

# Joint Optimization of Energy Consumption and Completion Time in Federated Learning

Xinyu Zhou<sup>1</sup>, Jun Zhao<sup>2</sup>, Huimei Han<sup>2</sup>, Claude Guet<sup>3</sup>

<sup>1</sup>School of Computer Science & Engineering and ERI@N, Nanyang Technological University, Singapore

<sup>2</sup>School of Computer Science & Engineering, Nanyang Technological University, Singapore

<sup>3</sup>School of Physical & Mathematical Sciences, Nanyang Technological University, Singapore  
xinyu003@e.ntu.edu.sg, junzhao@ntu.edu.sg, huimei.han@ntu.edu.sg, cguet@ntu.edu.sg

**Abstract**—Federated Learning (FL) is an intriguing distributed machine learning approach due to its privacy-preserving characteristics. To balance the trade-off between energy and execution latency, and thus accommodate different demands and application scenarios, we formulate an optimization problem to minimize a weighted sum of total energy consumption and completion time through two weight parameters. The optimization variables include bandwidth, transmission power and CPU frequency of each device in the FL system, where all devices are linked to a base station and train a global model collaboratively. Through decomposing the non-convex optimization problem into two subproblems, we devise a resource allocation algorithm to determine the bandwidth allocation, transmission power, and CPU frequency for each participating device. We further present the convergence analysis and computational complexity of the proposed algorithm. Numerical results show that our proposed algorithm not only has better performance at different weight parameters (i.e., different demands) but also outperforms the state of the art.

**Index Terms**—Federated learning, FDMA, resource allocation.

## I. INTRODUCTION

To create a model with high accuracy, machine learning (ML) needs a large size of high-quality training data. Due to limitations of user privacy, data security and competition among industries, usually the data is limited. Besides, the concern around data privacy is rising. Related laws and regulations are appearing, such as the General Data Protection Regulation (GDPR) of the European Union (EU), so it becomes more challenging to collect sensitive data together for learning a model. Therefore, federated learning (FL) could be a solution to collaboratively train an ML model while preserving data on its local device distributedly [1]. FL is a special distributed learning framework, which was proposed by Google in [2]. The algorithm in [2] was called FedAvg. In FedAvg, each device trains a model locally for a certain number of local iterations and uploads their parameters to a base station periodically to build a shared global model.

**Challenges and motivation.** On account of the huge number of parameters in prevalent ML models, FL has

high requirements of device memory, computational capability and communication bandwidth. The limited communication and computation resources in the real world bring several challenges to the deployment of FL: (1) Limited bandwidth results in long *latency* between clients and the server and consequently affects the convergence time of the global model. (2) A large amount of *energy* is needed because a satisfactory model requires plenty of local computation and communication rounds. (3) Due to the large size of current ML models, low-battery devices with limited computation resources have difficulties in joining an FL framework [3]. To improve the efficiency of FL, it is crucial to save both communication and computation resources, which can be achieved by reducing communication overhead [4], [5], [6] or finding optimal resource (e.g., transmission power, bandwidth, CPU frequency, etc.) allocation strategies [7], [3], [8]. This paper focuses on the latter solution. An optimal resource allocation strategy will allow us to reduce the execution latency and energy consumption, and accelerate the convergence of the global model. Therefore, it is essential to consider how to adjust bandwidth, transmission power and CPU clock frequency for each participating device to jointly optimize energy consumption and completion time (i.e., delay). Moreover, different scenarios have different energy and latency demands. For example, connected autonomous vehicles in a smart transportation system require low-latency control decisions, while personal computers may tolerate higher latency to utilize the idle time to train their model. Thus, it raises the following question: how to allocate the transmit power, bandwidth and CPU frequency to each user device to find the optimal trade-off between energy consumption and the total completion time of the FL system while accommodating to different scenarios?

Motivated by the above question, we consider a basic FL system via frequency-division multiple access (FDMA) and formulate a joint optimization of energy and delay problem over a wireless network. The objective function includes a weighted sum of the total energy consumption and the total completion time. Given a certain weight for a specific application scenario, we can find a trade-off between energy and execution latency. The participating devices in FL are under power and bandwidth constraints while pursuing as little energy and time consumption as possible.

Jun Zhao is supported in part by Nanyang Technological University (NTU) Startup Grant, and Singapore Ministry of Education Academic Research Fund under Grant Tier 2 MOE2019-T2-1-176. Huimei Han is supported in part by National Natural Science Foundation of China under Grant 62001419. Corresponding Author: Jun Zhao.

**Contributions.** In summary, the main contributions of this paper are listed as follows:

- We introduce the joint optimization of energy and delay framework into the design of a federated learning system via FDMA. The objective is to optimize a weighted sum of energy consumption and total completion time so that it can satisfy diverse requirements.
- The optimization problem includes a sum-of-ratio form of communication energy consumption, which is non-convex and cannot be solved through traditional optimization techniques. To tackle this problem, we first propose to decompose the original problem into two subproblems: the first subproblem is to minimize computation energy and completion time, and the second is to minimize the communication energy. The first one can be solved through its Lagrangian dual problem. The second subproblem is a sum-of-ratio minimization problem, and we transform it into a convex subtractive-form problem through a Newton-like method similar to [9] used for solving a general fractional programming problem. Convergence analysis and time complexity are provided.
- Numerical results show: 1) at different weight parameters (i.e., different requirements), our resource allocation algorithm is able to find a corresponding solution to jointly minimize energy consumption and completion time. 2) Our joint optimization framework’s performance is better than the communication-only optimization and computation-only optimization framework. 3) Our resource allocation scheme outperforms the state of the art, especially when there is a tight delay constraint.

The rest of this paper is organized as follows. Section II presents related works. In Section III, we introduce the FL model and the necessary notations. Problem formulation is provided in Section IV. We detail the proposed algorithm in Section V. Then Section VI provides the convergence and the computational complexity of the algorithm. Numerical results are shown in Section VII. Finally, Section VIII concludes the paper.

## II. RELATED WORK

In this section, we survey related research. Section II-A as the first subsection is devoted to federated learning (FL) over wireless networks which our paper is also about. Then in Section II-B, we elaborate on other FL studies. Finally, since our paper tackles resource allocation and appears in ICDCS, Section II-C presents related works on resource allocation in other distributed systems.

### A. Federated Learning over Wireless Networks

For simplicity, below we use WFL, short for wireless federated learning, to describe FL over wireless networks. In many papers [10], [8], [11], [12], [3], the base station (i.e., the aggregator) is also seen as an edge server, so the literature may also use the term federated edge learning (FEEL). Since our paper minimizes the weighted sum of two metrics: energy

consumption and delay, our coverage about WFL next is categorized into energy minimization, delay minimization, optimization of other metrics, weighted optimization, and non-optimization studies, respectively. Note that papers [8], [13], [3] minimizing the weighted sum of energy and delay are elaborated in the part about weighted optimization. During the discussion, we also compare our work with many studies including [7], [14], [10], [12], [15], [8], [13], [3].

**Minimizing the energy consumption in WFL.** Three representative papers [7], [12], [10] minimize the energy consumption only in WFL are explained as follows.

Yang *et al.* [7] minimize the sum of all devices’ energy consumption, which includes the local computation energy and the wireless transmission energy. In [7], devices access the wireless transmission media via FDMA, as in our paper. The optimization variables in [7] include the local training accuracy and each device’s transmission time per global round, bandwidth allocation, CPU frequency and transmit power. The constraints cover an upper bound on the total delay and a lower bound on each device’s transmission time per global round such that the local model can be uploaded.

The **major difference between [7] and our work** is that the optimization objective is just the energy consumption in [7], but is the weighted sum of the energy and delay in our paper. The consideration of both energy and delay in the objective function makes our optimization much more difficult to solve than [7], as explained below. Since we put the total completion time into the objective function, a part of our optimization problem (i.e., Subproblem 2 in (11) later) becomes an NP-complete problem—the sum-of-ratios problem [9]. While addressing a weaker optimization objective than our work, [7] has two more optimization variables than ours: the local training accuracy and each device’s transmission time per global round. Yet, since both two variables can be decoupled from others, their optimal values can be easily obtained as in Theorem 2 of [7]. The rest variables of [7] are the same as ours. However, if just minimizing the total energy consumption, it will be easy to decouple the optimization problem into two subproblems as [7] does. Since all subproblems in [7] are convex, [7] can easily obtain the optimal solution.

Different from [7] and our paper which both consider only CPU computing resources at devices, Li *et al.* [10] and Zeng *et al.* [12] incorporate devices’ GPU computation for WFL. In particular, [10] studies GPU only, while [12] uses CPU and GPU simultaneously. The energy consumption model for GPU is different from that of CPU used in our paper, so we omit the detailed discussions of [10], [12].

**Minimizing the delay in WFL.** Below we discuss three work [14], [15], [3] on delay minimization only in WFL.

Minimizing the delay by Yang *et al.* [14] is actually used to provide an initial feasible solution for the problem of minimizing the energy consumption subject to the delay constraint in [7]. In other words, the work done in [14] is a subroutine of that in [7]. We have discussed above how our optimization is more challenging to solve than that of [7], so our work is further more difficult than [14]. In [14], the FDMA bandwidth

allocation is not analyzed directly, but a bisection method is used to sandwich the minimum delay.

Distinct from [14] considering FDMA just like its extended version [7], the research [15] by Chen *et al.* employs orthogonal frequency-division multiple access (OFDMA) for delay minimization in WFL. In OFDMA, the radio is divided into 2-dimensional resource blocks (RBs) over time and frequency. The optimization variables in [15] include device selection matrices and RB allocation matrices, which all have binary elements. Hence, [15] handles binary integer programming, while the variables in [14] and our paper are all continuous.

**Optimizing other metrics in WFL.** Other metrics in WFL have also been optimized, such as the number of global aggregation rounds [16], the total number of bits sent from the devices to the aggregator [17], the admission data rates of all devices [18], the communication error [19], the total communication cost [20] with each communication link being assigned a cost number, the training loss [11], [21], [22], the learning efficiency [23], [24], and a metric capturing the long-term performance of WFL [25].

**Weighted optimization in WFL.** Below we review three work [8], [13], [3] on the weighted minimization of energy consumption and delay in WFL.

**The major difference between [8] and our paper** is that the channel access method in [8] for devices is time-division multiple access (TDMA), while our paper adopts FDMA. In addition, although the optimization variables in [8] also include parameters capturing local training and global training progresses respectively, the two parameters are optimized just by a numerical method: an exhaustive grid search. Other optimization variables in [8] include each device's CPU frequency and transmission time per global round (we do not discuss additional variables listed in [8] since they can be decided from variables already explained).

In [13], the optimization variables include the number of local iterations per global round, the number of global rounds, and the number of devices selected (via uniform sampling) per global round. Hence, [13] tackles integer programming, while the variables in [8] and our paper are all continuous.

In contrast to [8], [13] which have only a two-level hierarchy: the base station (often also an edge server) and the devices, [3] adopts a three-level hierarchy (introduced also by [26]): an cloud server, multiple edge servers, and the devices. **A detailed comparison between [3] and our paper is as follows.** Firstly, the optimization variables in both work include each device's bandwidth allocation via FDMA and CPU frequency, while [3] also considers edge association of devices, and our paper also optimizes each device's transmission power. Secondly, in our work, a challenge in solving the optimization arises from the coupling between the FDMA bandwidth allocation and the weighted minimization of energy and delay. Yet, [3] avoids such challenge by simplifying the Shannon formula for devices' data rates. In particular, in Equation (5) of [3], the noise power as the denominator inside the logarithm of the Shannon formula is forcefully assumed as a constant that does not scale with the allocated

bandwidth, while it should be the noise power spectral density times the bandwidth. Such extreme simplification in [3] makes the bandwidth allocation easy to solve. In this paper, we use the exact expression of the Shannon formula and hence run into a much more challenging optimization.

The work [27] also considers weighted optimization in WFL, but is not about energy consumption or delay.

**Studies other than optimizing metrics in WFL.** Beyond optimization studies, WFL has also been investigated in terms of scheduling policies [28], [29], [30], and mobility-based orchestration [31].

### B. Other Federated Learning Studies

In addition to federated learning over wireless networks elucidated in the previous subsection, we present other federated learning work in this subsection.

**Communication-efficient FL.** Researchers have enabled FL to be more communication-efficient, by transmitting clustered model updates [32], avoiding irrelevant updates [4], adaptive parameter freezing [6], and sketching [33].

**Security/privacy of FL.** Security of FL has been addressed by defending against Byzantine attacks [34], and enforcing device liability [35]. Privacy of FL is examined in [36], [37].

**Incentives in FL.** Designing incentive mechanisms for FL has been tackled in [38], [39].

**Machine learning approaches to improve FL.** FL systems have also benefited from other machine learning techniques including reinforcement learning [40], meta learning [41], [42], one-shot learning [43], and online learning [5].

**Applications of FL.** FL has been applied to various tasks including crowdsensing [44], [45], [46], RFID clone detection [47], and the classification of unmanned aerial vehicles (UAVs) [48].

### C. Resource Allocation in Other Distributed Systems

Resource allocation has also been studied in other distributed systems, such as blockchain [49], [50], cloud computing [51], [52], edge computing [53], distributed learning [54], [55], and wireless networks [56].

## III. SYSTEM MODEL

The system model is FL over wireless communications. We consider a star network with a base station, as shown in Fig. 1. We suppose that the base station serves  $N$  devices. The set of  $N$  devices is denoted by  $\mathcal{N}$ .

**Federated Learning.** Each device contains a dataset  $\mathcal{D}_n$  with  $D_n$  samples. The total dataset is  $\mathcal{D} = \bigcup_{n=1}^N \mathcal{D}_n$ . We have the local dataset  $\mathcal{D}_n = \bigcup_{i=1}^{D_n} \{\mathbf{x}_{ni}, y_{ni}\}$ .  $\mathbf{x}_{ni} \in \mathbb{R}^p$  is an input vector containing the features of each data sample, where  $\mathbb{R}$  denotes the set of real numbers.  $y_{ni}$  is the label of each data sample. Notations in this paper are listed in Table I. Bold symbols stand for vectors.

Then, the overall FL model is to solve the problem  $\min_{\mathbf{w} \in \mathbb{R}^p} F(\mathbf{w}) = \sum_{n=1}^N \frac{D_n}{D} l_n(\mathbf{w})$ , where  $D = \sum_{n=1}^N D_n$ .

Each local device trains its own model and the aim is to minimize the corresponding loss function defined at device  $n$

TABLE I: Notations

Description	Symbol
Total energy consumption	$E$
Maximum uplink bandwidth	$B$
Allocated bandwidth of device $n$	$B_n$
The number of samples on device $n$	$D_n$
Total number of samples on all devices	$D$
Transmission data size of device $n$	$d_n$
Maximum completion time in one global communication round	$T$
Maximum completion time	$\mathcal{T}$
Data transmission time of device $n$	$T_n^{up}$
Local computation time of device $n$	$T_n^{cmp}$
Data transmission rate of device $n$	$r_n$
Minimum transmission rate of device $n$	$r_n^{min}$
CPU frequency of device $n$	$f_n$
Minimum and maximum CPU frequency of device $n$	$f_n^{min}, f_n^{max}$
The number of global aggregation rounds	$R_g$
The number of local iteration per global round	$R_l$
The effective switched capacitance	$\kappa$
The number of CPU cycles per sample on device $n$	$c_n$
Local computation energy consumption of device $n$	$E_n^{cmp}$
Transmission energy consumption of device $n$	$E_n^{trans}$
The number of total devices	$N$
Weight parameters	$w_1, w_2$
Noise power spectral density	$N_0$
Transmission power of device $n$	$p_n$
Minimum and maximum transmission power of device $n$	$p_n^{min}, p_n^{max}$
Channel gain	$g$
Lagrangian multipliers	$\lambda, \tau, \mu$
Auxiliary variables	$\nu, \beta$

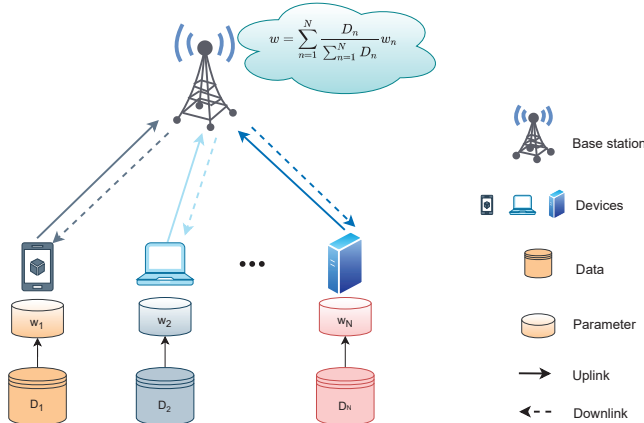


Fig. 1: Federated learning.

as  $l_n(\mathbf{w}) = \sum_{i=1}^{D_n} \frac{f_{ni}(\mathbf{w})}{D_n}$ , where  $\mathbf{w}$  is the parameters related to the whole global model. The function  $f_{ni}(\mathbf{w})$  measures the error between the model prediction and the ground truth label. Note that in each local iteration, each device  $n$  uses all of its  $D_n$  data samples, not just a subset.

As shown in Fig. 1, each device trains its own model locally for a certain number of iterations and then uploads its model parameter to the base station. The base station uses all the uploaded parameters to calculate the weighted average model parameter and broadcasts it to each device, where the weight for each device  $n$ 's parameter is  $\frac{D_n}{D}$ ; i.e., the fraction of device  $n$ 's number of data points over the total number of data points. The above uploading and broadcasting process is called global communication.

**FDMA.** In this paper, different devices communicate with the base station via FDMA. FL over FDMA has also been studied in [14], [10], [7], [3] (see Section II for detailed comparisons between these work and ours). FDMA is simple to implement even for mobile devices with limited computation capability [10]. We do not consider OFDMA since our optimization problem with FDMA is already difficult to tackle. Although [15] adopts OFDMA for FL, but it addresses delay minimization only and handles binary integer programming, while the variables in our paper are all continuous. In addition, TDMA over FL is investigated in [8], [23].

We will formulate a joint optimization of energy and completion time problem in FL via FDMA. The objective function is a weighted sum of total energy and time. Total energy consumption and total completion time are discussed in Sections III-A and III-B, respectively.

### A. Total Energy Consumption

The **total energy consumption**,  $E$ , comprises wireless network **transmission energy consumption** and **local computation energy consumption**. Following works [8], [3], [7], we do not consider the energy consumption of the base station (or the central server), either.

1) *Transmission energy consumption:* Generally, transmission time refers to the uplink time and the downlink time. However, since the output power of the base station is significantly larger than the maximum uplink transmission power of a user device, the downlink time is much smaller than the uplink time. Thus, we neglect the downlink time.

From the Shannon formula, we express the data transmission rate,  $r_n$ , as

$$r_n = B_n \log_2 \left( 1 + \frac{g_n p_n}{N_0 B_n} \right), \quad (1)$$

where  $B_n$  is the bandwidth allocated to user  $n$ ,  $N_0$  is the power spectral density of Gaussian noise,  $g_n$  is the channel gain between user device  $n$  and the base station and  $p_n$  is the transmission power. Note that there is no interference in FDMA since the channels assigned to users are separated by different frequencies.

Strictly speaking, the expression (1) describes just the maximum theoretical bit rate achievable in a wireless channel. The actual data rate depends on the spectral efficiency of the communication technology. Yet, we use (1) as the data rate for simplicity as in much prior work [14], [10], [7], [8], [23] on FL. Also, if “=” in (1) is replaced by “ $\leq$ ”, “=” will still be taken in our optimization problem to save the transmission time and hence the transmission energy.

The data transmission will happen in each global aggregation. If the transmission data size of user device  $n$  is  $d_n$ , the transmission time of user device  $n$  is

$$T_n^{up} = d_n / r_n. \quad (2)$$

Then, the **transmission energy consumption** of user device  $n$  is

$$E_n^{trans} = p_n T_n^{up}. \quad (3)$$

2) *Local computation energy consumption*: Following [8], and assuming that the CPU frequency  $f_n$  of device  $n$  does not change as time goes once it is optimized, we express the energy,  $E_n^{cmp'}$ , consumed in one **local** iteration of user device  $n$  as:

$$E_n^{cmp'} = \kappa c_n D_n f_n^2, \quad (4)$$

where  $c_n$  is the number of CPU cycles per sample on device  $n$ , and  $\kappa$  is the effective switched capacitance.

The **local computation energy consumption in one global iteration** is

$$E_n^{cmp} = \kappa R_l c_n D_n f_n^2, \quad (5)$$

where  $R_l$  is the number of local iterations and the total energy consumption is

$$E = R_g \sum_{n=1}^N (E_n^{trans} + E_n^{cmp}), \quad (6)$$

where  $R_g$  is the number of global rounds.

### B. Total Completion Time

The total completion time (i.e., the total delay),  $T$ , of the FL system model over wireless network comprises local computation time and data transmission time. The local computation time of device  $n$  in one **global** iteration,  $T_n^{cmp}$ , is

$$T_n^{cmp} = R_l \frac{c_n D_n}{f_n}. \quad (7)$$

Then with the transmission time given by  $T_n^{up} = \frac{d_n}{r_n}$ , the total completion time  $T$  becomes  $R_g \max_{n \in \mathcal{N}} \{T_n^{cmp} + T_n^{up}\}$ .

As we have mentioned, since the base station's output power is larger than the maximum uplink transmission power of a user device, the downlink time is much smaller than the uplink time. Hence, we ignore the downlink time.

## IV. OPTIMIZATION PROBLEM FORMULATION

To minimize the total energy consumption as well as the total completion time of the whole FL training process, we use a weighted sum method to formulate the optimization problem. As  $E_n^{cmp}$  in (5),  $E_n^{trans}$  in (3),  $T_n^{up}$  in (2) and  $T_n^{cmp}$  in (7) are highly related to  $p_n$ ,  $B_n$  and  $f_n$ , we choose  $p_n$ ,  $B_n$  and  $f_n$  as three optimization variables. Some works may consider more optimization variables such as the global communication rounds  $R_g$ , the local computations rounds  $R_l$  or the model accuracy. However,  $R_g$ ,  $R_l$  and the model accuracy are determined by the exact ML model run on each device. To make our optimization problem apply to flexible scenarios and broad use cases, we do not consider more optimization variables.

The optimization problem is as follows:

$$\min_{p_n, f_n, B_n} w_1 E + w_2 R_g \max_{n \in \mathcal{N}} \{T_n^{cmp} + T_n^{up}\}, \quad (8)$$

subject to:

$$p_n^{min} \leq p_n \leq p_n^{max}, \quad \forall n \in \mathcal{N}, \quad (8a)$$

$$f_n^{min} \leq f_n \leq f_n^{max}, \quad \forall n \in \mathcal{N}, \quad (8b)$$

$$\sum_{n=1}^N B_n \leq B, \quad (8c)$$

with  $w_1, w_2 \in [0, 1]$  denoting two weight parameters of total energy consumption and completion time and  $w_1 + w_2 = 1$ . If devices are in a low-battery condition, we can set  $w_1 = 1$  and  $w_2 = 0$  to allocate resources in the FL system. In contrast, if the FL system is time-sensitive, the weight parameters can be set as  $w_1 = 0$  and  $w_2 = 1$ .  $E$  is defined in (6). Constraint (8a) and constraint (8b) set the ranges of the transmission power and computation frequency of each device. Constraint (8c) refers to the uplink bandwidth constraint.

Note that the max function in (8) makes the optimization hard to solve. To circumvent this difficulty, we replace the max function by an auxiliary variable  $\mathcal{T}$  with an additional constraint; then, the problem writes as:

$$\min_{p_n, f_n, B_n, \mathcal{T}} w_1 R_g \sum_{n=1}^N (p_n T_n^{up} + \kappa R_l c_n D_n f_n^2) + w_2 R_g \mathcal{T}, \quad (9)$$

subject to:

$$(8a), (8b), (8c),$$

$$T_n^{cmp} + T_n^{up} \leq \mathcal{T}, \quad \forall n \in \mathcal{N}. \quad (9a)$$

Constraint (9a) is the total time constraint.

**Difficulty of solving problem (9)**. Since the formulation of  $T_n^{up}$  and constraint (9a) are not convex, problem (9) is non-convex and thus can be difficult to solve. Consequently, we must explore other ways to transform this problem to make it solvable. Additionally, the complex problem formulation increases the difficulty of transforming this problem.

### A. Problem Decomposition

To continuously transform problem (9) into a more simple form, we must decompose the problem. Because  $p_n$  and  $B_n$  are in the same term, we thereby separate problem (9) into two subproblems. One has  $p_n$  and  $B_n$  as optimization variables and the other has  $f_n$  and  $\mathcal{T}$  as optimization variables.

Two minimization problems are as follows:

a) *Subproblem 1*:

$$\min_{f_n, \mathcal{T}} w_1 R_g \sum_{n=1}^N \kappa R_l c_n D_n f_n^2 + w_2 R_g \mathcal{T} \quad (10)$$

subject to: (8b), (9a).

b) *Subproblem 2*:

$$\min_{p_n, B_n} w_1 R_g \sum_{n=1}^N p_n \frac{d_n}{B_n \log_2(1 + \frac{p_n g_n}{N_0 B_n})} \quad (11)$$

subject to: (8a), (8c), (9a).

Our strategy is to randomly choose feasible  $(\mathbf{p}, \mathbf{B})$  at first, and solve Subproblem 1 to get the optimal  $\mathbf{f}$ . Then, we solve Subproblem 2 to get the optimal  $(\mathbf{p}, \mathbf{B})$ . The whole process consists of iteratively solving Subproblems 1 and 2 until convergence.

## V. SOLUTION OF THE JOINT OPTIMIZATION PROBLEM

In this section, we will explain how to solve Subproblems 1 and 2.

### A. Solution to Subproblem 1

It is easy to verify that the function of variables  $f_n$  and  $\mathcal{T}$  in Eq. (10) is convex. To solve a convex optimization problem, a typical approach is to use the Karush–Kuhn–Tucker (KKT) conditions. For reader's convenience, we briefly describe KKT below while more details can be found in Section 5.5.3 of a book by Boyd et al. [57]. For a convex optimization problem, as long as its objective function and constraints are differentiable, and it satisfies Slater's condition<sup>1</sup>, KKT conditions are sufficient and necessary conditions to get the optimal solution.

Then, we treat the optimization problem by the KKT approach and introduce the Lagrange function as follows:

$$L_1(f_n, \mathcal{T}, \lambda) = w_1 R_g \sum_{n=1}^N \kappa R_l c_n D_n f_n^2 + w_2 R_g \mathcal{T} + \sum_{n=1}^N \lambda_n \left[ \left( \frac{R_l c_n D_n}{f_n} + T_n^{up} \right) - \mathcal{T} \right], \quad (12)$$

where  $\lambda$  is the multiplier associated with the inequality constraint (9a). After applying KKT conditions, we get

$$\frac{\partial L_1}{\partial f_n} = 2w_1 R_g R_l \kappa c_n D_n f_n - \lambda_n \frac{R_l c_n D_n}{f_n^2} = 0, \quad \forall n \in \mathcal{N}, \quad (13)$$

$$\frac{\partial L_1}{\partial \mathcal{T}} = w_2 R_g - \sum_n \lambda_n = 0, \quad (14)$$

$$\lambda_n \left[ \left( \frac{R_l c_n D_n}{f_n} + T_n^{up} \right) - \mathcal{T} \right] = 0, \quad (15)$$

from which we derive

$$f_n^* = \sqrt[3]{\frac{\lambda_n}{2w_1 R_g \kappa}}, \quad \text{and} \quad w_2 = \frac{\sum_n \lambda_n}{R_g}. \quad (16)$$

Then, the dual problem is

$$\max_{\lambda_n} \sum_{n=1}^N (2^{-\frac{2}{3}} + 2^{\frac{1}{3}}) h c_n D_n \lambda_n^{\frac{2}{3}} + T_n^{up} \lambda_n \quad (17)$$

subject to,

$$\sum_{n=1}^N \lambda_n = w_2 R_g, \quad (17a)$$

$$\lambda_n \geq 0, \quad (17b)$$

where  $h = R_l (w_1 \kappa R_g)^{\frac{1}{3}}$ . Obviously, this dual problem is a simple convex optimization problem. In this paper, we use CVX [58] to solve it and get the optimal  $\lambda^* = [\lambda_1^*, \dots, \lambda_n^*]$ . Then, we are able to calculate  $\mathbf{f}^*$  through equation (16) and get  $\mathbf{f}$  in the below,

$$\mathbf{f} = \min(\mathbf{f}^{min}, \max(\mathbf{f}^*, \mathbf{f}^{min})). \quad (18)$$

<sup>1</sup>In brief, Slater's condition states that the optimal solution must be interior of the feasible region.

### B. Subproblem 2

Subproblem 2 is a typical minimization sum-of-ratios problem and thus very hard to solve. Firstly, we define

$$G_n(p_n, B_n) := B_n \log_2 \left( 1 + \frac{p_n g_n}{N_0 B_n} \right) \quad (19)$$

to simplify the notations. Additionally, (9a) can be written as  $r_n \geq r_n^{min}$ , where  $r_n^{min} = \frac{d_n}{\mathcal{T} - \frac{R_l c_n D_n}{f_n}}$ . In fact,  $G_n(p_n, B_n)$  is the same as the data rate  $r_n$  defined in (1).

Then, we transform Subproblem 2 into the epigraph form through adding an auxiliary variable  $\beta_n$  and let  $\frac{p_n d_n}{G_n(p_n, B_n)} \leq \beta_n$ . Then, we get an equivalent problem called **SP2\_v1**.

$$\mathbf{SP2\_v1}: \min_{p_n, B_n, \beta_n} w_1 R_g \sum_{n=1}^N \beta_n \quad (20)$$

subject to,

$$(8a), (8c),$$

$$G_n(p_n, B_n) \geq r_n^{min}, \quad \forall n \in \mathcal{N}, \quad (20a)$$

$$p_n d_n - \beta_n G_n(p_n, B_n) \leq 0, \quad \forall n \in \mathcal{N}, \quad (20b)$$

where  $\beta_n$  is an auxiliary variable. Because  $\frac{p_n d_n}{G_n(p_n, B_n)} \leq \beta_n$  and  $G_n(p_n, B_n) > 0$ , we can get constraint (20b).

However, due to constraint (20b), problem **SP2\_v1** is non-convex, which is still intractable. To continue transforming this problem, we need to give the following lemma.

**Lemma 1.**  $G_n(p_n, B_n)$  is a concave function.

*Proof.* See Appendix A.  $\square$

The original objective function in Subproblem 2 is the sum of fractional functions. It is obvious that  $p_n d_n$  is convex, and Lemma 1 reveals  $G_n(p_n, B_n)$  is concave. With such trait, problem **SP2\_v1** can be transformed into a subtractive-form problem via the following theorem.

**Theorem 1.** If  $(\mathbf{p}^*, \mathbf{B}^*, \beta^*)$  is the solution of problem **SP2\_v1**, there exists  $\nu^*$  satisfying that  $(\mathbf{p}^*, \beta^*)$  is a solution of the following problem with  $\nu = \nu^*$  and  $\beta = \beta^*$ .

$$\mathbf{SP2\_v2}: \min_{p_n, B_n} \sum_{n=1}^N \nu_n (p_n d_n - \beta_n G_n(p_n, B_n)) \quad (21)$$

subject to: (8a), (8c), (20a).

Moreover, under the condition of  $\nu = \nu^*$  and  $\beta = \beta^*$ ,  $(\mathbf{p}^*, \mathbf{B}^*)$  satisfies the following equations.

$$\nu_n^* = \frac{w_1 R_g}{G_n(p_n^*, B_n^*)}, \quad n = 1, \dots, N, \quad (22)$$

$$\text{and } \beta_n^* = \frac{p_n^* d_n}{G_n(p_n^*, B_n^*)}, \quad n = 1, \dots, N. \quad (23)$$

*Proof.* The proof follows from Lemma 2.1 in [9].  $\square$

### C. Solution to Subproblem 2

**Theorem 1** indicates that problem **SP2\_v2** has the same optimal solution  $(\mathbf{p}^*, \mathbf{B}^*)$  with problem **SP2\_v1**. Consequently, we can first provide  $(\nu, \beta)$  to obtain a solution  $(p_n, B_n)$  through solving problem **SP2\_v2**. Then, the second step is to calculate  $(\nu, \beta)$  according to (22) and (23). The first step

can be seen as an inner loop to solve problem **SP2\_v2** and the second step is the outer loop to obtain the optimal  $(\boldsymbol{\nu}, \boldsymbol{\beta})$ . The iterative optimization algorithm is given in Algorithm 1, which is a Newton-like method. The original algorithm is provided in [9]. In Algorithm 1, we define

$$\phi_1(\boldsymbol{\beta}) = [-p_1 d_1 + \beta_1 G_1(p_1, B_1), \dots, -p_N d_N + \beta_N G_N(p_N, B_N)]^T, \quad (24)$$

$$\phi_2(\boldsymbol{\nu}) = [-w_1 R_g + \nu_1 G_1(p_1, B_1), \dots, -w_1 R_g + \nu_N G_N(p_N, B_N)]^T, \quad (25)$$

which are two vectors in  $\mathbb{R}^N$ , and

$$\phi(\boldsymbol{\beta}, \boldsymbol{\nu}) = [\phi_1^T(\boldsymbol{\beta}), \phi_2^T(\boldsymbol{\nu})]^T, \quad (26)$$

which is a vector in  $\mathbb{R}^{2N}$ . Additionally, the Jacobian matrices of  $\phi_1(\boldsymbol{\beta})$  and  $\phi_2(\boldsymbol{\nu})$  are

$$\phi_1'(\boldsymbol{\beta}) = \text{diag}(G_n(p_n, B_n)|_{n=1, \dots, N}), \quad (27)$$

$$\phi_2'(\boldsymbol{\nu}) = \text{diag}(G_n(p_n, B_n)|_{n=1, \dots, N}), \quad (28)$$

where  $\text{diag}()$  represents a diagonal matrix. [9] proves that when  $\phi(\boldsymbol{\beta}, \boldsymbol{\nu}) = \mathbf{0}$ , the optimal solution  $(\boldsymbol{\nu}, \boldsymbol{\beta})$  is achieved. Simultaneously, it satisfies the system of equations (22) and (23). Thus, a Newton-like method can be used to update  $(\boldsymbol{\nu}, \boldsymbol{\beta})$ , which is done in (31).

---

**Algorithm 1:** Optimization of  $(\boldsymbol{p}, \boldsymbol{B})$  and  $(\boldsymbol{\nu}, \boldsymbol{\beta})$

---

1 Initialize  $i = 0$ ,  $\xi \in (0, 1)$ ,  $\epsilon \in (0, 1)$ . Given feasible  $(\boldsymbol{p}^{(0)}, \boldsymbol{B}^{(0)})$ .

2 repeat

3 Calculate

$$\nu_n^{(i)} = \frac{w_1 R_g}{G_n(p_n^{(i)}, B_n^{(i)}), \beta_n^{(i)} = \frac{p_n d_n}{G_n(p_n^{(i)}, B_n^{(i)})}$$

for  $n = 1, \dots, N$ .

4 Obtain  $(\boldsymbol{p}^{(i+1)}, \boldsymbol{B}^{(i+1)})$  through solving **SP2\_v2** according to **Theorem 2** by using CVX given  $(\boldsymbol{\nu}^{(i)}, \boldsymbol{\beta}^{(i)})$ .

5 Let  $j$  be the smallest integer which satisfy

$$|\phi(\boldsymbol{\beta}^{(i)} + \xi^j \boldsymbol{\sigma}_1^{(i)}, \boldsymbol{\nu}^{(i)} + \xi^j \boldsymbol{\sigma}_2^{(i)})| \leq (1 - \epsilon \xi^j) |\phi(\boldsymbol{\beta}^{(i)}, \boldsymbol{\nu}^{(i)})|, \quad (29)$$

where

$$\begin{aligned} \boldsymbol{\sigma}_1^{(i)} &= -[\phi_1'(\boldsymbol{\beta}^{(i)})]^{-1} \phi_1(\boldsymbol{\beta}^{(i)}), \\ \boldsymbol{\sigma}_2^{(i)} &= -[\phi_2'(\boldsymbol{\nu}^{(i)})]^{-1} \phi_2(\boldsymbol{\nu}^{(i)}). \end{aligned} \quad (30)$$

6 Update

$$(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\nu}^{(i+1)}) = (\boldsymbol{\beta}^{(i)} + \xi^j \boldsymbol{\sigma}_1^{(i)}, \boldsymbol{\nu}^{(i)} + \xi^j \boldsymbol{\sigma}_2^{(i)}). \quad (31)$$

7 Let  $i \leftarrow i + 1$ .

8 until  $\phi(\boldsymbol{\beta}^{(i)}, \boldsymbol{\nu}^{(i)}) = \mathbf{0}$  or the number of iterations reaches the maximum iterations  $i_0$ ;

---

Until now, the rest problem we need to focus on is problem **SP2\_v2**. Remember that  $\boldsymbol{\nu}$  and  $\boldsymbol{\beta}$  are fixed in this problem, which can be seen as constants. Given Lemma 1,  $p_n d_n$  is convex, and three constraints (8a) (8c) (20a) are convex, the

objective function (21) in problem **SP2\_v2** is convex. Thus, KKT conditions are the sufficient and necessary optimality conditions for finding the optimal solution.

Before applying KKT conditions, we write down the partial Lagrangian function of problem **SP2\_v2**:

$$\begin{aligned} L_2(p_n, B_n, \tau_n, \mu) &= \\ &\sum_{n=1}^N \nu_n (p_n d_n - \beta_n B_n \log_2(1 + \frac{p_n g_n}{N_0 B_n})) \\ &- \sum_{n=1}^N \tau_n (B_n \log_2(1 + \frac{p_n g_n}{N_0 B_n}) - r_n^{min}) + \mu (\sum_{n=1}^N B_n - B), \end{aligned} \quad (32)$$

where  $\tau_n|_{n=1, \dots, N}$  and  $\mu$  are non-negative Lagrangian multipliers.

After applying KKT conditions to problem **SP2\_v2**, we get

$$\frac{\partial L_2}{\partial p_n} = \nu_n (d_n - \frac{\beta_n g_n}{N_0 (1 + \vartheta_n) \ln 2}) - \frac{\tau_n g_n}{N_0 (1 + \vartheta_n) \ln 2} = 0, \quad (33)$$

$$\begin{aligned} \frac{\partial L_2}{\partial B_n} &= -(\nu_n \beta_n + \tau_n) \log_2(1 + \vartheta_n) + \frac{(\nu_n \beta_n + \tau_n) p_n g_n}{(1 + \vartheta_n) \ln 2 N_0 B_n} \\ &+ \mu = 0, \end{aligned} \quad (34)$$

$$-\tau_n (B_n \log_2(1 + \vartheta_n) - r_n^{min}) = 0, \quad (35)$$

$$\mu (\sum_{n=1}^N B_n - B) = 0, \quad (36)$$

where  $\vartheta_n = \frac{p_n g_n}{N_0 B_n}$  for  $n = 1, \dots, N$ .

With (33)–(36), **Theorem 2** is given to find the optimal bandwidth and transmission power for each device.

**Theorem 2.** The optimal bandwidth  $\boldsymbol{B}$  and transmission power  $\boldsymbol{p}$  are expressed as

$$B_n^* = \begin{cases} \frac{r_n^{min}}{\log_2(1 + \Lambda_n)}, & \text{if } \tau_n \neq 0, n = 1, \dots, N, \\ \text{Solution to problem (A.6)}, & \text{if } \tau_n = 0, \end{cases} \quad (37)$$

$$p_n^* = \min(p_n^{max}, \max(\Gamma(B_n), p_n^{min})) \quad \text{when } \mu \neq 0, n = 1, \dots, N, \quad (38)$$

where

$$\Lambda_n = \frac{(\nu_n \beta_n + \tau_n) g_n}{N_0 d_n \nu_n \ln 2}, \quad (39)$$

$$\Gamma(B_n) = (\frac{(\nu_n \beta_n + \tau_n) g_n}{N_0 d_n \nu_n \ln 2} - 1) \frac{N_0 B_n}{g_n}. \quad (40)$$

*Proof.* See Appendix B.  $\square$

Then, we have solved problem **SP2\_v2**.

#### D. Resource Allocation Algorithm

Until now, we have already solved Subproblem 1 and Subproblem 2. Here we give the complete resource allocation algorithm in Algorithm 2.

Algorithm 2 first initializes a feasible solution according to the range of power  $\boldsymbol{p}$  and bandwidth  $\boldsymbol{B}$  and the sum of the bandwidth of each device cannot exceed  $B$ . Then, solving problem (17) with given  $(\boldsymbol{p}, \boldsymbol{B})$  can make sure we get an

---

**Algorithm 2:** Resource Allocation Algorithm

---

- 1 Initialize  $sol^{(0)} = (\mathbf{p}^{(0)}, \mathbf{B}^{(0)}, \mathbf{f}^{(0)})$  of problem (9), iteration number  $k = 1$ .
  - 2 **repeat**
  - 3     Solve Subproblem 1 through solving problem (17) with CVX given  $(\mathbf{p}^{(k-1)}, \mathbf{B}^{(k-1)})$ . Obtain  $\mathbf{f}^{(k)}$  according to equation (18).
  - 4     Solve Subproblem 2 through Algorithm 1 and obtain  $(\mathbf{p}^{(k)}, \mathbf{B}^{(k)})$  with CVX.
  - 5      $sol^{(k)} = (\mathbf{p}^{(k)}, \mathbf{B}^{(k)}, \mathbf{f}^{(k)})$ .
  - 6     Set  $k \leftarrow k + 1$ .
  - 7 **until**  $|sol^{(k)} - sol^{(k-1)}| \leq \epsilon_0$  or the number of iterations achieves  $K$ ;
- 

optimal  $\mathbf{f}$  at this step. Subproblem 2 does not include  $\mathbf{f}$  and it is an optimization problem which focuses on  $\mathbf{p}$  and  $\mathbf{B}$ .

## VI. CONVERGENCE AND TIME COMPLEXITY

In this section, we analyze the convergence of our resource allocation algorithm (Algorithm 2) and present the time complexity of Algorithm 2.

### A. Convergence Analysis

In Algorithm 2, given  $(\mathbf{p}^{(k-1)}, \mathbf{B}^{(k-1)})$  at iteration  $k$ , we can obtain an optimal solution  $\mathbf{f}^{(k)}$  of problem (17).

Then, Algorithm 1 is called in Algorithm 2, so next we discuss the convergence of Algorithm 1. To simplify the notation, we define  $\alpha := (\beta, \nu)$ .

According to Theorem 3.2 in [9], if the following three conditions hold: (i)  $\phi(\alpha)$  is differentiable and satisfies the Lipschitz condition in a solution set  $\Omega$ , (ii) assume a number  $L > 0$  such that for  $\alpha_1, \alpha_2 \in \Omega$ ,

$$\|\phi'(\alpha_1) - \phi'(\alpha_2)\| \leq L\|\alpha_1 - \alpha_2\|, \quad (41)$$

and (iii) there exists a number  $M > 0$  so that for each  $\alpha \in \Omega$ ,

$$\|[\phi'(\alpha)]^{-1}\| \leq M, \quad (42)$$

then no matter at which starting point  $\alpha^0 \in \Omega$ , (29) and (31) in Algorithm 1 will converge to a unique solution of  $\phi(\alpha) = 0$  in a linear rate. In addition, the convergence rate in the solution's neighborhood is quadratic.

$\phi(\alpha)$  in Algorithm 1 is actually linear and it is easy to prove it satisfies the Lipschitz condition and the conditions (41) (42). Therefore, Algorithm 1 converges with a linear rate at any starting point  $\alpha^0 \in \Omega$  and a quadratic convergence rate of the solution's neighborhood.

Therefore, iteratively solving Subproblem 1 to get  $\mathbf{f}$  and Subproblem 2 to get  $(\mathbf{p}, \mathbf{B})$  will converge eventually.

### B. Time Complexity

For Algorithm 1, we can use floating point operations (flops) to count the time complexity. One real addition/multiplication/division is counted as one flop. We mainly focus on analyzing Steps 3–6 because they take a major part of computational complexity. Step 3 needs  $2N$  flops. Step 4

utilizes CVX to solve problem *SP2\_v2*. Besides, CVX invokes an interior-point algorithm to solve Semidefinite Programming (SDP) problems. Thus, given the solution accuracy  $\epsilon_1 > 0$ , the worst time complexity is  $\mathcal{O}(N^{4.5} \log \frac{1}{\epsilon_1})$  [59]. Additionally, Step 5 will take  $\mathcal{O}((j+1)N)$  flops. In fact,  $j$  is a small number usually. If  $j = 0$  (i.e.,  $\xi^j = 1$ ), Step 6 is just the Newton method. Otherwise, it is a Newton-like method. Therefore, the convergence rate is superlinear or quadratic [9]. Obviously, Step 6 takes  $4N$  flops. Since the number of iterations is not larger than  $i_0$ , the time complexity of Algorithm 1 will not exceed  $\mathcal{O}(i_0(N^{4.5} \log \frac{1}{\epsilon_1} + (j+1)N))$ .

As for Algorithm 2, it solves two subproblems successively through using the convex package CVX in each iteration. Steps 3 and 4 both use CVX to solve two subproblems. As we have mentioned above, the time complexity caused by CVX is  $\mathcal{O}(N^{4.5} \log \frac{1}{\epsilon_1})$  given a solution accuracy  $\epsilon_1 > 0$ . Step 4 invokes Algorithm 1, and we have already analyzed the time complexity of Algorithm 1. With the aforementioned time complexity of Algorithm 1, the overall complexity is  $\mathcal{O}(K(i_0 + 1)N^{4.5} \log \frac{1}{\epsilon_1} + Ki_0(j+1)N)$ .

## VII. EXPERIMENTS

We will conduct extensive experiments to illustrate the effectiveness of our proposed algorithm. We first present the parameter setting of our experiments in Section VII-A, and then report various experimental results in other subsections.

### A. Parameter Setting

For most experimental settings, we use the settings used in [7]. In the experiments,  $N$  denoting the number of devices is 50. The devices are uniformly located in a circular area of size  $500 \text{ m} \times 500 \text{ m}$  and the center is a base station. The channel's pass loss is modeled as  $128.1 + 37.6 \log(\text{distance})$  along with the standard deviation of shadow fading, which is 8 dB, and the unit of distance is kilometer. The power spectral density of Gaussian noise  $N_0$  is  $-174 \text{ dBm/Hz}$ .

Additionally, the default number of local iterations  $R_l$  is set as 10 and the number of global aggregations  $R_g$  is 400. The data size for each device to upload is set as 28.1 kbits. Besides, each device has 500 samples. The number of CPU cycles per sample,  $c_n$ , is chosen randomly from  $[1, 3] \times 10^4$ . The effective switched capacitance  $\kappa$  is  $10^{-28}$ . The maximum frequency  $f_n^{max}$  and the maximum transmission power  $p_n^{max}$  for all devices are 2 GHz and 12 dBm, respectively. The minimum transmission power  $p_n^{min}$  is set as 0 dBm. The total bandwidth  $B$  is 20 MHz.

### B. How does Our Optimization Algorithm Perform with respect to System Parameters?

In this section, we study the effect of system parameters on our optimization problem. The system parameters include: weight parameters, the number of devices, global rounds & local iterations, and the radius of the circular area for generating different devices.

**Weight parameters.** In our optimization problem, we have two weight parameters  $w_1$  and  $w_2$ . The sum of them always

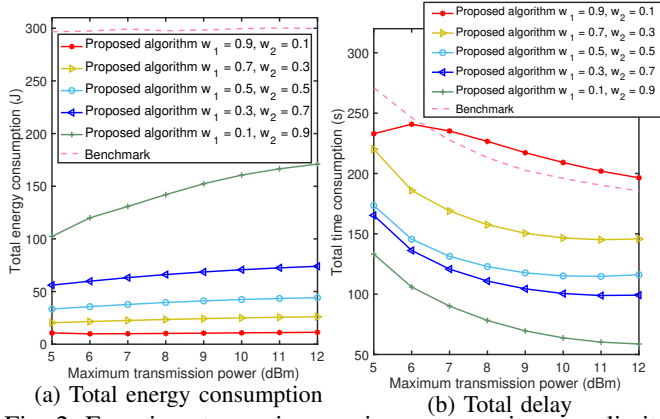


Fig. 2: Experiments varying maximum transmit power limits.

equals to 1. If  $w_1$  is greater than  $w_2$ , energy consumption will take a more dominant position in our optimization problem. Analogously, if  $w_2$  is larger than  $w_1$ , it is more important to optimize time instead of energy.

In Fig. 2a, Fig. 2b, Fig. 3a and Fig. 3b, we compare five pairs of  $(w_1, w_2) = (0.9, 0.1), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7), (0.1, 0.9)$  with a benchmark algorithm. There are six lines in those figures.  $(w_1, w_2) = (0.9, 0.1)$  and  $(0.7, 0.3)$  are to simulate devices under a low-battery condition, so the weight parameter  $w_1$  of the energy consumption is larger than the weight parameter  $w_2$  of the time consumption.  $(w_1, w_2) = (0.5, 0.5)$  represents the normal scenario, which means no preference for the time or energy. Besides,  $(w_1, w_2) = (0.1, 0.9)$  and  $(0.3, 0.7)$  stand for the time-sensitive scenario, so  $w_2$  is larger than  $w_1$ . In brief, the algorithms in our comparison include:

- Our proposed algorithm with weight parameters  $(w_1 = 0.9, w_2 = 0.1)$ ,  $(w_1 = 0.7, w_2 = 0.3)$ ,  $(w_1 = 0.5, w_2 = 0.5)$ ,  $(w_1 = 0.3, w_2 = 0.7)$  and  $(w_1 = 0.1, w_2 = 0.9)$ .
- The benchmark algorithm: 1) When comparing energy consumption and completion time at different maximum transmission power limits, for the  $n$ -th device, randomly select the CPU frequency  $f_n$  from 0.1 to 2 GHz and set  $p_n = p^{max}$ ,  $B_n = \frac{B}{N}$ . 2) When comparing at different maximum CPU frequencies, randomly select the transmission power  $p_n$  between 0 and 12 dBm and set  $f_n = f^{max}$ ,  $B_n = \frac{B}{N}$ .

To be more general, we run our algorithm at each pair of weight parameters 100 times and take the average value of energy and time consumption. At each time, we will generate 50 users randomly. Readers may wonder why the simulation plots do not show confidence intervals. This is because at each time, we only generate users randomly. For the same group of users, our algorithm will produce the same result for each operation if we do not change any system parameter. We only want to observe the overall performance of our optimization algorithm at different pairs of weight parameters  $(w_1, w_2)$ .

Fig. 2a and Fig. 2b show the results of total energy consumption and total time consumption with five pairs of  $(w_1, w_2)$  at different maximum transmit power limits, respec-

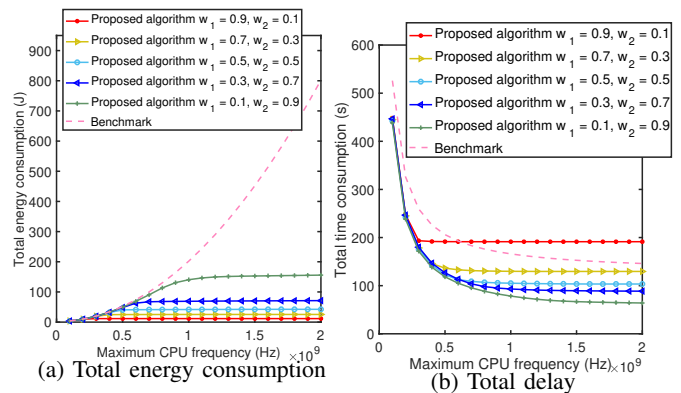


Fig. 3: Experiments with different maximum CPU frequencies.

tively. It can be observed clearly that when  $w_1$  becomes larger and  $w_2$  decreases, the total energy consumption becomes smaller and the total time becomes larger. The reason is that if  $w_1$  (resp.,  $w_2$ ) increases, our optimization algorithm will pay more attention to minimizing the energy cost (resp., time consumption).

Additionally, in Fig. 2a, the green, dark blue, light blue, yellow and red lines are all below the benchmark line with a wide gap, which means our algorithm obtains better total energy consumption even at  $w_1 = 0.1$  (this means the algorithm mainly focuses on minimizing the total time consumption instead of energy).

Correspondingly, Fig. 2b reveals that our time optimization is still better than the benchmark except for  $w_1 = 0.9$  and  $w_2 = 0.1$  after  $p^{max} = 7$  dBm. We analyze that it is caused by finding a better solution to minimize the total energy consumption through sacrificing the total time cost.

Fig. 3a and Fig. 3b compare the total energy consumption and delay at different maximum CPU frequencies. Obviously, with the increase of the maximum CPU frequency, energy consumption of the benchmark algorithm is rising, which can be seen in Fig. 3a. Intuitively, if the CPU frequency is set at the maximum, the time consumption will be reduced. Therefore, in Fig. 3b, the total time consumption of the benchmark algorithm is decreasing as CPU frequency increases. However, it only performs better than our algorithm when  $w_1 = 0.9$  and  $w_2 = 0.1$  after the maximum CPU frequency = 0.6 GHz. The reason is our algorithm mainly focuses on optimizing the energy when  $w_1 = 0.9$  and  $w_2 = 0.1$ , so our algorithm's energy consumption is the smallest in Fig. 3a.

Besides, five lines related to our algorithm in Fig. 3a and Fig. 3b all enter a stationary phase after a certain maximum CPU frequency. This is because we have already found the optimal frequency for each device at that weight parameters. Thus, there will not be any change with the increasing  $f^{max}$ .

**The number of devices.** We also explore the effect of the number of devices on the total energy consumption and the time consumption. We assume the total number of samples is 25000 and distribute all the data equally to each device.

It can be observed in Fig. 4a that with the increase of the number of devices, the total energy consumption decreases. This is caused by the decreasing number of samples allocated

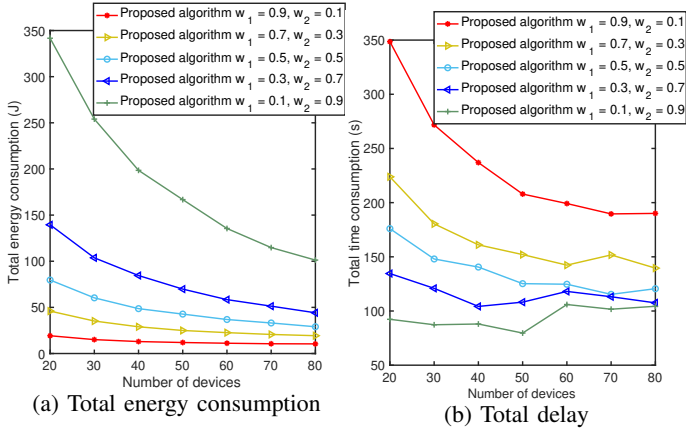


Fig. 4: Experiments with different numbers of devices.

to each device as the number of devices increases, which saves the computation energy. In Fig. 4b, the overall trend is decreasing. As we have mentioned, because we increase the number of devices, the number of samples assigned to each device decreases, and the computation time cost decreases. While, for the green line at the bottom in Fig. 4b, the total delay increases slightly as the number of devices increases. Since our algorithm optimizes a weighted sum of the energy and delay, sometimes the increase in the delay is to achieve lower energy consumption.

**The radius of the circular area.** Fig. 5a and Fig. 5b show the change in energy and time consumption as the area used for generating users randomly increases. There are three lines in each figure representing 20, 50, 80 devices, respectively. Fig. 5a does not reveal a clear correlation between energy consumption and the radius. Since the number of samples of each device is always 500 in this setting, as the number of devices increases, the overall computation energy consumption must increase unless the CPU frequency is set lower. Whereas, if the CPU frequency/transmission power is set lower, the computation time/transmission time cost becomes larger. This explains why the total energy consumption of these three lines suddenly drops at 0.9 km. Since the CPU frequency/transmission power is set lower, the total energy consumption decreases at 0.9 km. Fig. 5b clearly shows the time consumption and the radius are positively correlated. Intuitively, the transmission time will increase as the radius increases if the transmission power is kept constant. To save the time consumption, our optimization algorithm cannot always put the energy consumption first. Therefore, at 1.3 km, the CPU frequency and transmission power much be set larger than the previous settings so that the time consumption will not rise much higher.

**The communication/computation rounds.** We choose five kinds of global communication rounds  $R_g$ : 50, 100, 200, 300, 400. Fig. 6a and Fig. 6b show how the energy consumption and time cost change with the increase in the number of local iterations  $R_l$  per global round. It can be seen that as  $R_l$  and  $R_g$  become larger, the total energy and time consumption both increase. They are positively correlated.

**The number of samples on each device.** If we keep all the system parameters consistent and just vary the number

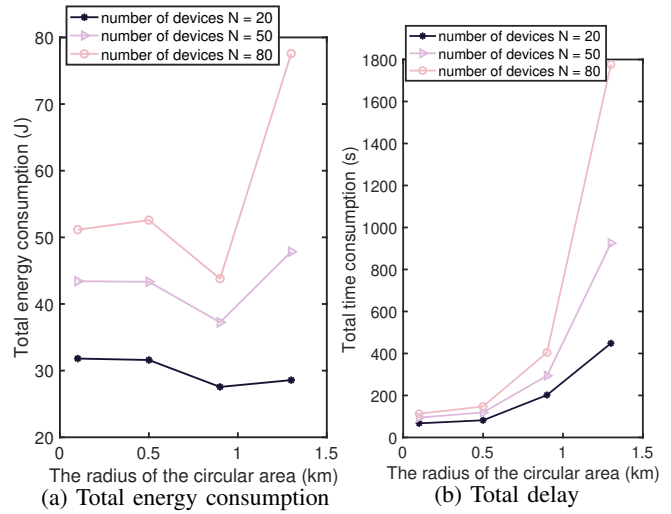


Fig. 5: Experiments with different radii of the circular area.  $w_1 = w_2 = 0.5$ .

of samples on each device, we find that  $D_n$  is positively correlated with time and energy. This result is consistent with the analytical expression of the optimization objective. Due to the space limitation, we do not provide experiments where different devices have various numbers of samples.

### C. Joint Communication & Computation Optimization vs. Communication/Computation Optimization Only

To study the domination relationship between the transmission energy and computation energy, we compare our algorithm with the following schemes at the aspect of total energy consumption at different maximum completion time limits. The maximum transmission power is  $p^{max} = 10$  dBm. We denote the maximum delay of the whole training process as  $T$ .  $T$  is fixed to compare these three schemes, and we set  $w_1 = 1$  and  $w_2 = 0$  in our algorithm.

- **Communication optimization only:** Each device's computation frequency is set as a fixed value. We optimize only the transmission power and bandwidth allocated to each device. To guarantee there is a feasible solution, we set the fixed frequency value for each device as  $\frac{R_g R_l c_n D_n}{T - R_g \max(d_n/r_n)}$ , which is derived from constraint (9a), and  $r_n$  is calculated from the initial bandwidth and transmission power.
- **Computation optimization only:** Each device's transmission power and bandwidth are fixed and we optimize only the CPU frequency. The transmission power and bandwidth of device  $n$  are set as  $p_n = p^{max}$  and  $B_n = \frac{B}{2N}$ . Such setting gives better simulation results than letting  $p_n$  be  $\frac{p^{max}}{2}$  and/or  $B_n$  be  $\frac{B}{N}$ , and is also used in the source code of [7].

Apparently, it can be observed from Fig. 7 that our proposed algorithm performs better than the other two schemes. In addition, transmission energy takes the lead of the total energy consumption. Only optimizing the computation energy cannot exceed the performance of only optimizing the transmission energy. Furthermore, we are able to discover the conflicting

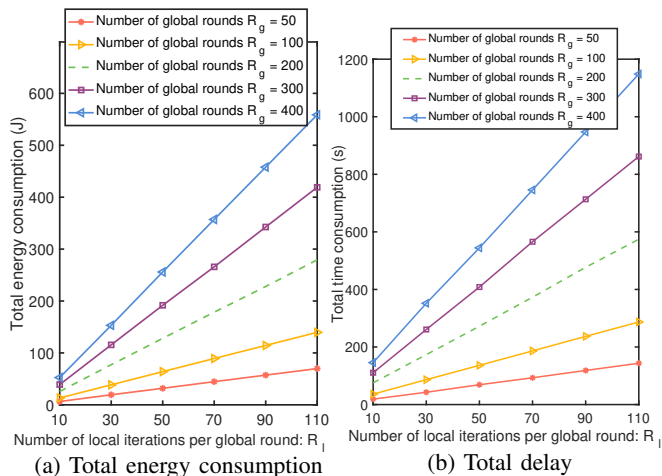


Fig. 6: Experiments with different computation rounds.  $w_1 = w_2 = 0.5$ .

relationship between maximum completion time and total energy consumption. As maximum completion time increases, there will be better solutions to reduce total energy consumption. However, if we relax the constraint of the maximum completion time  $T$  too much, the optimal solution will reach the minimum limit of transmission power and CPU frequency. That is the reason why the gap of these three lines becomes small when  $T$  is large.

#### D. Superiority of Our Algorithm over Existing Work

Prior to our study, the state of the art for FL with FDMA is [7] by Yang *et al.* on energy optimization subject to the delay constraint. In the comparison below, we refer to their method as Scheme 1. Scheme 1 has two more optimization variables, which are local accuracy and the delay. To be fair, we only use the method (Algorithm 3 in [7]) that they use to optimize the transmission power, the bandwidth and the CPU frequency of each user. Since the optimization of the local accuracy and delay is a separate section in Scheme 1, it will not affect the comparison between our method and Scheme 1.

The maximum completion time is not included in the objective function of the optimization problem in Scheme 1. Instead, it appears as a constraint. Therefore, to be fair, we set a fixed maximum completion time  $T$  to compare our optimization algorithm with Scheme 1. With a fixed maximum completion time  $T$ , our minimization problem (9) is under the condition that  $w_1 = 1$  and  $w_2 = 0$ . Besides, the initial transmission power and bandwidth of device  $n$  are set as  $p_n = p^{max}$ ,  $B_n = \frac{B}{2N}$ . The comparison result is shown in Fig. 8. From Fig. 8, we can find that each line of our proposed algorithm is below the corresponding line of Scheme 1, which reveals that our algorithm's performance is better. Additionally, as the maximum completion time  $T$  decreases, the total energy consumption gap between our algorithm and Scheme 1 gets larger, thereby indicating that our proposed algorithm leads to a better solution when the maximum completion time  $T$  is small. Therefore, our proposed algorithm will show more superiority for the training scenario with a restricted delay.

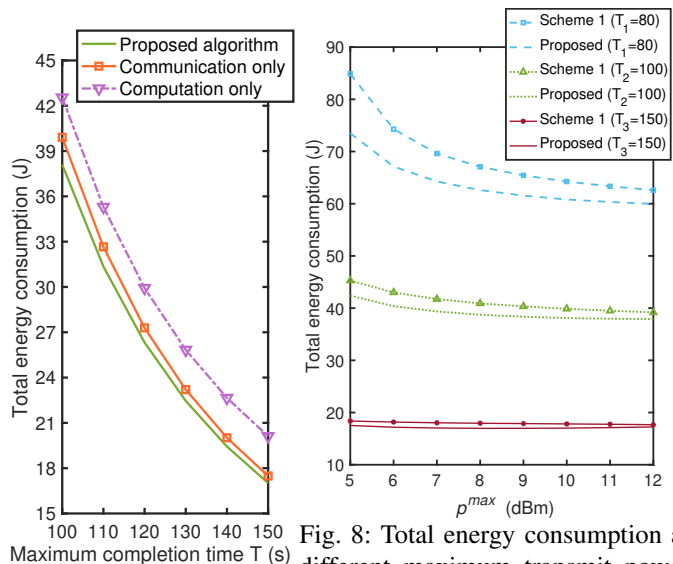


Fig. 7: Total energy consumption at different maximum completion time  $T_1 = 80$  s,  $T_2 = 100$  s and maximum completion  $T_3 = 150$  s. The number of users is  $N = 50$ .

At  $T_3 = 150$  s, the line of our algorithm first decreases and then slightly increases. This is because our initial transmission power is initialized as  $p_n = p^{max}$  at the start of our optimization algorithm. After finding the optimal solution at a specific  $p^{max}$ , the subsequent total energy consumption will slightly increase due to the different starting point from the previous step as  $p^{max}$  continues increasing.

#### VIII. CONCLUSION

In our paper, we have investigated the bandwidth allocation, transmission power control and CPU frequency of each participating device in federated learning. Through introducing two weight parameters  $w_1$  and  $w_2$ , we optimized a weighted sum of energy consumption and total completion time. By adjusting two weight parameters, the corresponding resource allocation strategy can be found. Additionally, this makes our resource allocation scheme more flexible and amenable to different requirements of the FL system. In the experiments, we can observe that our resource allocation algorithm improves the state of the art especially when there is a tight constraint in terms of the total completion time.

#### REFERENCES

- [1] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *IEEE Trans. on Knowledge & Data Engineering*, 2021.
- [2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.
- [3] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, "HFEL: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning," *IEEE Trans. on Wireless Comm.*, vol. 19, no. 10, pp. 6535–6548, 2020.

- [4] W. Luping, W. Wei, and L. Bo, "CMFL: Mitigating communication overhead for federated learning," in *IEEE ICDCS*, 2019, pp. 954–964.
- [5] P. Han, S. Wang, and K. K. Leung, "Adaptive gradient sparsification for efficient federated learning: An online learning approach," in *IEEE ICDCS*, 2020, pp. 300–310.
- [6] C. Chen, H. Xu, W. Wang, B. Li, B. Li, L. Chen, and G. Zhang, "Communication-efficient federated learning with adaptive parameter freezing," in *IEEE ICDCS*, 2021.
- [7] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. on Wireless Comm.*, vol. 20, no. 3, pp. 1935–1949, 2021.
- [8] C. T. Dinh, N. H. Tran, M. N. Nguyen, C. S. Hong, W. Bao, A. Y. Zomaya, and V. Gramoli, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. on Networking*, vol. 29, no. 1, pp. 398–409, 2021.
- [9] Y. Jong, "An efficient global optimization algorithm for nonlinear sum-of-ratios problem," *Optimization Online*, 2012.
- [10] L. Li, D. Shi, R. Hou, H. Li, M. Pan, and Z. Han, "To talk or to work: Flexible communication compression for energy efficient federated learning over heterogeneous mobile edge devices," in *IEEE INFOCOM*, 2021.
- [11] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Comm.*, vol. 37, no. 6, pp. 1205–1221, 2019.
- [12] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing," *IEEE Trans. on Wireless Comm.*, vol. 20, no. 12, pp. 7947–7962, 2021.
- [13] B. Luo, X. Li, S. Wang, J. Huang, and L. Tassiulas, "Cost-effective federated learning design," in *IEEE INFOCOM*, 2021.
- [14] Z. Yang, M. Chen, W. Saad, C. S. Hong, M. Shikh-Bahaei, H. V. Poor, and S. Cui, "Delay minimization for federated learning over wireless communication networks," in *ICML Workshop*, 2020.
- [15] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. on Wireless Comm.*, vol. 20, no. 4, pp. 2457–2471, 2020.
- [16] T. Zeng, O. Semiari, M. Mozaffari, M. Chen, W. Saad, and M. Bennis, "Federated learning in the sky: Joint power allocation and scheduling with UAV swarms," in *IEEE ICC*, 2020.
- [17] T. Wang, Y. Liu, X. Zheng, H.-N. Dai, W. Jia, and M. Xie, "Edge-based communication optimization for distributed federated learning," *IEEE Trans. on Network Science & Engineering*, 2021.
- [18] X. Wang, X. Zhong, Y. Yang, and T. Yang, "CFLMEC: Cooperative federated learning for mobile edge computing," in *IEEE ICC*, 2022.
- [19] J. Zhang, N. Li, and M. Dedeoglu, "Federated learning over wireless networks: A band-limited coordinated descent approach," in *IEEE INFOCOM*, 2021.
- [20] Y. Deng, F. Lyu, J. Ren, Y. Zhang, Y. Zhou, Y. Zhang, and Y. Yang, "SHARE: Shaping data distribution at edge for communication-efficient hierarchical federated learning," in *IEEE ICDCS*, 2021, pp. 24–34.
- [21] 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. on Wireless Comm.*, vol. 20, no. 1, pp. 269–283, 2020.
- [22] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. on Wireless Comm.*, vol. 20, no. 1, pp. 453–467, 2020.
- [23] Y. He, J. Ren, G. Yu, and J. Yuan, "Importance-aware data selection and resource allocation in federated edge learning system," *IEEE Trans. on Vehicular Technology*, vol. 69, no. 11, pp. 13 593–13 605, 2020.
- [24] J. Ren, G. Yu, and G. Ding, "Accelerating dnn training in wireless federated edge learning systems," *IEEE Journal on Selected Areas in Comm.*, vol. 39, no. 1, pp. 219–232, 2020.
- [25] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. on Wireless Comm.*, vol. 20, no. 2, pp. 1188–1200, 2020.
- [26] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *IEEE ICC*, 2020.
- [27] C. Feng, Z. Zhao, Y. Wang, T. Q. Quek, and M. Peng, "On the design of federated learning in the mobile edge computing systems," *IEEE Trans. on Communications*, vol. 69, no. 9, pp. 5902–5916, 2021.
- [28] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. on Communications*, vol. 68, no. 1, pp. 317–333, 2019.
- [29] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. on Wireless Comm.*, vol. 19, no. 11, pp. 7690–7703, 2020.
- [30] C. Wang, Y. Yang, and P. Zhou, "Towards efficient scheduling of federated mobile devices under computational and statistical heterogeneity," *IEEE Trans. on Parallel and Distributed Systems*, vol. 32, no. 2, pp. 394–410, 2020.
- [31] D. Deveaux, T. Higuchi, S. Uçar, C.-H. Wang, J. Härri, and O. Altintas, "On the orchestration of federated learning through vehicular knowledge networking," in *IEEE Vehicular Networking Conference (VNC)*, 2020.
- [32] L. Cui, X. Su, Y. Zhou, and Y. Pan, "Slashing communication traffic in federated learning by transmitting clustered model updates," *IEEE Journal on Selected Areas in Comm.*, vol. 39, no. 8, pp. 2572–2589, 2021.
- [33] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora, "FetchSGD: Communication-efficient federated learning with sketching," in *International Conference on Machine Learning*, 2020, pp. 8253–8265.
- [34] H. Guo, H. Wang, T. Song, Y. Hua, X. Lv, X. Jin, Z. Xue, R. Ma, and H. Guan, "Siren: Byzantine-robust federated learning via proactive alarming," in *Proceedings of the ACM Symposium on Cloud Computing*, 2021, p. 47–60.
- [35] F. Malandrino and C. F. Chiasserini, "Toward node liability in federated learning: Computational cost and network overhead," *IEEE Communications Magazine*, vol. 59, no. 9, pp. 72–77, 2021.
- [36] J. Wu, Q. Xia, and Q. Li, "Efficient privacy-preserving federated learning for resource-constrained edge devices," in *IEEE Conference on Mobility, Sensing and Networking (MSN)*, 2021, pp. 191–198.
- [37] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.
- [38] R. Zeng, S. Zhang, J. Wang, and X. Chu, "FMore: An incentive scheme of multi-dimensional auction for federated learning in MEC," in *IEEE ICDCS*, 2020, pp. 278–288.
- [39] R. Zhou, J. Pang, Z. Wang, J. C. Lui, and Z. Li, "A truthful procurement auction for incentivizing heterogeneous clients in federated learning," in *IEEE ICDCS*, 2021, pp. 183–193.
- [40] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-iid data with reinforcement learning," in *IEEE INFOCOM*, 2020, pp. 1698–1707.
- [41] S. Yue, J. Ren, J. Xin, D. Zhang, Y. Zhang, and W. Zhuang, "Efficient federated meta-learning over multi-access wireless networks," *IEEE Journal on Selected Areas in Comm.*, 2022.
- [42] S. Lin, G. Yang, and J. Zhang, "A collaborative learning framework via federated meta-learning," in *IEEE ICDCS*, 2020, pp. 289–299.
- [43] Y. Zhou, G. Pu, X. Ma, X. Li, and D. Wu, "Distilled one-shot

federated learning,” *arXiv*, 2020.

- [44] Y. Wang, Z. Su, N. Zhang, and A. Benslimane, “Learning in the air: Secure federated learning for UAV-assisted crowdsensing,” *IEEE Trans. on Network Science & Engineering*, vol. 8, no. 2, pp. 1055–1069, 2020.
- [45] Y. Wang, Z. Su, T. Luan, R. Li, and K. Zhang, “Federated learning with fair incentives and robust aggregation for uav-aided crowdsensing,” *IEEE Trans. on Network Science & Engineering*, 2021.
- [46] Y. Liu, Z. Ma, X. Liu, S. Ma, S. Nepal, R. Deng, and K. Ren, “Boosting privately: Federated extreme gradient boosting for mobile crowdsensing,” in *IEEE ICDCS*, 2020.
- [47] M. Piva, G. Maselli, and F. Restuccia, “The tags are alright: Robust large-scale RFID clone detection through federated data-augmented radio fingerprinting,” in *ACM MobiHoc*, 2021, pp. 41–50.
- [48] G. Reus-Muns and K. R. Chowdhury, “Classifying uavs with proprietary waveforms via preamble feature extraction and federated learning,” *IEEE Trans. on Vehicular Technology*, vol. 70, no. 7, pp. 6279–6290, 2021.
- [49] Y. Huang, J. Zhang, J. Duan, B. Xiao, F. Ye, and Y. Yang, “Resource allocation and consensus of blockchains in pervasive edge computing environments,” *IEEE Trans. on Mobile Computing*, 2021.
- [50] Y. K. Cheung, S. Leonardos, G. Piliouras, and S. Sridhar, “From griefing to stability in blockchain mining economies,” *arXiv*, 2021.
- [51] J. R. Gunasekaran, C. S. Mishra, P. Thinakaran, B. Sharma, M. T. Kandemir, and C. R. Das, “Cocktail: A multidimensional optimization for model serving in cloud,” in *USENIX NSDI*, 2022, pp. 1041–1057.
- [52] S. Guan, A. Boukerche, and A. Loureiro, “Novel sustainable and heterogeneous offloading management techniques in proactive cloudlets,” *IEEE Trans. on Sustainable Computing*, vol. 6, no. 2, pp. 334–346, 2021.
- [53] T. Tan and G. Cao, “Deep learning video analytics on edge computing devices,” in *IEEE SECON*, 2021.
- [54] R. Mayer and H.-A. Jacobsen, “Scalable deep learning on distributed infrastructures: Challenges, techniques, and tools,” *ACM Comput. Surv.*, vol. 53, no. 1, 2021.
- [55] Y. Yang, R. G. D’Oliveira, S. El Rouayheb, X. Yang, H. Seferoglu, and Y. Chen, “Secure coded computation for efficient distributed learning in mobile IoT,” in *IEEE SECON*, 2021.
- [56] S. Tripathi, C. Puligheddu, C. F. Chiasserini, and F. Mungari, “A context-aware radio resource management in heterogeneous virtual RANs,” *IEEE Trans. on Cognitive Commu. & Netw.*, vol. 8, no. 1, pp. 321–334, 2022.
- [57] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
- [58] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” 2014.
- [59] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, “Semidefinite relaxation of quadratic optimization problems,” *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

#### APPENDIX A: PROOF OF LEMMA 1

It is obvious that  $K(p_n, B_n) = p_n d_n$  is an affine function with  $B_n = 0$ . Thus,  $K(p_n, B_n)$  is convex.

We calculate the Hessian matrix of  $G(p_n, B_n)$ :

$$Hessian(G) = \begin{bmatrix} -\frac{g_n^2}{B_n N_0^2 (\frac{g_n p_n}{B_n N_0} + 1)^2} & \frac{g_n^2 p_n}{B_n^2 N_0^2 (\frac{g_n p_n}{B_n N_0} + 1)^2} \\ \frac{g_n^2 p_n}{B_n^2 N_0^2 (\frac{g_n p_n}{B_n N_0} + 1)^2} & -\frac{g_n^2 p_n^2}{B_n^3 N_0^2 (\frac{g_n p_n}{B_n N_0} + 1)^2} \end{bmatrix}.$$

For any vector  $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^2$ , we have  $\mathbf{x}^T Hessian(G) \mathbf{x} = -\frac{(x_1 g_n B_n - x_2 g_n p_n)^2}{B_n^3 N_0^2 (\frac{g_n p_n}{B_n N_0} + 1)^2} \leq 0$ . Because

$\mathbf{x}^T * Hessian(G) * \mathbf{x} \leq 0$  for all  $\mathbf{x}$ ,  $Hessian(G)$  is a negative semidefinite matrix. Therefore,  $G(p_n, B_n)$  is a concave function. **Lemma 1** is proved.

#### APPENDIX B: PROOF OF THEOREM 2

We derive the relationship between  $p$  and  $B$  from (33), so

$$p_n = \left( \frac{(\nu_n \beta_n + \tau_n) g_n}{N_0 d_n \nu_n \ln 2} - 1 \right) \frac{N_0 B_n}{g_n}. \quad (\text{A.1})$$

From (35) and (A.1), we have  $B_n = \frac{r_n^{min}}{\log_2 \left( \frac{(\nu_n \beta_n + \tau_n) g_n}{N_0 d_n \nu_n \ln 2} \right)}$ , if  $\tau_n \neq 0$ . Substituting (A.1) in (32), we replace  $p_n$  by  $B_n$  in the Lagrangian function:  $L_3(B_n, \tau_n, \mu) = B_n \left( \frac{\nu_n \beta_n + \tau_n}{\ln 2} - \frac{\nu_n d_n N_0}{g_n} - (\nu_n \beta_n + \tau_n) \times \log_2 \left( \frac{(\nu_n \beta_n + \tau_n) g_n}{N_0 d_n \nu_n \ln 2} \right) + \mu \right) + \sum_{n=1}^N r_n^{min} * \tau_n - \mu B$ . The corresponding dual problem writes as

$$\max_{\tau_n, \mu} g(\tau_n, \mu) = \sum_{n=1}^N r_n^{min} * \tau_n - \mu B \quad (\text{A.2})$$

subject to  $\frac{a_n}{\ln 2} - j_n - a_n \log_2 \left( \frac{a_n}{j_n \ln 2} \right) + \mu = 0, \mu \geq 0, \tau_n \geq 0$ , (A.3)

where  $a_n = \nu_n \beta_n + \tau_n$  and  $j_n = \frac{\nu_n d_n N_0}{g_n}$ . Besides, constraint (A.3) derives the relationship between  $\tau_n$  and  $\mu$ :

$$\tau_n = \frac{(\mu - j_n) \ln 2}{W \left( \frac{\mu - j_n}{e * j_n} \right)} - \nu_n \beta_n, \quad (\text{A.4})$$

where  $W(\cdot)$  is Lambert  $W$  function and  $e$  is the mathematical constant. Given (A.4), the dual function is simplified as

$$g(\mu) = \sum_{n=1}^N r_n^{min} * \left( \frac{(\mu - j_n) \ln 2}{W \left( \frac{\mu - j_n}{e * j_n} \right)} - \nu_n \beta_n \right) - \mu B. \quad (\text{A.5})$$

Take the first derivative and we get  $g'(\mu) = \sum_{n=1}^N \frac{r_n^{min} \ln 2}{W \left( \frac{\mu - j_n}{e * j_n} \right)} \left( 1 - \frac{1}{W \left( \frac{\mu - j_n}{e * j_n} \right) + 1} \right) - B$ . Additionally,  $g''(\mu) = -\sum_{n=1}^N \frac{r_n^{min} \ln 2 W \left( \frac{\mu - j_n}{e * j_n} \right)}{(\mu - j_n) (1 + W \left( \frac{\mu - j_n}{e * j_n} \right))^3} \leq 0$  due to  $\frac{\mu - j_n}{e * j_n} \geq -\frac{1}{e}$ .

Therefore,  $g'(\mu)$  is a monotone decreasing function, and  $g(\mu)$  is concave, so  $g(\mu)$  reaches maximum when  $g'(\mu) = 0$ . Bisection method can be used to find  $\mu$  satisfying  $g'(\mu) = 0$ . Naturally,  $\tau_n = \max((A.4), 0)$ . Note  $\tau_n \neq 0$  implies  $B_n = \frac{r_n^{min}}{\log_2 \left( \frac{(\nu_n \beta_n + \tau_n) g_n}{N_0 d_n \nu_n \ln 2} \right)}$ . We denote the sum of bandwidth of these devices by  $B_{\tau_n \neq 0}$  and the set of these devices by  $\mathcal{N}_{\tau_n \neq 0}$ , with  $N_{\tau_n \neq 0}$  being the cardinality of  $\mathcal{N}_{\tau_n \neq 0}$ .

We use (A.1) to replace  $p_n$  by  $B_n$  in problem **SP2\_v2** of (21), and remove those devices whose bandwidth is calculated in the previous step and numbering the left devices. Then the new problem becomes

$$\min_{B_n} \sum_{n=1}^{N - N_{\tau_n \neq 0}} \left( \frac{\nu_n \beta_n}{\ln 2} - \frac{N_0 d_n \nu_n}{g_n} - \nu_n \beta_n \log_2 \left( \frac{\beta_n g_n}{N_0 d_n \ln 2} \right) \right) B_n \quad (\text{A.6})$$

subject to,

$$p_n^{min} \leq \left( \frac{\beta_n g_n}{N_0 d_n \ln 2} - 1 \right) \frac{N_0 B_n}{g_n} \leq p_n^{max}, \quad \forall n \in \mathcal{N} \setminus \mathcal{N}_{\tau_n \neq 0},$$

$$\sum_{n=1}^{N - N_{\tau_n \neq 0}} B_n \leq B - B_{\tau_n \neq 0},$$

where  $\mu$  and  $\tau_n$  have already been solve in previous steps.

Then the problem **SP2\_v2** becomes a convex optimization problem with just one variable. Thus, we can use the convex problem solver CVX [58] to solve problem (A.6).