

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enabling Large-Scale Federated Learning over Wireless Edge Networks

Thinh Quang Dinh*, Diep N. Nguyen*, Dinh Thai Hoang*, Pham Tran Vu[†], and Eryk Dutkiewicz*

*University of Technology Sydney

[†]Ho Chi Minh City University of Technology

Email: {Thinh.Dinh, Diep.Nguyen, Hoang.Dinh, Eryk.Dutkiewicz}@uts.edu.au, {ptvu}@hcmut.edu.vn

Abstract—Major bottlenecks of large-scale Federated Learning (FL) networks are the high costs for communication and computation. This is due to the fact that most of current FL frameworks only consider a star network topology where all local trained models are aggregated at a single server (e.g., a cloud server). This causes significant overhead at the server when the number of users are huge and local models’ sizes are large. This paper proposes a novel edge network architecture which decentralizes the model aggregation process at the server, thereby significantly reducing the aggregation latency of the whole network. In this architecture, we propose a highly-effective in-network computation protocol consisting of two components. First, an in-network aggregation process is designed so that the majority of aggregation computations can be offloaded from cloud server to edge nodes. Second, a joint routing and resource allocation optimization problem is formulated to minimize the aggregation latency for the whole system at every learning round. The problem turns out to be NP-hard, and thus we propose a polynomial time routing algorithm which can achieve near optimal performance with a theoretical bound. Numerical results show that our proposed framework can dramatically reduce the network latency, up to 4.6 times. Furthermore, this framework can significantly decrease cloud’s traffic and computing overhead by a factor of K/M , where K is the number of users and M is the number of edge nodes, in comparison with conventional baselines.

Index Terms—Mobile Edge Computing, Federated Learning, In-network Computation

I. INTRODUCTION

The last decade has witnessed the adoption of machine learning (ML) and artificial intelligence (AI) as the core engines of intelligent systems [1]. Under most ML-based frameworks, raw data are collected and trained at centralized cloud servers, raising concerns in user privacy, latency, and network overhead. Federated Learning (FL) has recently emerged as a potential distributed learning solution to these issues [2]. Under FL, mobile users (MUs), instead of sharing their raw data with the server, can build and learn their local learning models. After that, they only need to send these local model parameters to the centralized server [3]. By doing so, the MUs can iteratively download the new global model from the server, update their local models using their local training data, and then upload their new local trained models to the server for the model aggregation. This process is repeated until the global model converges or after a predefined number of learning rounds reaches.

However, given its distributed setting, communication and computation costs are the two major bottlenecks of FL [2]–[4]. In addition, due to a huge demands of advanced AI-

based mobile applications, learning tasks are more and more complicated with very large data sizes. For example, with a large model like Visual Geometry Group-16 (VGG-16), each user needs to update about 500 TB of data until the global model is converged [5]. Since conventional FL models use star network topologies, during the model aggregation step, the cloud generally needs to connect with a huge number of users. In such a case, aggregation operations at the cloud incur (a) high transmission latency, (b) high traffic overhead and (c) high computational overhead in term of processing and memory resources. To overcome these challenges, edge computing (EC) has recently emerged as a great potential solution by “moving” computing resources closer to end users [3]. Since edge nodes possess both computation and communication capacities, edge networks can decentralize the model aggregation computations at the cloud server in very large scale FL networks. To that end, it is critical to develop a distributed in-network aggregation functionality implemented at edge networks’ components in order to address current challenges of FL.

In-network computation (INC) is a process of gathering, processing data at intermediate nodes then routing the processed data through a multi-hop network [6]. INC has been well-studied for distributed data clusters such as MapReduce [7], Pregel [8] and DryadLINQ [9]. Three basic components of an in-network computation solution are: suitable networking protocols, effective aggregation functions, and efficient methods for data representation [6]. The early work of Liu *et al.* [10] proposed to aggregate/average users’ models at an edge node that later sends these intermediate model parameters to the cloud server. However, in this work, users are assumed to connect directly to a single edge node without any alternative paths. In practice, due to dense deployment of edge networks [11], a given MU can associate one or another or even with multiple nearby edge nodes. As a result, the problems of network routing and resource allocation for the model aggregation in FL under EC become more challenging.

Given the above, this paper proposes a novel edge network architecture aiming at minimizing the aggregation latency of FL processes. This architecture allows the cloud node to decentralize its aggregation process to the edge nodes. To accomplish that network functionality, we design an in-network computation protocol which consists of two components: an in-network aggregation process and a network routing algorithm. Specifically, the in-network aggregation process guides on how packets are processed at edge nodes and cloud

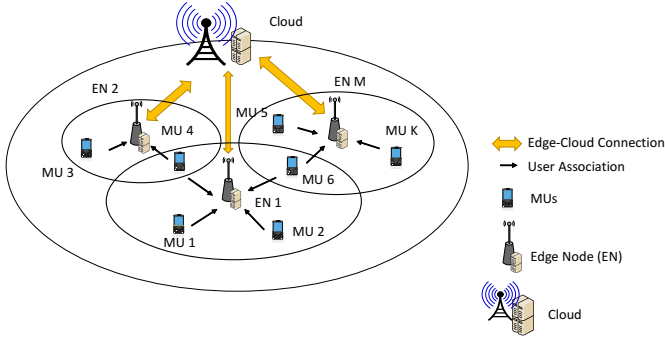


Fig. 1. FL-enabled Edge Computing Network Architecture.

node and how the cloud decentralizes the model aggregation process of FL. Then, we formulate the joint routing and resource allocation optimization problem aiming to minimize the network's aggregation latency. The problem turns out to be NP-hard. We thus propose an effective algorithm based on randomized rounding techniques, which provably achieves an approximation guarantee. Finally, simulation results show that our proposed solutions significantly reduce not only the network's aggregation latency but also the cloud node's traffic and computing overhead.

II. SYSTEM MODEL

As illustrated in Fig. 1, let's consider a set of K mobile users MUs, denoted by $\mathcal{K} = \{1, \dots, K\}$ with local datasets $\mathcal{D}_k = \{\mathbf{x}_i \in \mathbb{R}^d, y_i\}_{i=1}^{n_k}$ with n_k data points. Let $\mathcal{M} = \{0, \dots, M\}$ denote the set of edge nodes (ENs). They can be co-located with small cell base stations which have communications and computing capacities [12]. These ENs are connected with a macro base stations, equipped with a cloud server, denoted as EN 0. Each user can be associated with one or more ENs.

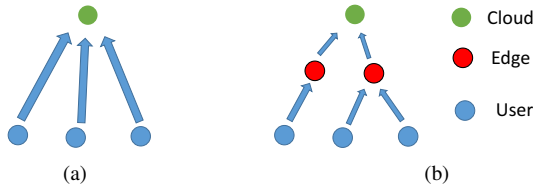


Fig. 2. The logical view of (a) conventional network model and (b) multi-tier edge network model.

A. Federated Learning

To construct the shared global model, the goal is to find the model parameters $\mathbf{w} \in \mathbb{R}^d$ which minimize the following global loss function in a distributed manner:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ P(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n l_i(\mathbf{x}_i^T \mathbf{w}) + \xi r(\mathbf{w}) \right\}, \quad (1)$$

where $n = \sum_{i=1}^K n_k$, ξ is the regularizing parameter, $r(\mathbf{w})$ is a deterministic penalty function and l_i is the loss function at data sample i [13]. Here, we also use notation ψ for the global model.

To solve (1), a Federated Learning framework introduced in [2] is performed as following. At each iteration t , the cloud broadcasts the global model ψ^t to all the MUs. Based on the latest global model, each MU learns its local parameters \mathbf{w}_k^t according to the Stochastic Gradient Descent update rule aiming at minimizing the objective function $P(\mathbf{w})$ by only using local information and the parameter value in ψ^t [2]:

$$\mathbf{w}_k^t = \psi^t - \eta(\nabla l_i(\psi^t) + \nabla r(\psi^t)). \quad (2)$$

The resulting local model updates are forwarded to the cloud for computing the new global model as follows:

$$\psi^{t+1} = \frac{1}{n} \sum_{k=1}^K n_k \mathbf{w}_k^t. \quad (3)$$

We summarize the procedures of the FL framework as follows:

1. Global Model Broadcasting: The cloud broadcasts the latest global model ψ^t to the MUs.
2. Local Model Updating: Each MU performs local training following (2).
3. Global Model Aggregation: Local models are then sent back to the cloud. The new value of global model is computed following (3).
4. Steps 1-3 are repeated until convergence.

B. Communication Model

We then introduce the communication model for multi-user access. For each FL iteration, the cloud node will select a set of users \mathcal{K}^t at each iteration t ¹. All users consent about their models' structure, such as a specific neural network design. Hence, let D denote the data size of model parameters, which is fixed and identical for all users, where D is proportional to the cardinality of \mathbf{w} [5].

1) *Global Model Broadcasting*: Since the downlink communication capacity of the cloud node is much larger than that of an edge node, all users will listen to the cloud node at the model broadcasting step. Let W^d denote the downlink communication capacity of the cloud node. The latency for broadcasting the global model is $T^d = \frac{D}{W^d}$.

2) *Global Model Aggregation*: Let a_{km} be the aggregation routing variable, where

$$a_{km} = \begin{cases} 1 & \text{if MU } k \text{'s is associated with EN } m, \forall m \in \mathcal{M}, \\ 0 & \text{otherwise.} \end{cases}$$

Here, we assume that an MU is not allowed to transmit data directly to the macro base station to reduce the uplink traffic overhead. However, MUs can listen to the downlink channel in network broadcast messages. Let $\mathbf{a}_m = [a_{1m}, a_{2m}, \dots, a_{Km}]^T$ denote the uplink association vector of edge node m . We let $\mathbf{A} = \{a_{km}\} \in \{0, 1\}^{K \times M}$ denote the uplink association matrix, and $\tilde{\mathbf{a}} = [\mathbf{a}_0^T, \mathbf{a}_1^T, \dots, \mathbf{a}_M^T]^T$ denote the column vector corresponding to \mathbf{A} . Let \mathcal{K}_m^t denote the set of user associated with edge node m , then we have

¹The learner selection in FL can be based on the quality or significance of information or location learners [3]. Here, how to select best MUs at each learning round is out of the scope of this paper.

$\bigcup_m \mathcal{K}_m^t = \mathcal{K}^t$, and $|\mathcal{K}^t| = \sum_m |\mathcal{K}_m^t|$, where $|\cdot|$ denotes the cardinality of a set.

Let r_{km} denote the uplink data rate between MU k and edge node m . Let $\mathbf{r}_m = [r_{1m}, r_{2m}, \dots, r_{Km}]^T$ denote the uplink bandwidth allocation vector corresponding to edge node m . We denote $\mathbf{R} = r_{km} \in \mathbb{R}^{K \times M}$ as the uplink bandwidth allocation matrix. Let B_m^{fr} , and B_m^{bk} denote the uplink fronthaul and backhaul capacity of edge node m . Then, uplink communication latency between edge node m and its associated users is the longest latency of a given user:

$$T_m^{\text{u,fr}} = \max_{k \in \mathcal{K}_m} \left\{ \frac{D a_{km}}{r_{km}} \right\}, \text{ where } r_{km} \leq B_m^{\text{fr}}, \forall m \in \mathcal{M} \setminus \{0\}. \quad (4)$$

After edge nodes receive local models, each edge node can perform its aggregation computation, then send the aggregated result to the cloud node. Alternatively, edge nodes just forward received models to the cloud. Let γ_m denote the transmission latency between edge node m and the cloud node. Without in-network aggregation functionality, γ_m is computed as followed

$$\gamma_m = \frac{D \sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{bk}}}. \quad (5)$$

The uplink aggregation latency of users associated with edge node m is

$$T_m^{\text{u}} = T_m^{\text{u,fr}} + \gamma_m. \quad (6)$$

III. IN-NETWORK AGGREGATION DESIGN

We now introduce the in-network computation protocol where edge nodes support the cloud node for averaging users' local models. First, we design the user packet which plays a role of data representation in a in-network computation protocol. Let $\phi_k^t = \{\phi_k^t[0], \phi_k^t[1]\}$ denote the local message of users k at iteration t such that

$$\{\phi_k^t[0], \phi_k^t[1]\} = \{n_k, \mathbf{w}_k^t\}. \quad (7)$$

A. In-Network Aggregation Process

First, consider the following in-network aggregation (INA) process at edge nodes and the cloud node that helps decentralize the aggregation process at the cloud node. Let χ_m^t denote the average local model of edge node m such that

$$\chi_m^t = \frac{1}{\sum_{k \in \mathcal{K}_m} \phi_k^t[0]} \sum_{k \in \mathcal{K}_m} \phi_k^t[0] \phi_k^t[1]. \quad (8)$$

Let $\varphi_m^t = \{\varphi_m^t[0], \varphi_m^t[1]\}$ denote the message edge node m sends to the cloud node such that

$$\{\varphi_m^t[0], \varphi_m^t[1]\} = \left\{ \sum_{k \in \mathcal{K}_m} \phi_k^t[0], \chi_m^t \right\}. \quad (9)$$

To conserve the result of (3), the global model is computed as follows:

$$\psi^{t+1} = \frac{\sum_m \varphi_m^t[0] \varphi_m^t[1]}{\sum_m \varphi_m^t[0]}. \quad (10)$$

Theorem 1. *The edge network architecture as well as the INA process reduce the traffic and computing overhead at the cloud*

node by a factor of K/M in comparison with conventional star network topologies.

Proof. The proof is omitted here for brevity. \square

B. Revised Latency Model

With the proposed INA process, let γ'_m denote the transmission latency between an edge node m and the cloud node. If there is no user associate with an edge node m , i.e., $\sum_k a_{km} = 0$, γ'_m is zero. Otherwise, since edge node m only needs to send its aggregated model, computed in (8), to the cloud node, γ'_m is computed as follows

$$\gamma'_m = \min \left\{ \frac{D}{B_m^{\text{bk}}}, \frac{D \sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{bk}}} \right\}. \quad (11)$$

IV. NETWORK ROUTING AND RESOURCE ALLOCATION FRAMEWORK FOR FL

In this section, we aim to minimize the total uplink aggregation latency by jointly optimizing (a) which edge node a user should send its local model directly to and (b) the optimal data rates for wireless connections between the users and the edge nodes. The total uplink aggregation latency is computed as follows

$$T^{\text{u}}(\mathbf{A}, \mathbf{R}) = \max_m \left\{ T_m^{\text{u,fr}} + \gamma'_m \right\}. \quad (12)$$

The aggregation latency-minimized routing framework is formulated as followed

$$\mathcal{P}_1 : \min_{\mathbf{A}, \mathbf{R}} T^{\text{u}}(\mathbf{A}, \mathbf{R}),$$

$$\text{s.t. } \sum_{m=0}^M a_{km} = 1, \forall k \in \mathcal{K}^t, \quad (13a)$$

$$\sum_{k \in \mathcal{K}_m} r_{km} \leq B_m^{\text{fr}}, \forall m \in \mathcal{M} \setminus \{0\}, \quad (13b)$$

$$a_{km} \in \{0, 1\}, \quad (13c)$$

$$r_{km} \in [0, B_m^{\text{fr}}], \forall m \in \mathcal{M} \setminus \{0\}. \quad (13d)$$

The constraints (13a) guarantee that a user can associate with only one edge node in one iteration. The constraints (13b) ensure that total users' data rates associated with each edge node must not exceed its bandwidth capacity. \mathcal{P}_1 is a mixed-integer nonlinear programming, which is NP-hard.² We will propose a highly efficient random rounding solution for practical implementation in the next section.

V. RANDOMIZED ROUNDING BASED SOLUTION

In this section, we present an approximation algorithm for the main problem that leverages a randomized rounding technique [14]. Firstly, \mathcal{P}_1 is transformed to an equivalent integer linear program (ILP). Then, by relaxing the integer constraints, \mathcal{P}_1 becomes a linear programming which can be solved by linear solvers. We first observe that:

Lemma 1. *Given any uplink association matrix \mathbf{A} , with $|\mathcal{K}_m| = \sum_{k \in \mathcal{K}_m} a_{km} > 0$, for problem \mathcal{P}_1 , at each edge*

²The proof is omitted here for brevity.

node m , the uplink latency for users associated with edge node m satisfies

$$T_m^u = \max_{k \in \mathcal{K}_m} \left\{ D \frac{a_{km}}{r_{km}} \right\} + \min \left\{ \frac{D}{B_m^{\text{bk}}}, \frac{D \sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{bk}}} \right\} \\ \geq \frac{D|\mathcal{K}_m|}{B_m^{\text{fr}}} + \min \left\{ \frac{D}{B_m^{\text{bk}}}, \frac{D|\mathcal{K}_m|}{B_m^{\text{bk}}} \right\}. \quad (14)$$

The equality happens when $r_{1m} = \dots = r_{|\mathcal{K}_m|m} = \frac{B_m^{\text{fr}}}{|\mathcal{K}_m|}$.

Proof. The proof is omitted here for brevity. \square

Following Lemma 1, the network operator hence only needs to optimize the uplink association matrix while the uplink data rates for users associated with edge nodes will be allocated in a fairness manner. If $|\mathcal{K}_m| = 0$, we arbitrarily set the values of \mathbf{r}_m and T_m^u to be 0. As a result, \mathcal{P}_1 is reduced to

$$\mathcal{P}_2 : \min_{\mathbf{A}} \max_m \left\{ D \frac{\sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{fr}}} \right. \\ \left. + \min \left\{ \frac{D}{B_m^{\text{bk}}}, \frac{D \sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{bk}}} \right\} \right\}, \\ \text{s.t. (13a) and (13c),} \quad (15)$$

where the the optimal solution in \mathcal{P}_1 can be computed from optimal solution in \mathcal{P}_2 as the following lemma:

Lemma 2. Let \mathbf{A}^{**} denote the optimal solution in \mathcal{P}_2 . Following Lemma 1, the optimal solution in \mathcal{P}_1 , $\{\mathbf{A}^*, \mathbf{R}^*\}$ is computed as follows

$$\mathbf{A}^* = \mathbf{A}^{**}, \\ r_{1m}^* = \dots = r_{|\mathcal{K}_m|m}^* = \frac{B_m^{\text{fr}}}{|\mathcal{K}_m|}. \quad (16)$$

Proof. The proof is omitted here for brevity. \square

The proposed algorithm is described in detail below and summarized in Algorithm 1. First, we introduce auxiliary variables y and $\gamma = \{\gamma_1, \dots, \gamma_M\}$ into \mathcal{P}_2 such that

$$y \geq \max_m \left\{ D \frac{\sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{fr}}} + \gamma_m \right\}, \quad (17)$$

$$\gamma_m \leq \min \left\{ \frac{D}{B_m^{\text{bk}}}, \frac{D \sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{bk}}} \right\}, \forall m \quad (18)$$

Problem \mathcal{P}_2 is then equivalently transformed to

$$\mathcal{P}_3 : \min_{y, \gamma, \mathbf{A}} y, \\ \text{s.t. } y \geq D \frac{\sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{fr}}} + \gamma_m, \forall m, \quad (19a)$$

$$\gamma_m \leq \frac{D}{B_m^{\text{bk}}}, \forall m \in \mathcal{M} \setminus \{0\}, \quad (19b)$$

$$\gamma_m \leq \frac{D \sum_{k \in \mathcal{K}_m} a_{km}}{B_m^{\text{bk}}}, \forall m \in \mathcal{M} \setminus \{0\}, \quad (19c)$$

(13a) and (13c).

Algorithm 1 Randomized Routing Algorithm for Low Latency Federated Learning

Input: $D, B_m^{\text{fr}}, B_m^{\text{bk}}$, and \mathcal{K}^t .

Output: $\mathbf{A}^{(\text{Alg})}, \mathbf{R}^{(\text{Alg})}$

- 1: Solve \mathcal{P}_4 to achieve \mathbf{A}^\dagger .
- 2: **if** \mathbf{A}^\dagger is binary **then**
- 3: $\mathbf{A}^{(\text{Alg})} = \mathbf{A}^\dagger$
- 4: **else**
- 5: **for** $k = 1$ to $k = K$ **do**
- 6: $a_{km}^{(\text{Alg})} = 1$ with probability a_{km}^\dagger with exclusive manner based on constraints (13a)
- 7: **end for**
- 8: **end if**
- 9: Then,

$$r_{1m}^{(\text{Alg})} = \dots = r_{|\mathcal{K}_m|m}^{(\text{Alg})} = \frac{B_m^{\text{fr}}}{|\mathcal{K}_m|}, \forall m \text{ such that } |\mathcal{K}_m| \neq 0.$$

The Algorithm 1 starts by solving the Linear Relaxation (LR) of \mathcal{P}_3 . Specifically, it relaxes the variables a_{km} to be fractional, rather than integer. The Linear Relaxation of \mathcal{P}_3 can be expressed as follows:

$$\mathcal{P}_4 : \min_{y, \gamma, \mathbf{A}} y, \\ \text{s.t. (19a) - (19c), (13a), and } a_{km} \in [0, 1]. \quad (20)$$

Let $z^\dagger = [\tilde{\mathbf{a}}^\dagger, y^\dagger, \gamma_1^\dagger, \dots, \gamma_M^\dagger]$ denote the optimal solution of \mathcal{P}_4 . First, vector $\tilde{\mathbf{a}}^\dagger$ should be transformed to an equivalent fractional matrix \mathbf{A}^\dagger , whose elements are in $[0, 1]$ by a “reshape” operation. The term “reshape” means to change the size of a vector or a matrix while its number of elements is unchanged. \mathbf{A}^\dagger is the optimal solution to \mathcal{P}_2 , if all components of \mathbf{A}^\dagger are binary. Otherwise, to obtain binary matrix $\mathbf{A}^{(\text{Alg})}$, for each row of \mathbf{A}^\dagger , we perform a randomization by setting the element a_{km} to 1 with probability a_{km}^\dagger . The decision is done in an exclusive manner for satisfying constraints (13a). It means that for each row k , only one element of the row is one, the rest are zeros. The random decision is made independently for all k . By doing this procedure, the matrix $\mathbf{A}^{(\text{Alg})}$ is achieved. Then, $r_{1m}^{(\text{Alg})} = \dots = r_{|\mathcal{K}_m|m}^{(\text{Alg})} = \frac{B_m^{\text{fr}}}{|\mathcal{K}_m|}, \forall m$. The complexity of this algorithm is $O(\nu^{3.5} \Omega^2)$, where $\nu = KM + M + 1$.

Theorem 2. The aggregation latency returned by Algorithm 1 is at most $\frac{2 \ln K}{y^\dagger} + 3$ times higher than that of the optimal with high probability, where K is the number of MUs and y^\dagger is the lower bound of the aggregation latency which can be obtained in polynomial time.

Proof. The proof is omitted here for brevity. \square

VI. NUMERICAL RESULTS

In this section, simulations are conducted to show the performance of the proposed algorithm. We consider a similar setup as in [12], depicted in Fig. 3. Here, $M = 9$ edge nodes are regularly deployed in a grid network inside a $500 \times 500 \text{ m}^2$ area. $K = 1000$ mobile users are distributed uniformly at random over the EN coverage regions (each of 150m radius). In our simulation, without loss of generality, all K users’ models are aggregated in one learning iteration. The cloud

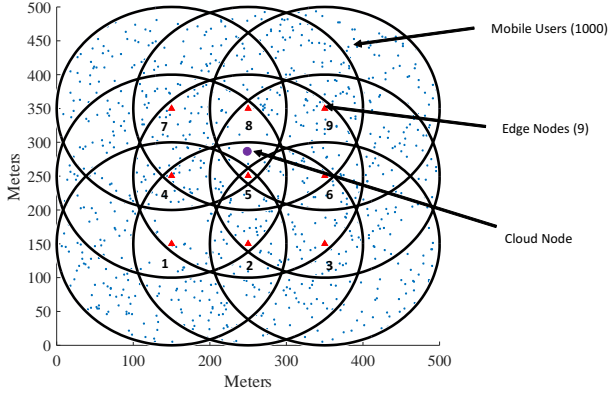


Fig. 3. The network setup.

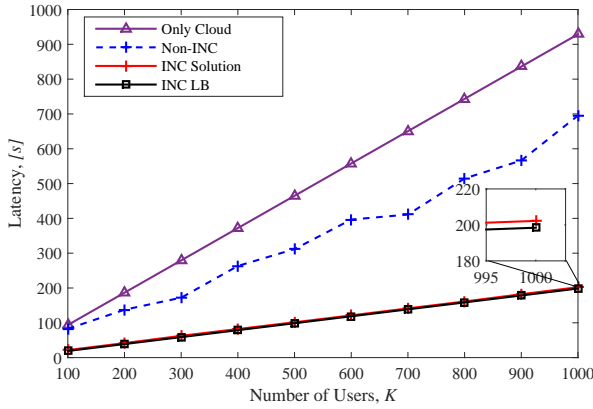


Fig. 4. Algorithm comparison with respect to different number of users.

node’s coverage contains all mobile users. For each edge node m , we set the the uplink fronthaul capacity to $B_m^{fr} = 1\text{Gbps}$, the backhaul capacity to $B_m^{bk} = 1\text{Gbps}$. These settings are inspired by Wifi IEEE 802.11ac standards [15], and data centers interconnection using optical fibers [16]. We also set the cloud downlink capacities $W^d = 2\text{Gbps}$. These values may be changed during the evaluations. For model aggregation, by default, we investigate our system using ResNet152’s model size, i.e., $D = 232\text{ MB}$ [17]. In later simulations, we also investigate our system with different model sizes.

A. Algorithm Comparison - Latency Reduction

Fig. 4 compares the aggregation latency of different algorithms versus the number of users K in one learning iteration. The proposed INC protocol is compared with three other baseline methods, namely:

1. *Only Cloud*: K users send their models to the cloud node via its hypothetical uplink wireless channel with $W^u = 2\text{Gbps}$.
2. *INC Solution*: K users can associate with the cloud node and edge nodes with INC protocol. The network routing problem \mathcal{P}_3 is solved by using Algorithm 1.
3. *Non-INC*: In this scenario, without the proposed INA process, K users are associated with their nearest edge

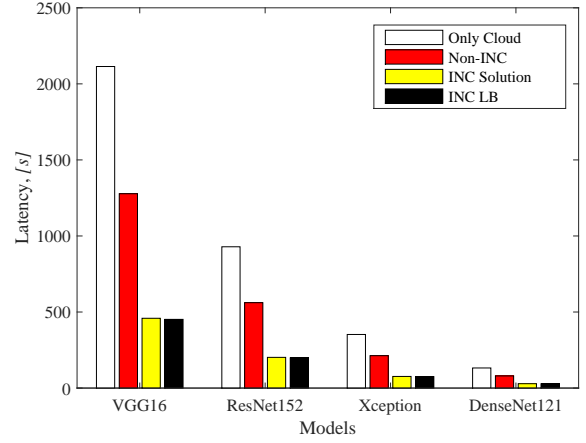


Fig. 5. Algorithm comparison with respect to different models.

nodes regardless of their capacities. The latency between edge node and cloud node is computed by following (5).

4. *INC LB*: In this scenario, we use Linear Relaxation to solve \mathcal{P}_3 . This scenario will provide the lower bound of network latency if the proposed INC protocol is considered.

As can be observed in Fig. 4, our proposed algorithm can achieve near optimal performance. When $K = 1000$, the latency obtained by the proposed solution is approximately 1.9% higher than that of the *INC LB*. It implies that our proposed solution can achieve the performance almost the same as that of the lower bound solution. *Only Cloud* has the worst performance. For example, when $K = 1000$, the aggregation latency of *Only Cloud* is 928.9s which is 133% higher than that of the second worst one, *Non-INC*, 695.3s. We also observe that when $K = 1000$, *Only Cloud* and *Non-INC* are 459% and 343% higher than that of our proposed solution, i.e., *INC Solution*. Last but not least, the gaps between our proposed algorithm and *Only Cloud* and *Non-INC* enlarge as the number of users K increases. This clearly shows that our proposed solution is significantly beneficial for very large scale federated learning networks.

In Fig. 5, consider *Only Cloud*, *Non-INC* and our *INC solution*, we evaluate the aggregation latency in different models in one learning iteration. They are VGG16, ResNet152, Xception and DenseNet121 whose model sizes are 528 MB, 232 MB, 88 MB and 33 MB, respectively [17]. Here, we choose the default setting with $K = 1000$. We observe that with different models, the proposed solution, *INC solution*, is significantly lower than those of the *Only Cloud* and *Non-INC*. For example, with VGG16, the aggregation latency of *INC solution* is 4.6 times and 2.7 times lower than those of *Only Cloud* and *Non-INC*, respectively.

B. Traffic and Computation Reduction at the Cloud Node

In this part, using ResNet152’s model setting, we investigate the uplink traffic and the number of models needed to be aggregated at the cloud node in one learning iteration. We compare three schemes: *Only Cloud*, *Non-INC* and our *INC solution*. The number of models needed to be aggregated at

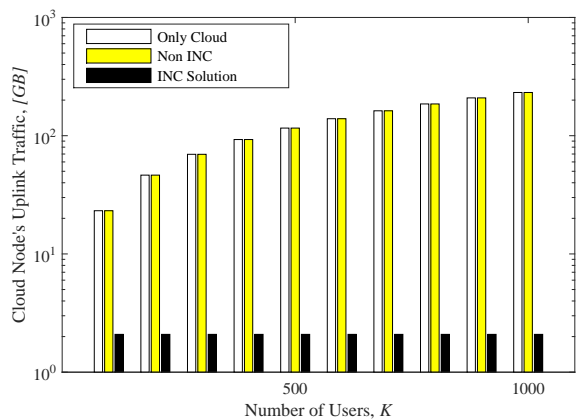


Fig. 6. Cloud node's uplink traffic and computing load.

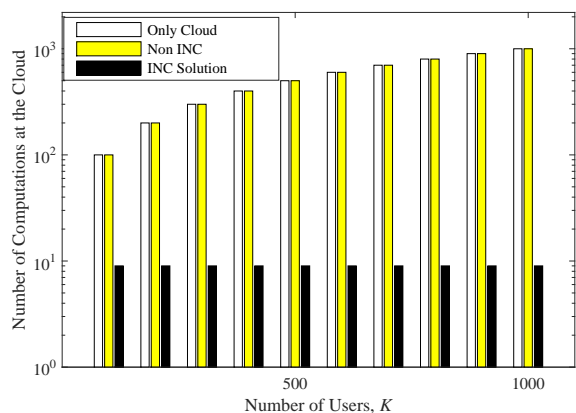


Fig. 7. Cloud node's uplink traffic and computing load.

the cloud node is proportional to the number of computations here. In Fig. 6 and Fig. 7, the uplink traffic and the number of computations of *Non-INC* at the cloud node are equal to those of *Only Cloud*. It is because all models need to be sent to the cloud before being aggregated and edge nodes only forward the models from users to the cloud without the proposed INA process. Meanwhile, with *INC solution*, the two metrics are significantly reduced by remaining unchanged at low value. The reason is that the cloud only collects aggregated models from edge nodes which are fixed. For example, when $K = 1000$, the traffic is 2.32GB for our scheme and 232GB for the other two. As a result, our scheme can keep the traffic and computing load at the cloud very low even with a large number of users.

VII. CONCLUSION

In this paper, we propose a novel edge network architecture aiming at minimizing the aggregation latency of FL processes. This architecture is able to decentralize the model aggregation process of cloud node to edge nodes. To achieve that network functionality, we design an in-network computation protocol consisting of in-network aggregation process and routing protocol. The in-network aggregation process is to enhance learning processes through leveraging computations

at the edges and cloud. We also formulate a joint routing and resource allocation optimization problem to minimize the network's aggregation latency. As the optimization problem is NP-Hard, we propose a highly-effective solution based on random rounding with provable performance guarantee. Our simulation results show that the proposed algorithm can achieve near optimal network latency and outperform some other baseline schemes such as *Only Cloud*, *Non-INC LB*. We also show that the INC protocol can help the cloud node significantly decrease not only its network latency but also its traffic load and computing load.

REFERENCES

- [1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.
- [2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 54, Apr. 2017, pp. 1273–1282.
- [3] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y. C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, Third Quarter 2020.
- [4] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE INFOCOM*, Paris, France, Jun. 2019, pp. 1387–1395.
- [5] Wojciech Samek and Deniz Gunduz, "Distributed deep learning: Concepts, methods & applications in wireless networks," 2020, IEEE GLOBECOM 2020 Tutorial. [Online]. Available: <http://www.federated-ml.org/tutorials/globecom2020/part2.pdf>
- [6] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-network aggregation techniques for wireless sensor networks: a survey," *IEEE Wireless Commun.*, vol. 14, no. 2, pp. 70–87, Apr. 2007.
- [7] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.
- [8] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: A system for large-scale graph processing," in *Proc. ACM SIGMOD*, Indianapolis, Indiana, USA, 2010, pp. 135–146.
- [9] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. K. Gunda, and J. Currey, "DryadLINQ: A system for general-purpose distributed data-parallel computing using a high-level language," in *Proc. USENIX Conf. Oper. Syst. Des. Implement. (OSDI)*, Dec. 2008, pp. 1–14.
- [10] L. Liu, J. Zhang, S. H. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *Proc. IEEE ICC*, Dublin, Ireland, Jul. 2020, pp. 1–6.
- [11] M. Chen and Y. Hao, "Task offloading for mobile edge computing in software defined ultra-dense network," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 587–597, Mar. 2018.
- [12] K. Poularakis, J. Llorca, A. M. Tulino, I. Taylor, and L. Tassiulas, "Service placement and request routing in mec networks with storage, computation, and communication constraints," *IEEE/ACM Trans. Netw.*, vol. 28, no. 3, pp. 1047–1060, Jun. 2020.
- [13] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, Apr. 2010.
- [14] R. Motwani and P. Raghavan, "Randomized algorithms," *ACM Comput. Surveys*, vol. 28, no. 1, pp. 33–37, 1996.
- [15] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, "A tutorial on IEEE 802.11ax high efficiency WLANs," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 197–216, Sep. 2019.
- [16] Y. Cheng, "Optical interconnects for next generation datacenters," Ph.D. dissertation, KTH Royal Institute of Technology, 2019.
- [17] Keras. Keras applications. [Online]. Available: <https://keras.io/api/applications/>