsymbol{w}}_{r,E}^{(f_l)}}{N_r^{(l+1,l)} + N_r^{(l-1,l)} + \tilde{N}_r^{(f_l)}}. \tag{3}$$

For boundary cells, the missing neighbor's model is taken as zero. so (3) holds for all $l \in \mathcal{L}$.

Noted that $\boldsymbol{w}_{r,E}^{(l-1,l)}$ may contain not only the locally aggregated model of cell $l-1$ but also models relayed from other cells through multi-hop transmissions. For instance, the model of cell i can be forwarded along the path $\mathcal{P}_{i\to(l-1)} = \{\text{ES } i, \text{ROC } b_{i,i+1}, \text{ES } (i+1), \ldots, \text{ROC } b_{l-2,l-1}, \text{ES } (l-1)\}$, where each relay node aggregates the received model with its own before forwarding, provided that the timing constraint is satisfied. We denote $\boldsymbol{p}_r^{(f_l)} = [p_r^{(1,l)}, \ldots, p_r^{(L,l)}]$, where $p_r^{(i,l)} \in \{0,1\}$ indicates whether cell i's model participates in ES l's aggregation. Based on (3), we have:

$$\begin{aligned}\boldsymbol{w}_{r+1}^{(f_l)} &= \frac{1}{N_{r,E}^{(f_l)}}\Bigg[\sum_{i=1}^{l-1} p_r^{(i,l)}\left(\tilde{N}_r^{(f_i)}\tilde{\boldsymbol{w}}_{r,E}^{(f_i)} + n^{(b_{i,i+1})}\boldsymbol{w}_{r,E}^{(b_{i,i+1})}\right) \\ &\quad + p_r^{(l,l)}\tilde{N}_r^{(f_l)}\tilde{\boldsymbol{w}}_{r,E}^{(f_l)} \\ &\quad + \sum_{i=l+1}^{L} p_r^{(i,l)}\left(\tilde{N}_r^{(f_i)}\tilde{\boldsymbol{w}}_{r,E}^{(f_i)} + n^{(b_{i-1,i})}\boldsymbol{w}_{r,E}^{(b_{i-1,i})}\right)\Bigg] \\ &= \sum_{i=1}^{L}\frac{p_r^{(i,l)}\hat{N}_r^{(f_i)}\hat{\boldsymbol{w}}_{r,E}^{(f_i)}}{\sum_{i=1}^{L} p_r^{(i,l)}\hat{N}_r^{(f_i)}},\end{aligned} \tag{4}$$

where $\hat{\boldsymbol{w}}_{r,E}^{(f_i)}$ is denoted as

$$\hat{\boldsymbol{w}}_{r,E}^{(f_i)} = \sum_{k\in\hat{\mathcal{K}}_r^{(f_i)}}\frac{n^{(k)}}{\hat{N}_r^{(f_i)}}\boldsymbol{w}_{r,E}^{(k)}, \tag{5}$$

where $\hat{N}_r^{(f_i)} = \sum_{k\in\hat{\mathcal{K}}_r^{(f_i)}} n^{(k)}$, and the client set $\hat{\mathcal{K}}_r^{(f_i)}$ satisfies:

$$\hat{\mathcal{K}}_r^{(f_i)} = \begin{cases}\mathcal{S}_i \cup b_{i,i+1}, & 1 \le i < l \\ \mathcal{S}_i, & i = l \\ \mathcal{S}_i \cup b_{i-1,i}, & l+1 \le i \le L.\end{cases} \tag{6}$$

C. Latency Analysis

Several key event timings are critical to our algorithm:

- **Broadcast time** (t_l^{cast}): Time for ES l to broadcast its model to its covered clients.
- **Cell update time** (t_l^{comp}): Time for all clients in a cell to complete local computations and uploading model updates for aggregation.
- **Relay start time** ($t_{l,j}^{\text{start}}$): Time when ES l begins transmitting its local aggregated ES model to a neighboring cell $j \in \{l-1, l+1\}$ through the ROC $b_{l,j}$.
- **Communication transmission time** ($t_{l,j}^{\text{com}}$): Time taken for ES l to transmit its model to a neighboring cell j through the ROC $b_{l,j}$.
- **Model aggregation time** (t_l^{agg}): Time when ES l final aggregates all the received ES models.

Assume a total spectrum budget of B for model transmission. Adjacent cells operate on disjoint sub-bands, each allocated $B/2$ bandwidth. Each ES broadcasts the model to its clients and uniformly allocates the available bandwidth among them for local model uploads. After aggregation, the ES reclaims the bandwidth and forwards the model to neighboring ESs via ROCs. The communication latency can be given by

$$t_{l,l+1}^{\text{com}} = \frac{M}{\frac{B}{4}\left[\log\left(1+\frac{4\delta_{l,l+1}P}{BN_0}\right) + \log\left(1+\frac{4\delta_{l,l+1}p}{BN_0}\right)\right]}, \tag{7}$$

where M is the model size, and $\delta_{l,l+1}$ represents the channel gain from ES l to $l+1$, with P and p being the transmission powers of the ES and client, respectively. N_0 is the noise power spectral density. The relay start time is constrained by

$$t_{l,j}^{\text{start}} \ge t_l^{\text{cast}} + t_l^{\text{comp}}, \quad \forall j \in \{l-1, l+1\}. \tag{8}$$

The aggregation time t_l^{agg} is determined by:

$$t_l^{\text{agg}} = \max\left\{p^{(j,l)}\left(t_{j,l}^{\text{start}} + t_{j,l}^{\text{com}}\right) \;\middle|\; j \in \{l-1, l, l+1\}\right\}. \tag{9}$$

III. THEORETICAL RESULTS

In this paper, we consider a C-class classification problem over a compact feature space $\mathcal{X}$ with label space $\mathcal{Y} = \{1, 2, \ldots, C\}$. Following [8], we define the population loss under the cross-entropy as $\ell(\boldsymbol{w}) = \sum_{i=1}^{C} P_{y=i}\,\mathbb{E}_{\boldsymbol{x}|y=i}\big[\log f_i(\boldsymbol{x}, \boldsymbol{w})\big]$, where $P_{y=i}$ denotes the probability of class i, and f_i is the predicted probability of class i parameterized by the model weights [8].

Now we introduce the widely used assumptions in FL:

Assumption 1. *$\nabla_{\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x}|y=i}\log f_i(\boldsymbol{x}, \boldsymbol{w})$ is $\lambda_{\boldsymbol{x}|y=i}$-Lipschitz for each class $i \in [C]$, i.e., for all $\boldsymbol{v}$ and $\boldsymbol{w}$, $\|\nabla_{\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x}|y=i}\log f_i(\boldsymbol{x}, \boldsymbol{w}) - \nabla_{\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x}|y=i}\log f_i(\boldsymbol{x}, \boldsymbol{v})\| \le \lambda_{\boldsymbol{x}|y=i}\|\boldsymbol{w} - \boldsymbol{v}\|$.*

Theorem 1. *Let Assumption 1 holds, assume $\sum_{i=1}^{C} P_{y=i}\lambda_{x|y=i} = \lambda$, and set $\boldsymbol{w}^*$ to the optimal global model. If the learning rate satisfies $\eta_{r,e} = \frac{1}{(r+1)(E-1)}$ $\forall$ r, e, then we have*

$$\begin{aligned}&\ell(\boldsymbol{w}_R^{(f_l)}) - \ell(\boldsymbol{w}^*) \\ \le&\frac{\lambda}{2}\Bigg[\sum_{j=1}^{L} D_{R-1}^{(j)}\left\|\boldsymbol{w}_{R-1}^{(f_j)} - \boldsymbol{w}_{R-1}^{(c)}\right\| + \underbrace{\frac{\beta_{R-1}}{R}}_{\epsilon^{intra}}\end{aligned}$$

$$+\underbrace{\frac{\sum_{j=1}^{3}\hat{N}_{R-1}^{(f_j)}H^{(j)}\left(\sum_{i=1}^{C}\left\|P_{y=i}^{(c_j)}-P_{y=i}^{(c)}\right\|\right)}{NR(E-1)}+F_{R-1}^{(l)}}_{\epsilon^{inter}}\Bigg], \tag{10}$$

where $F_{R-1}^{(l)}=\sum_{j=1}^{L}\left|\frac{p_{R-1}^{(j,q)}\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L}p_{R-1}^{(j,q)}\hat{N}_{R-1}^{(f_j)}}-\frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L}\hat{N}_{R-1}^{(f_j)}}\right|\left\|\hat{\boldsymbol{w}}_{R-1,E}^{(f_j)}\right\|$, *and* $\beta_{R-1}=\frac{E-1}{N}\sum_{j=1}^{L}\sum_{k\in\hat{\mathcal{K}}_r^{(f_j)}}n^{(k)}\Big(\sum_{i=1}^{C}|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}|\Big)\Big(\prod_{e=1}^{E-1}a_{R-1,e}^{(k)}\Big)g_{\max}\big(\boldsymbol{w}_{R-1}^{(c_j)}\big)$.

Proof. Please refer to the Appendix. □

ϵ^{intra} quantifies the intra-cell heterogeneity, mainly reflected by the divergence between each client's local data distribution and that of the entire cell, i.e., $\sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right|$ for $k\in\hat{\mathcal{K}}_r^{(f_j)}$, $\forall j$. This term vanishes as $R\to\infty$. ϵ^{inter} captures the inter-cell heterogeneity across different cells and consists of two parts. The first part corresponds to the divergence between each cell's distribution and that of the entire system, i.e., $\sum_{i=1}^{C}\left\|P_{y=i}^{(c_j)}-P_{y=i}^{(c)}\right\|$, which converges to zero as $R\to\infty$. The second part, $F_{R-1}^{(l)}$, arises from the model aggregation procedure in our algorithm. As the number of external cell models incorporated into $\boldsymbol{w}_r^{(f_l)}$ increases, i.e., more elements in $\boldsymbol{p}_r^{(f_l)}=[p_r^{(1,l)},\ldots,p_r^{(L,l)}]$ equals to 1, $F_r^{(l)}$ approaches zero. In the extreme case where every ES model $\boldsymbol{w}_r^{(f_l)}$ aggregates all ES models at each round, i.e., $\boldsymbol{p}_r^{(f_l)}$, $\forall\ l$ is an all-one vector, the process degenerates into centralized FL with full cloud aggregation, and thus $F_r^{(l)}=0$. To sum up, the loss error is significantly influenced by the probability of each cell participating in the aggregation of other cell models in each round.

IV. Problem Formulation and Solution

Based on **Theorem 1**, we need to minimize $F_r^{(l)}$ for all l, which is equivalent to maximizing the propagation range of each ES model within a limited time. This is achieved by optimizing the routing decisions and the relay start times of each ES, enabling broader model sharing among cells.

The transmission path from cell j to l is defined as path $=\{\text{ES } j,\text{ES } j+1,\ldots,\text{ES } l\}$. Successful transmission requires: (1) a valid link between adjacent nodes, i.e., $p^{(q,q+1)}=1$, (2) each model arriving before the next forwarding step, i.e., $t_{q,q+1}^{\text{start}}+t_{q,q+1}^{\text{com}}\le t_{q+1,q+2}^{\text{start}}$.

We define the indicator variable:

$$s^{(q,q+1)}=\mathbb{1}\Big\{t_{q,q+1}^{\text{start}}+t_{q,q+1}^{\text{com}}\le t_{q+1,q+2}^{\text{start}}\Big\}. \tag{11}$$

Then, we have

$$p^{(j,l)}=p^{(l-1,l)}\prod_{q=j}^{l-2}\big(s^{(q,q+1)}p^{(q,q+1)}\big),\ j\le l-1, \tag{12}$$

$$p^{(j,l)}=p^{(l+1,l)}\prod_{q=j}^{l+2}\big(s^{(q,q-1)}p^{(q,q-1)}\big),\ l+1\le j\le L. \tag{13}$$

We denote the optimization problem as

$$\begin{aligned}\textbf{P1:}\quad \min_{\boldsymbol{t}^{\text{start}},\boldsymbol{p}}\quad & \sum_{l=1}^{L}\sum_{j=1}^{L}\left|\frac{p^{(j,l)}\hat{N}^{(f_j)}}{\sum_{j=1}^{L}p^{(j,l)}\hat{N}^{(f_j)}}-\frac{\hat{N}^{(f_j)}}{\sum_{j=1}^{L}\hat{N}^{(f_j)}}\right|\\ \text{s.t.}\quad & (8),\ (9),\ (11)-(13),\\ & t_l^{\text{agg}}\le T_{\max},\quad \forall l\in\mathcal{L}.\end{aligned} \tag{14}$$

where $\boldsymbol{t}^{\text{start}}=\{t_{l,j}^{\text{start}}\},\ \forall\ l,j,\ \boldsymbol{p}=\{\boldsymbol{p}^{(f_l)}\},\ \forall\ l$ and $T_{\max}$ denotes the maximum duration of each round.

Since the transmission processes in opposite directions (toward left and right ESs) are independent in the constraints, the variables for each direction can be optimized separately. After removing the absolute value, **P1** can be reformulated to

$$\begin{aligned}\textbf{P2:}\quad \max_{\boldsymbol{t}_{\text{right}}^{\text{start}},\boldsymbol{p}_{\text{right}}}\quad & \sum_{l=1}^{L}\sum_{j=1}^{l}\Big(p^{(j,l)}-\mathbb{1}\Big\{p^{(j,l)}=0\Big\}\Big)\hat{N}^{(f_j)}\\ +\max_{\boldsymbol{t}_{\text{left}}^{\text{start}},\boldsymbol{p}_{\text{left}}}\quad & \sum_{l=1}^{L}\sum_{j=l+1}^{L}\Big(p^{(j,l)}-\mathbb{1}\Big\{p^{(j,l)}=0\Big\}\Big)\hat{N}^{(f_j)}\\ \text{s.t.}\quad & (8),\ (9),\ (11)-(13),\\ & t_l^{\text{agg}}\le T_{\max},\quad \forall l\in\mathcal{L}.\end{aligned} \tag{15}$$

where $\boldsymbol{t}_{\text{right}}^{\text{start}}=\{t_{l,j}^{\text{start}}\}$, $\boldsymbol{p}_{\text{right}}=\{p^{(l,j)}\},\forall\ l\le j$, $\boldsymbol{t}_{\text{left}}^{\text{start}}=\{t_{l,j}^{\text{start}}\}$, $\boldsymbol{p}_{\text{left}}=\{p^{(l,j)}\},\ \forall\ l>j$. We illustrate the algorithm using the first problem as an example. Firstly $\{p^{(j,j+1)}\}$ for $1\le j\le L-1$ can be determined straightforwardly. If $t_{j,j+1}^{\text{local}}+t_{j,j+1}^{\text{com}}\le T_{\max}$, then $p^{(j,j+1)}=1$. This ensures that ES $j+1$ at least can receive the model from ES j. If $p^{(j,l)}=1$, all intermediate relays from ES j to ES l must also be active. Thus, maximizing the objective is equivalent to finding the longest feasible relay paths. To achieve this, each ES j transmits its cell model to ES $j+1$ once intra-cell aggregation is completed. After receiving the model, ES $j+1$ forwards the aggregated model at $\max\{t_{j,j+1}^{\text{start}}+t_{j,j+1}^{\text{com}},\ t_{j+1}^{\text{cast}}+t_{j+1}^{\text{comp}}\}$, ensuring $s^{(j,j+1)}=1$ as defined in (11). This process continues until $s_{q,q+1}=0$, where no feasible $t_{q,q+1}^{\text{start}}$ satisfies $t_{q,q+1}^{\text{start}}+t_{q,q+1}^{\text{com}}\le T_{\max}$. The resulting longest path from ES j is $\{j,j+1,\ldots,q\}$. Accordingly, the optimal transmission path for each ES l is $\mathcal{P}_{\text{right}}^{(q,l)}=\{q,q+1,\ldots,l\}$, where q is the earliest ES that can successfully transmit its model to ES l. Conflicts arise when two paths share intermediate nodes. We model this as a conflict graph $\mathcal{G}(\mathcal{V},\mathcal{E})$, where vertices represent paths weighted by the total data volume of cells along this path, which we denote as $D^{(q,l)}$ for path $\mathcal{P}^{(q,l)}$, and edges denote conflicts. The problem is then equivalent to selecting a maximum-weight independent set. For small networks, exhaustive search provides the optimum. For larger networks, we propose a Conflict-Graph-Based Local Search Algorithm. The algorithm consists of three steps as follows.

Step 1. Generate an initial independent set via a greedy selection of non-conflicting high-weight vertices.

Step 2. Sort all paths by their end node numbers in ascending order, and find the sub-optimal and non-conflicted path for each node between the end node of one path and

Algorithm 1 Conflict-Graph-Based Local Search

Input: Routing decision $\boldsymbol{p}_{\text{right}}$ (or $\boldsymbol{p}_{\text{left}}$), paths $\mathcal{P}^{(q,l)}$ with weights $D^{(q,l)}$.
Output: Relay start times $\boldsymbol{t}_{\text{right}}^{\text{start}}$ (or $\boldsymbol{t}_{\text{left}}^{\text{start}}$).

```
Initialize independent set I^ini greedily by using Step 1.
Compute objective U(I^ini).
for each i ∈ I^ini do
    Replace i by feasible j to form I^new, s.t. {j, I \ i} still remain independent.
    Calculate C(I^new) and U(I^new) by using step 2.
    if U(I^new) > U(I^final) then
        Update I^final = I^new.
    end if
end for
Calculate C(I^final) and U(I^final) by using step 2.
for each node in path P^(q,l) ∈ C(I^final) do
    t^start_{j,j+1} = max(t^start_{j-1,j} + t^com_{j-1,j}, t^cast_j + t^comp_j).
end for
return t^start_right (or t^start_left).
```

the start node of the next path. Then adding these paths into $\mathcal{C}\left(\mathcal{I}^{\text{ini}}\right)$ which we named as the full paths set of $\mathcal{I}^{\text{ini}}$, and calculate the objective of **P2**, which we denote as $U\left(\mathcal{I}^{\text{ini}}\right)$.

Step 3. Iteratively refine the set by local search update higher $U\left(\mathcal{I}^{\text{final}}\right)$.

The detailed algorithm is summarized in **Algorithm 1**.

V. Performance Evaluation

A. Simulation Setup

We evaluate a multi-server network with L ESs and $K = 60$ clients, each ES covering a circular area of radius 600m, where lients are randomly distributed. The wireless channel model includes Rayleigh small-scale fading and large-scale path-loss $(128.1 + 37.6\log_{10}(d(\text{km})))$. Experiments are conducted on MNIST and CIFAR-10 with heterogeneous data distributions: each client has samples from two classes, and each ES is restricted to five classes, creating strong imbalance. The maximum per-round time $T_{\max}$ is aligned with FedOC, i.e., the maximum time for for fair comparison. For the MNIST dataset, we adopt a lightweight Convolutional Neural Network (CNN) with 21840 model parameters [3]. For CIFAR-10, we employ a deeper six-layer CNN with 1.14 million parameters [4]. Key parameters are summarized in Table II.

We compare our method with the following baselines:

- **HFL [3]**: Hierarchical FL without overlapping regions.
- **FedMES [5]**: OCs train local models based on the aggregated multiple ES models.
- **FL-EOCD [9]**: After caching and training on aggregation of received ES models, each OC uploads a model that combines its local update with the cached ES models.
- **FedOC [2]**: ESs directly forward aggregated models to neighbors via ROCs after local aggregation, without transmission time optimization.

TABLE II: Simulation Parameters

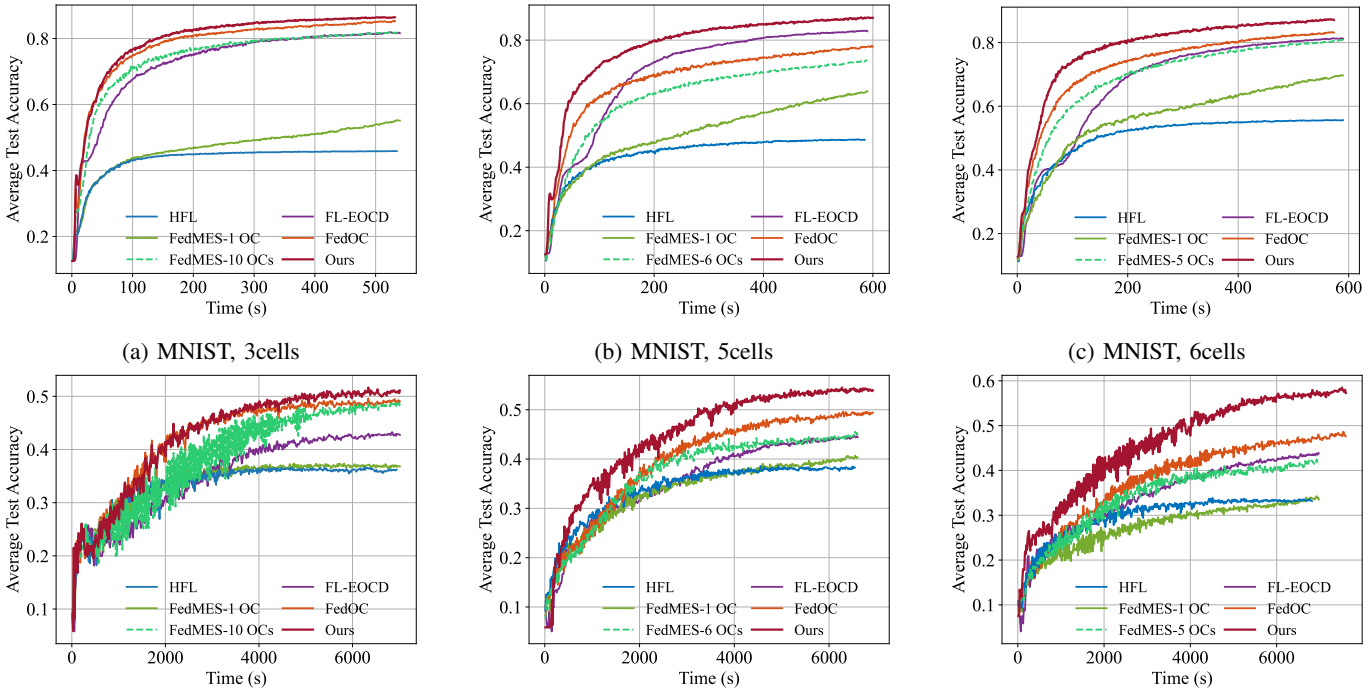
Parameters	**Values**
Number of clients, K	60
Number of iterations, R	500
Number of local training epochs, $U_k(r)$	5
Client's transmit power, p (W)	1
ES's transmit power, P (W)	5
Total channel bandwidth, B (MHz)	50
Client's one-epoch update time (seconds)	MNIST: $[0.1, 0.2]$, CIFAR-10: $[1, 2]$
Noise power spectral density, N_0 (dBm/Hz)	-174
The number of model parameters	MNIST: 21840, CIFAR-10: 1.14 million
Batch size	20
Initial learning rate, η	MNIST: 0.01, CIFAR-10: 0.1
Exponential decay factor of η in SGD	MNIST: 0.995, CIFAR-10: 0.992

TABLE III: The average number of clients aggregated per cell.

Dataset	**Cells**	**FedOC**	**Ours**
MNIST	3 cells	47.67	**53.00**
	5 cells	31.40	**50.00**
	6 cells	26.76	**50.00**
CIFAR-10	3 cells	46.67	**55.33**
	5 cells	31.40	**53.40**
	6 cells	27.79	**34.35**

As shown in Table III, we compare the average number of clients aggregated per cell between FedOC and our method. Under the same round time $T_{\max}$, our approach enables each cell to receive significantly more ES models, including those from distant cells, without extra communication links. This confirms the effectiveness of our algorithm.

Fig. 2 compares the accuracy of different benchmarks across various numbers of cells. HFL and FedMES with one OC only perform intra-cell aggregation in each round, thus their models plateau at a lower accuracies. When the OCs increase to $\lfloor K/(2L) \rfloor$ clients per region, the accuracy of FedMES increases significantly, benefiting from its reliance on multiple OCs to aggregate ES models, which indirectly enhances the generalization of each ES model. However, relying solely on the limited aggregation within OCs provides only constrained gains in model generalization, especially when the number of cells increases. FL-EOCD fares better: by integrating other cells' models, albeit with a delay, it achieves higher accuracy than HFL/FedMES in this scenario. In some MNIST cases (e.g., 5 cells) it can even outperform FedOC. However, its advantage diminishes on the more challenging CIFAR-10 task due to model staleness. In contrast, FedOC leverages real-time transmission between neighboring ESs via ROCs, achieving performance second only to our algorithm in most cases. Our algorithm builds upon FedOC by incorporating optimizations for the routing and transmission times of model exchanges, enabling models to be transmitted to more distant ESs without



(a) MNIST, 3cells (b) MNIST, 5cells (c) MNIST, 6cells

(d) CIFAR-10, 3cells (e) CIFAR-10, 5cells (f) CIFAR-10, 6cells

Fig. 2: Average test accuracy versus training time

adding extra communication links or time. This optimization accelerates model fusion, leading to faster convergence and higher model accuracy.

VI. Conclusion

In this paper, we presented a cloud-free multi-server FL framework leveraging overlapping clients as relays to enable multi-hop model dissemination among ESs without introducing additional communication links. We derived a new convergence bound of loss error of our algorithm, which characterizes the relationship between each ES model propagation depth among all ESs and the convergence error. Building upon this, we formulate an optimization problem maximizing the propagation range of edge models under latency constraints by jointly scheduling inter-server route decisions and transmission times. To solve this problem, we proposed a conflict-graph-based local search algorithm that models path dependencies and transmission conflicts as a maximum-weight independent set problem. By combining greedy initialization and iterative local refinement, it efficiently derives near-optimal relay schedules with low computational complexity. Experimental results demonstrated that the proposed framework consistently outperforms existing schemes, achieving faster convergence and higher accuracy. The gains become more pronounced as the number of cells increases, confirming that optimized relay scheduling enables deeper dissemination of models across multiple ESs and thereby accelerates convergence.

Appendix
Proof of the Theorem 1

Proof. From (6), the set $\hat{\mathcal{K}}_r^{(f_i)}$ is not only associated with cell i but also depends on cell l, which complicates the analysis. Since this difference only arises from the assignment of a single ROC and does not affect the overall client data of each cell, we approximate by assigning each ROC to the left cell in the mathematical formulation. That is,

$$\hat{\mathcal{K}}_r^{(f_i)} = \mathcal{S}_i \cup b_{i,i+1}, \quad 1 \le i < L-1. \tag{16}$$

Under these definitions, each ROC is mathematically associated with a unique cell. As shown in (4) the edge model of each ES l is the weighted sum of all participating ES models. Since the population loss $\ell(\boldsymbol{w})$ satisfies $\ell(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x},y\sim P}\left[\sum_{i=1}^{C} \mathbb{1}_{y=i} \log f_i(\boldsymbol{x},\boldsymbol{w})\right] = \sum_{i=1}^{C} P_{y=i}\mathbb{E}_{\boldsymbol{x}|y=i}\left[\log f_i(\boldsymbol{x},\boldsymbol{w})\right]$, the local update in (1) can be written as

$$\boldsymbol{w}_{r,e+1}^{(k)} = \boldsymbol{w}_{r,e}^{(k)} - \eta_{r,e} \sum_{i=1}^{C} P_{y=i}^{(k)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i}\left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{r,e}^{(k)})\right], \tag{17}$$

where $P_{y=i}^{(k)}$ represents the data proportion of class i on clients k.

To derive the convergence bound of the loss function, we introduce a cell-centralized SGD algorithm, where each ES collects the datasets of its associated clients, performs local model training, and then uploads the updates to the CS for

aggregation. Let $\boldsymbol{w}_r^{(c)}$ denote the global model of r-th round, and $\boldsymbol{w}_{r,e+1}^{(c_l)}$ represent the e-th update in the r-th round under the centralized setting. Then cell-centralized SGD performs the following update

$$\boldsymbol{w}_{r,e+1}^{(c_l)} = \boldsymbol{w}_{r,e}^{(c_l)} - \eta_{r,e} \sum_{i=1}^{C} P_{y=i}^{(c_l)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{r,e}^{(c_l)}) \right], \tag{18}$$

where $e = 0, 1, \ldots, E-1$, and $P_{y=i}^{(c_l)}$ represents the data distribution of participating clients in cell l. After E times update, the aggregated model for next round $r+1$ in centralized settings can be written as

$$\boldsymbol{w}_{r+1}^{(c)} = \sum_{i=1}^{L} \frac{\hat{N}_r^{(f_i)} \boldsymbol{w}_{r,E}^{(c_i)}}{\sum_{i=1}^{L} \hat{N}_r^{(f_i)}}. \tag{19}$$

Combining **Assumption 1** with $\sum_{i=1}^{C} P_{y=i} \lambda_{x|y=i} = \lambda$, we obtain that $\ell(\boldsymbol{w})$ is λ-smooth, i.e., $\|\nabla \ell(\boldsymbol{w}) - \nabla \ell(\boldsymbol{v})\| \leq \lambda \|\boldsymbol{w} - \boldsymbol{v}\|, \ \forall \ \boldsymbol{v}, \boldsymbol{w}$. Then we have

$$\begin{aligned} \ell(\boldsymbol{w}_R^{(f_l)}) - \ell(\boldsymbol{w}^*) &\leq \frac{\lambda}{2} \left\| \boldsymbol{w}_R^{(f_l)} - \boldsymbol{w}^* \right\| \\ &= \frac{\lambda}{2} \left\| \boldsymbol{w}_R^{(f_l)} - \boldsymbol{w}_R^{(c)} + \boldsymbol{w}_R^{(c)} - \boldsymbol{w}^* \right\| \\ &\leq \frac{\lambda}{2} \underbrace{\left\| \boldsymbol{w}_R^{(f_l)} - \boldsymbol{w}_R^{(c)} \right\|}_{A_1} + \frac{\lambda}{2} \underbrace{\left\| \boldsymbol{w}_R^{(c)} - \boldsymbol{w}^* \right\|}_{A_2}, \end{aligned} \tag{20}$$

We next focus on bounding A_1. Based on the definitions of $\boldsymbol{w}_R^{(f_l)}$ and $\boldsymbol{w}_R^{(c)}$ in (4) and (19), we have

$$\begin{aligned} A_1 &= \left\| \boldsymbol{w}_R^{(f_l)} - \boldsymbol{w}_R^{(c)} \right\| \\ &= \left\| \sum_{j=1}^{L} \frac{p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)} \hat{\boldsymbol{w}}_{R-1,E}^{(f_j)}}{\sum_{j=1}^{L} p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}} - \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)} \boldsymbol{w}_{R-1,E}^{(c_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \right\| \\ &= \left\| \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \left(\frac{p_{R-1}^{(j,l)} \sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}} \hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} \right. \right. \\ &\quad \left. \left. - \hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} + \hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} - \boldsymbol{w}_{R-1,E}^{(c_j)} \right) \right\| \\ &= \left\| \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \left[\left(\frac{p_{R-1}^{(j,l)} \sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}} - 1 \right) \hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} \right. \right. \\ &\quad \left. \left. + \left(\hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} - \boldsymbol{w}_{R-1,E}^{(c_j)} \right) \right] \right\| \\ &\leq \sum_{j=1}^{L} \left| \frac{p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}} - \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \right| \left\| \hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} \right\| \\ &\quad + \underbrace{\left\| \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \left(\hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} - \boldsymbol{w}_{R-1,E}^{(c_j)} \right) \right\|}_{B_1}. \end{aligned} \tag{21}$$

We substitute (5), (18) into B_1, and we have

$$\begin{aligned} & B_1 \\ &= \left\| \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \left(\hat{\boldsymbol{w}}_{R-1,E}^{(f_j)} - \boldsymbol{w}_{R-1,E}^{(c_j)} \right) \right\| \\ &\quad \cdot \sum_{i=1}^{C} P_{y=i}^{(k)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-1}^{(k)}) \right] \Bigg) - \boldsymbol{w}_{R-1,E-1}^{(c_j)} \\ &\quad + \eta_{R-1,E-1} \sum_{i=1}^{C} P_{y=i}^{(c_j)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-1}^{(c_j)}) \right] \Bigg] \Bigg\| \\ &\overset{1}{=} \left\| \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \sum_{k=1}^{\hat{K}_r^{(f_j)}} \frac{n^{(k)}}{\hat{N}_{R-1}^{(f_j)}} \left[\left(\boldsymbol{w}_{R-1,E-1}^{(k)} - \boldsymbol{w}_{R-1,E-1}^{(c_j)} \right) \right. \right. \\ &\quad + \eta_{R-1,E-1} \sum_{i=1}^{C} P_{y=i}^{(k)} \left(\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-1}^{(k)}) \right] \right. \\ &\quad \left. \left. - \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-1}^{(c_j)}) \right] \right) \right] \Bigg\| \\ &\overset{2}{\leq} \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \sum_{k=1}^{\hat{K}_r^{(f_j)}} \frac{n^{(k)}}{\hat{N}_{R-1}^{(f_j)}} \cdot \\ &\quad \left(1 + \eta_{R-1,E-1} \sum_{i=1}^{C} P_{y=i}^{(k)} \lambda_{\boldsymbol{x}|y=i} \right) \\ &\quad \cdot \left\| \boldsymbol{w}_{R-1,E-1}^{(k)} - \boldsymbol{w}_{R-1,E-1}^{(c_j)} \right\| \end{aligned} \tag{22}$$

where equality 1 holds because for each class $i \in \{1, 2, \ldots, C\}$, $P_{y=1}^{(c_j)} = \sum_{k=1}^{\hat{K}_r^{(f_j)}} \frac{n^{(k)}}{\hat{N}_R^{(f_j)}} P_{y=1}^{(k)}$, i.e., the data distribution over all the clients in cell j is the same as the data distribution over the whole cell population. Inequality 2 holds because of the **Assumption 1**. Inspired by [8], we give the bounding of $\left\| \boldsymbol{w}_{R-1,E-1}^{(k)} - \boldsymbol{w}_{R-1,E-1}^{(c_j)} \right\|$ for client $k \in \hat{\mathcal{K}}_r^{(c_j)}$ in cell j:

$$\begin{aligned} & \left\| \boldsymbol{w}_{R-1,E-1}^{(k)} - \boldsymbol{w}_{R-1,E-1}^{(c_j)} \right\| \\ &\overset{3}{=} \left\| \left(\boldsymbol{w}_{R-1,E-2}^{(k)} - \boldsymbol{w}_{R-1,E-2}^{(c_j)} \right) \right. \\ &\quad + \left(\eta_{R-1,E-2} \sum_{i=1}^{C} P_{y=i}^{(k)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(c_j)}) \right] \right. \\ &\quad \left. \left. - \eta_{R-1,E-2} \sum_{i=1}^{C} P_{y=i}^{(k)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(k)}) \right] \right) \right\| \\ &\leq \left\| \boldsymbol{w}_{R-1,E-2}^{(k)} - \boldsymbol{w}_{R-1,E-2}^{(c_j)} \right\| \\ &\quad + \eta_{R-1,E-2} \left\| \sum_{i=1}^{C} P_{y=i}^{(k)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(k)}) \right] \right. \\ &\quad \left. - \sum_{i=1}^{C} P_{y=i}^{(k)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(c_j)}) \right] \right. \end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{C} P_{y=i}^{(k)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i}\left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(c_j)})\right] \\
&-\sum_{i=1}^{C} P_{y=i}^{(c_j)} \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i}\left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(c_j)})\right] \Big\| \\
\leq & \left\|\boldsymbol{w}_{R-1,E-2}^{(k)}-\boldsymbol{w}_{R-1,E-2}^{(c_j)}\right\| \\
&+\eta_{R-1,E-2}\Bigg[\Bigg\| \sum_{i=1}^{C} P_{y=i}^{(k)}\Big(\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i}\left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(k)})\right] \\
&-\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i}\left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(c_j)})\right]\Big)\Bigg\|+ \\
&\Bigg\| \sum_{i=1}^{C}\left(P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right) \nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i}\left[\log f_i(\boldsymbol{x}, \boldsymbol{w}_{R-1,E-2}^{(c_j)})\right] \Bigg\|\Bigg] \\
\overset{4}{\leq} & \left(1+\eta_{R-1,E-2} \sum_{i=1}^{C} P_{y=i}^{(k)} \lambda_{x|y=i}\right)\left\|\boldsymbol{w}_{R-1,E-2}^{(k)}-\boldsymbol{w}_{R-1,E-2}^{(c_j)}\right\| \\
&+\eta_{R-1,E-2} \sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right| g_{\max}(\boldsymbol{w}_{R-1,E-2}^{(c_j)}) \\
\overset{5}{=} & a_{R-1,E-2}^{(k)}\left\|\boldsymbol{w}_{R-1,E-2}^{(k)}-\boldsymbol{w}_{R-1,E-2}^{(c_j)}\right\| \\
&+\eta_{R-1,E-2} \sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right| g_{\max}(\boldsymbol{w}_{R-1,E-2}^{(c_j)}) \qquad (23)
\end{aligned}
$$

where equality 3 holds because (17) and (18). Inequality 4 holds because the **Assumption 1** and we denote $g_{\max}\left(\boldsymbol{w}_{r,q}^{(c_j)}\right) = \max_{i=1}^{C}\left\|\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{x}|y=i} \log f_i\left(\boldsymbol{x}, \boldsymbol{w}_{r,q}^{(c_j)}\right)\right\|$ for $r \in \{1,2,\ldots,R\}$ and $q \in \{0,1,\ldots,E\}$. equality 5 holds because we denote $a_{r,q}^{(k)} = 1+\eta_{r,q} \sum_{i=1}^{C} P_{y=i}^{(k)} \lambda_{\boldsymbol{x}|y=i}$ for $r \in \{1,2,\ldots,R\}$ and $q \in \{0,1,\ldots,E\}$. From (23) we deduce the relation between $\left\|\boldsymbol{w}_{R-1,E-1}^{(k)}-\boldsymbol{w}_{R-1,E-1}^{(c_j)}\right\|$ and $\left\|\boldsymbol{w}_{R-1,E-2}^{(k)}-\boldsymbol{w}_{R-1,E-2}^{(c_j)}\right\|$. By repeatedly using this relation, we can recursively obtain:

$$
\begin{aligned}
&\left\|\boldsymbol{w}_{R-1,E-1}^{(k)}-\boldsymbol{w}_{R-1,E-1}^{(c_j)}\right\| \\
\leq & a_{R-1,E-2}^{(k)}\left\|\boldsymbol{w}_{R-1,E-2}^{(k)}-\boldsymbol{w}_{R-1,E-2}^{(c_j)}\right\| \\
&+\sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right| \eta_{R-1,E-2} g_{\max}(\boldsymbol{w}_{R-1,E-2}^{(c_j)}) \\
\leq & \left(\prod_{q=0}^{E-2} a_{R-1,q}^{(k)}\right)\left\|\boldsymbol{w}_{R-1,0}^{(k)}-\boldsymbol{w}_{R-1,0}^{(c_j)}\right\|+\sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right| \\
&\cdot \sum_{q=0}^{E-2} \eta_{R-1,q}\left(\prod_{t=q+1}^{E-2} a_{R-1,t}^{(k)}\right) g_{\max}\left(\boldsymbol{w}_{R-1,q}^{(c_j)}\right) \\
\leq & \left(\prod_{q=0}^{E-2} a_{R-1,q}^{(k)}\right)\left\|\boldsymbol{w}_{R-1}^{(f_j)}-\boldsymbol{w}_{R-1}^{(c)}\right\|+\sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right| \\
&\cdot \sum_{q=0}^{E-2} \eta_{R-1,q}\left(\prod_{t=q+1}^{E-2} a_{R-1,t}^{(k)}\right) g_{\max}\left(\boldsymbol{w}_{R-1,q}^{(c_j)}\right), \qquad (24)
\end{aligned}
$$

where we assume $\prod_{a}^{b} a_{r,e}^{(k)} = 1,\ \forall b < a,\ r$ and e. Institute (24) into (22), we have

$$
\begin{aligned}
B_1 \leq & \sum_{j=1}^{L} \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \sum_{k=1}^{\hat{K}_r^{(f_j)}} \frac{n^{(k)}}{\hat{N}_{R-1}^{(f_j)}}\Bigg[\left(\prod_{e=0}^{E-1} a_{R-1,e}^{(k)}\right) \\
&\cdot\left\|\boldsymbol{w}_{R-1}^{(f_j)}-\boldsymbol{w}_{R-1}^{(c)}\right\|+\sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right| \\
&\cdot \sum_{d=0}^{E-2} \eta_{R-1,d}\left(\prod_{e=d+1}^{E-1} a_{R-1,e}^{(k)}\right) g_{\max}\left(\boldsymbol{w}_{R-1,d}^{(c_j)}\right)\Bigg] \\
= & \sum_{j=1}^{L}\left(D_{R-1}^{(j)}\left\|\boldsymbol{w}_{R-1}^{(f_j)}-\boldsymbol{w}_{R-1}^{(c)}\right\|+G_{R-1}^{(j)}\right),
\end{aligned} \qquad (25)
$$

where $D_{R-1}^{(j)} = \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \sum_{k=1}^{\hat{K}_r^{(f_j)}} \frac{n^{(k)}}{\hat{N}_{R-1}^{(f_j)}}\left(\prod_{e=0}^{E-1} a_{R-1,e}^{(k)}\right)$, and $G_{R-1}^{(j)} = \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \sum_{k=1}^{\hat{K}_r^{(f_j)}} \frac{n^{(k)}}{\hat{N}_{R-1}^{(f_j)}}\left(\sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right|\right) \sum_{e=0}^{E-2} \eta_{R-1,d}\left(\prod_{d=e+1}^{E-1} a_{R-1,d}^{(k)}\right) g_{\max}\left(\boldsymbol{w}_{R-1,e}^{(c_j)}\right)$. Plug (25) into (21), we have

$$
\begin{aligned}
A_1 = & \left\|\boldsymbol{w}_{R}^{(f_l)}-\boldsymbol{w}_{R}^{(c)}\right\| \\
\leq & \sum_{j=1}^{L}\left|\frac{p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}}-\frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}}\right|\left\|\hat{\boldsymbol{w}}_{R-1}^{(f_j)}\right\| \\
&+\sum_{j=1}^{L}\left(D_{R-1}^{(j)}\left\|\boldsymbol{w}_{R-1}^{(f_j)}-\boldsymbol{w}_{R-1}^{(c)}\right\|+G_{R-1}^{(j)}\right) \\
\overset{6}{=} & \sum_{j=1}^{L} D_{R-1}^{(j)}\left\|\boldsymbol{w}_{R-1}^{(f_j)}-\boldsymbol{w}_{R-1}^{(c)}\right\|+F_{R-1}^{(l)}+G_{R-1}
\end{aligned} \qquad (26)
$$

where equality 6 holds because we denote

$$
F_{R-1}^{(l)} = \sum_{j=1}^{L}\left|\frac{p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} p_{R-1}^{(j,l)} \hat{N}_{R-1}^{(f_j)}}-\frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}}\right|\left\|\hat{\boldsymbol{w}}_{R-1,E}^{(f_j)}\right\|, \qquad (27)
$$

and $G_{R-1} = \sum_{j=1}^{L} G_{R-1}^{(j)}$.

In the following, we first derive a theoretical upper bound for G_{R-1}. We denote

$$
G_{R-1}^{(j)} = \sum_{e=0}^{E-2} \eta_{R-1,e}\, \beta_{R-1,e}^{(j)}, \qquad (28)
$$

and

$$
\beta_{R-1,e}^{(j)} = \frac{\hat{N}_{R-1}^{(f_j)}}{\sum_{j=1}^{L} \hat{N}_{R-1}^{(f_j)}} \sum_{k=1}^{\hat{K}_r^{(f_j)}} \frac{n^{(k)}}{\hat{N}_{R-1}^{(f_j)}}\left(\sum_{i=1}^{C}\left|P_{y=i}^{(k)}-P_{y=i}^{(c_j)}\right|\right)
$$

$$\cdot \left(\prod_{d=e+1}^{E-1} a_{R-1,d}^{(k)} \right) g_{\max}\left(\boldsymbol{w}_{R-1,e}^{(c_j)} \right). \tag{29}$$

We assume that there exist constants $g_{\max}$ and $\beta_r^{(j)}$ such that $g_{\max}\left(\boldsymbol{w}_{r,e}^{(c_j)}\right) \leq g_{\max}$ and $\beta_{r,e}^{(j)} \leq \beta_r^{(j)}$, $\forall r,\ e,\ j$. Moreover, we denote $\sum_{j=1}^{L} \beta_r^{(j)} = \beta_r$. Since $a_{r,q}^{(k)} = 1 + \eta_{r,q} \sum_{i=1}^{C} P_{y=i}^{(k)} \lambda_{\boldsymbol{x}|y=i} > 1,\ \forall r,\ q$, based on (22), we further assume $\beta_{\min} \leq \beta_r \leq \beta_{\max}$. Then we have

$$G_{R-1} = \sum_{e=0}^{E-2} \eta_{R-1,e} \sum_{j=1}^{L} \beta_{R-1,e}^{(j)} \leq \sum_{e=0}^{E-2} \eta_{R-1,e} \beta_{R-1} \leq \frac{\beta_{R-1}}{R}, \tag{30}$$

Finally, we will give the upper bound of $A2$. $A2$ is the model divergence between the cell-centralized SGD algorithm and $\boldsymbol{w}^*$. If we see each cell as a client, the cell-centralized SGD is the standard FedAvg. Under the assumption that $\boldsymbol{w}^*$ is the final model of the global centralized SGD, where data in all clients are collected in the CS to train a global model. Under the assumptions of **Theorem 1**, [8] gives the bound of A_2 as

$$\begin{aligned} A_2 &= \left\| \boldsymbol{w}_R^{(c)} - \boldsymbol{w}^* \right\| \\ &\leq \frac{\sum_{j=1}^{3} \hat{N}_{R-1}^{(f_j)} H^{(j)} \left(\sum_{i=1}^{C} \left\| P_{y=i}^{(c_j)} - P_{y=i}^{(c)} \right\| \right)}{NR(E-1)}. \end{aligned} \tag{31}$$

where N is the total data volume of all clients and $H^{(j)} = \sum_{e=0}^{E-1} \left(\prod_{d=e+1}^{E-1} a_{R-1,d}^{(c_j)} \right) g_{\max}(\boldsymbol{w}_{R-1,e}^{(c)})$.

Substituting (26), (30) and (31) into (20), we have

$$\begin{aligned} &\ell(\boldsymbol{w}_R^{(f_l)}) - \ell(\boldsymbol{w}^*) \\ \leq &\frac{\lambda}{2} \Bigg[\sum_{j=1}^{L} D_{R-1}^{(j)} \left\| \boldsymbol{w}_{R-1}^{(f_j)} - \boldsymbol{w}_{R-1}^{(c)} \right\| + \underbrace{\frac{\beta_{R-1}}{R}}_{\epsilon^{\text{inter}}} \\ &+ \underbrace{\frac{\sum_{j=1}^{3} \hat{N}_{R-1}^{(f_j)} H^{(j)} \left(\sum_{i=1}^{C} \left\| P_{y=i}^{(c_j)} - P_{y=i}^{(c)} \right\| \right)}{NR(E-1)} + F_{R-1}^{(l)}}_{\epsilon^{\text{inter}}} \Bigg] \end{aligned} \tag{32}$$

which completes the proof. □

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 1273–1282, iSSN: 2640-3498.

[2] G. Zhu, Y. Deng, X. Chen, H. Zhang, Y. Fang, and T. F. Wong, "ESFL: Efficient Split Federated Learning over Resource-constrained Heterogeneous Wireless Devices," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 27 153–27 166, 2024.

[3] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-Edge-Cloud Hierarchical Federated Learning," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. Dublin, Ireland: IEEE, Jun. 2020, pp. 1–6.

[4] X. Chen, G. Zhu, Y. Deng, and Y. Fang, "Federated Learning Over Multihop Wireless Networks With In-Network Aggregation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4622–4634, Jun. 2022.

[5] D.-J. Han, M. Choi, J. Park, and J. Moon, "FedMes: Speeding Up Federated Learning With Multiple Edge Servers," *IEEE J. Select. Areas Commun.*, vol. 39, no. 12, pp. 3870–3885, Dec. 2021.

[6] Z. Qu, X. Li, J. Xu, B. Tang, Z. Lu, and Y. Liu, "On the Convergence of Multi-Server Federated Learning With Overlapping Area," *IEEE Trans. on Mobile Comput.*, vol. 22, no. 11, pp. 6647–6662, Nov. 2023.

[7] Y. Ji, Z. Chen, X. Zhong, Y. Ma, S. Zhang, and Y. Fang, "Fedoc: Multi-server fl with overlapping client relays in wireless edge networks," 2025. [Online]. Available: https://arxiv.org/abs/2509.19398

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated Learning with Non-IID Data," Jul. 2022, arXiv:1806.00582 [cs].

[9] M. S. Al-Abiad, M. Obeed, M. J. Hossain, and A. Chaaban, "Decentralized Aggregation for Energy-Efficient Federated Learning via D2D Communications," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3333–3351, Jun. 2023.