Efficient Bandwidth Allocation and Computation Configuration in Industrial IoT

HUANG Rui, LI Huilin, ZHANG Yongmin
(Central South University, Changsha 410012, China)

DOI: 10.12142/ZTECOM.202301007
https://kns.cnki.net/kcms/detail/34.1294.TN.20230213.1639.003.html, published online February 14, 2023
Manuscript received: 2022-12-01

Abstract: With the advancement of the Industrial Internet of Things (IoT), the rapidly growing demand for data collection and processing poses a huge challenge to the design of data transmission and computation resources in the industrial scenario. Taking advantage of improved model accuracy by machine learning algorithms, we investigate the inner relationship of system performance and data transmission and computation resources, and then analyze the impacts of bandwidth allocation and computation resources on the accuracy of the system model in this paper. A joint bandwidth allocation and computation resource configuration scheme is proposed and the Karush-Kuhn-Tucker (KKT) conditions are used to get an optimal bandwidth allocation and computation configuration decision, which can minimize the total computation resource requirement and ensure the system accuracy meets the industrial requirements. Simulation results show that the proposed bandwidth allocation and computation resource configuration scheme can reduce the computing resource usage by 10% when compared to the average allocation strategy.

Keywords: bandwidth allocation; computation resource management; industrial IoT; system accuracy

Citation (IEEE Format): R. Huang, H. L. Li, and Y. M. Zhang, "Efficient bandwidth allocation and computation configuration in industrial IoT," *ZTE Communications*, vol. 21, no. 1, pp. 55–63, Mar. 20223. doi: 10.12142/ZTECOM.202301007.

1 Introduction

In recent years, the advances in computation, communication and application design and the rapid development of the Internet of Things (IoT) have been driving the realization of intelligence and automation in the industry[1-2]. Through various IoT devices, a large number of data can be collected, such as images, sounds and temperatures, to judge the operating status of equipment and efficient follow-up maintenance/management strategies are then made. For the traditional Industrial IoT with the cloud, the system performance may be affected by the network performance and computing capability of the cloud since data should be transmitted to a remote cloud via the Internet for processing. With the development of industry, more and more data needs to be collected and processed in real time, leading to explosive growth in communication overhead and computation requirements, which brings significant challenges in the design of Industrial IoT, especially with high-reliability requirements.

To solve data transmission issues in the Industrial IoT, the communication framework has been updated to improve the speed and reliability of data transmission[3-4] and a new wireless transmission system framework has also been proposed to help design an operable and effective end-to-end wireless solution[5]. Moreover, many works have focused on improving data transmission technologies, such as time slot frequency hopping technologies[6] and clustering of data transmission[7]. Besides optimizing the communication framework of the Industrial IoT, some researchers have considered and studied the energy consumption, delay, cost and other parameters associated with the data transmission issues; for example, Ref. [8] proposed a bandwidth allocation strategy based on deep reinforcement learning algorithm and Ref. [9] enables the control of transmission energy consumption for the dynamic change of bandwidth. To deal with the large and unstable communication latency, an edge computing system with computing resources deployed at the network edge has been introduced into the Industrial IoT and become a potential mainstream solution[10-12]. These works alleviate the problem of insufficient wireless resources.

To solve the computation resource issues, there exists a lot of work focusing on the optimization of task offloading performance for cloud computing/edge computing or collaborative edge-cloud computing, such as computation delay, energy consumption, resource efficiency and data quality, to guarantee the quality of computation service[13-16]. With the gradual deepening of machine learning research, it has been discov-

This work has been supported in part by the National Natural Science Foundation of China under Grant No. 62172445 and in part by the Young Talents Plan of Hunan Province, China.
Corresponding author: ZHANG Yongmin

ered that the number of training epochs directly affects the accuracy of the system model after training[17 - 18]. Considering that allocated computation resources can determine the training epochs in a given time scale, some researchers have investigated the relationships among accuracy of the system model, the number of processed data, the number of computation resources, the training speed/delay and the energy consumption, and made use of machine learning based algorithms to further improve the performance of the computation system[19 - 22]. In such a way, the performance of the computation system can be further improved.

However, most of the current works do not consider the inner relationship between the bandwidth allocation and the computation resource management incurred by the data transmission and just assume that the IoT devices transmit all of collected data to the edge server via access points (AP). The AP needs to try its best to forward the data to the edge server and the edge server processes all the received data making use of its available computation resources. Unfortunately, with the explosive increase of IoT devices, it is difficult for the existing Industrial IoT system to carry on such a heavy workload, which may lead to network congestion, even network crash when wireless communication resources are exhausted. Therefore, to solve the data transmission problem, it is worthwhile to consider the inner relationship among the computation resources, the accuracy of the system model and the data transmission. Making up for the shortage of wireless resources by increasing the computing resources can guarantee system accuracy.

In this paper, we aim at the scenario of resource management in the Industrial IoT, which can allocate the wireless communication resources by the AP and train a high-accuracy model by computation resources at the edge server. First, we model the available channel bandwidth for each IoT device based on the allocated bandwidth and the distance between the IoT device and the AP. Second, we formulate the bandwidth allocation and computation configuration as a resource requirement minimization problem. Then, we analyze the relationship among the transmitted data, the computation resources and the system accuracy, and design a heuristic algorithm to obtain the optimal computation resources allocation and communication resources management to each IoT device. The contributions of this paper can be summarized as follows:

• The bandwidth allocation and computation resource management problem for Industrial IoT is formulated as a cost minimization problem with the given accuracy requirement.

• The relationship among the accuracy of the system model, the transmitted data and the computation resources is investigated and an efficient bandwidth allocation and resource management scheme is designed to satisfy the system requirement with a minimal resource requirement.

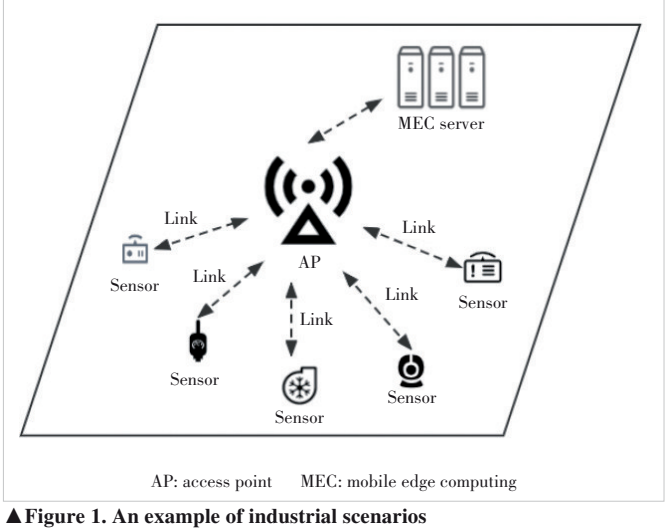
• Simulation results show the proposed algorithm can minimize the resource requirement with a performance guarantee.

The rest of the paper is organized as follows. Section 2 presents the system model and problem formulation. An algorithm that can minimize the resource requirement with performance guarantee is proposed in Section 3. An operational performance analysis is demonstrated based on simulation results in Section 4. Finally, Section 5 concludes our work.

2 System Model and Problem Formulation

Fig. 1 shows an industrial scenario of the Industrial IoT system. In this system, there are N IoT devices and the set of IoT devices is denoted as $\boldsymbol{N} = \{1, 2, \cdots, N\}$, and several APs (small cell base stations or Wi-Fi APs) with edge servers. Generally, an IoT device collects monitoring data and transmits the data to the edge server via AP using the wireless communication technology, the AP allocates the available wireless bandwidth to each IoT device and forwards the monitoring data to the edge server for processing, and the edge server processes the monitoring data using machine learning models to satisfy the requirement of system performance. Here, we assume that each IoT device can adjust its monitoring data according to the available wireless bandwidth, and the connections of the IoT devices and APs are given due to specified monitoring objects. Thus, for ease of description, we only focus on optimizing communication bandwidth allocation and computation resource management strategies in a single AP with an edge server with multiple IoT device scenarios in this paper, but the results can be extended to multiple APs with multiple edge servers based on the deployment of the inference model. Note that we mainly focus on the allocation of communication resources and the configuration of computation resources at the edge server.

Considering the time-varying feature of industrial scenarios, one optimization period can be divided into T time segments $\boldsymbol{T} = \{1, 2, \cdots, T\}$, and t represents the t-th time segment. The



▲Figure 1. An example of industrial scenarios

accuracy requirement of a system model for one IoT device during one time segment is a constant and can be changed at different time segments.

2.1 Communication Model

Generally, all IoT devices need to send their real-time monitoring data to the edge server for processing via the AP. It means that several IoT devices will transmit their data to the AP simultaneously. To mitigate interference among IoT devices, some effective interference cancellation techniques, such as orthogonal frequency-division multiple access (OFDMA) and time division multiple access (TDMA), can be used by the AP. In this paper, we assume that OFDMA is used for wireless communications. Besides the interference, signal pass loss is another important factor that affects data transmission. According to Refs. [23] and [24], the pass loss can be formulated as a function of transmission distance with a path loss exponent $2 \leqslant \alpha \leqslant 4$. Let $h_{n,t}$ denote the small-scale channel gain from the n-th mobile device to the AP during time segment t. The achievable data transmission rate for IoT device n during time segment t, denoted by $R_{n,t}$, can be given by

$$R_{n,t} = w_{n,t} \log_2\left(1 + \frac{P_{n,t}\left|h_{n,t}\right|^2}{\left(d_{n,t}\right)^{\alpha}\sigma^2}\right), n = 1, 2, \cdots, N, \tag{1}$$

where $w_{n,t}$ denotes the allocated bandwidth for IoT device n during time segment t, $P_{n,t}$ denotes the transmission power of IoT device n during time segment t, $d_{n,t}$ denotes the distance between IoT device n and AP, and σ^2 denotes the background noise power. Generally, $w_{n,t}$ is determined by AP, $d_{n,t}$ and σ^2 are constants, and the value of $P_{n,t}$ can be calculated by the power control algorithms[24-25]. Due to the limitation of AP's wireless communication resources, the total bandwidths that can be allocated to IoT devices have an upper bound, denoted by $\bar{W}$. Thus, we have

$$\sum_n w_{n,t} \leqslant \bar{W}. \tag{2}$$

It is obvious that the bandwidth allocation strategy of the AP should consider the distance $d_{n,t}$ for IoT device n and the requirements of all the IoT devices. The data set of IoT device n that have been transmitted to the edge server during time segment t, denoted by $D_{n,t}$, is

$$D_{n,t} = R_{n,t} * t, \tag{3}$$

where t is the total number of time units in one time segment.

2.2 Computation Model

The edge server can make use of the received data and the machine learning based algorithm to train a high-accuracy system model for each IoT device. We define accuracy as the opposite of the loss function in a training model based on federated learning. In general, the performance of the system model achieved by the machine learning algorithm can be affected by multiple factors, including feature selection, user-defined parameters, data sets and computation resources for training. In this paper, we mainly consider the impact of the data set and the computation resources on the training results and intend to find an appropriate data set and computing resources to satisfy the requirements of system accuracy.

According to Refs. [17] and [19], the accuracy of the system model through training of traditional machine learning algorithms generally tends to increase with the increase of the data set. At the same time, due to the noise $\delta_{n,t}$ in the data set and the limitation of the model capacity, the accuracy growth rate of the system model will gradually slow down until it becomes stable[26]. Besides, the precision of a machine learning algorithm, such as a neural network, has a logarithmic relationship with the number of training epochs[17]. Thus, with the increase of computation resources, the number of training epochs within the limited time scale can be increased, which can improve the precision of the system model in a logarithmic form. Thus, in this paper, the accuracy of the system model for IoT device n during time segment t, denoted by $\xi_{n,t}$, can be modeled as

$$\xi_{n,t} = a \log_{10}\left(\frac{C_{n,t}}{C_{\text{epoch}}} * \frac{D_{n,t}}{D_{\text{unit}}} + b\right) + \delta_n, \tag{4}$$

where a and b are accuracy parameters based on the machine learning algorithm and $0 \leqslant a,b \leqslant 1$, C_{epoch} denotes the computation resources required for training the data set for one epoch, D_{unit} denotes a reference unit of the data set for training, and δ_n is the influence factor of the noise in the data set on the accuracy. Generally, $\delta_{n,t}$ $(-1 \leqslant \delta_n < 0)$ is a constant affected by the noise during time segment t[27].

It can be found that both the data set and computation resources can affect the accuracy of the system model. Because of this, it provides the industrial IoT with an opportunity to solve the wireless resource issues by managing the computation resources.

2.3 Problem Formulation

In this paper, we intend to design an efficient bandwidth allocation and computation resource management scheme for the Industrial IoT to satisfy the accuracy requirement of each IoT devices. Let $\underline{\xi_{n,t}}$ denote the accuracy requirement of IoT device n during time segment t. Thus, we have

$$\xi_{n,t} \geqslant \underline{\xi_{n,t}}, \ \forall n, t. \tag{5}$$

To achieve the accuracy requirements of each IoT device, the AP allocates its available communication resources to IoT

devices for data transmission while the edge server manages the computation resources for data processing. In other words, when wireless communication resources are scarce/costly, more computation resources can be used to improve the accuracy of the system model. Otherwise, more computation resources can be saved to keep the accuracy of the system model at a given level.

Considering that the wireless communication resources for each AP are limited, our objective function is to minimize the total computation resources requirement, which can both minimize the operating cost and identify the bottleneck of the system performance. The bandwidth allocation and the computation resource management problem can be formulated as following

$$\mathbf{P1}: \min_{\vec{w},\vec{C}} \sum_{n} C_{n,t}, \tag{6}$$

$$\text{s.t.} \sum_{n} w_{n,t} \leqslant \bar{W}, \quad \forall t, \tag{7}$$

$$\xi_{n,t} \geqslant \underline{\xi}_{n,t}, \quad \forall n, t, \tag{8}$$

where $\vec{\boldsymbol{w}} = \left\{w_{n,t}, \forall n,t\right\}$ is the set of the bandwidth allocation of the AP and $\vec{\boldsymbol{C}} = \left\{C_{n,t}, \forall n,t\right\}$ is the set of the computation resources for data processing. The objective of Problem **P1** is to obtain the optimal bandwidth allocation, which can minimize the total number of computation resources. The first constraint ensures the sum of the bandwidth resources that are allocated to the IoT devices does not exceed the total numer of the available bandwidth resources of the AP. The second constraint guarantees that the accuracy of the system model for each IoT device can meet the industrial requirements.

3 Optimal Bandwidth Allocation and Computation Configuration Scheme

To solve this problem, from the perspective of the edge server, we study the relationship among accuracy $\xi_{n,t}$, computation resources $C_{n,t}$, and data set $D_{n,t}$ of a specific IoT device n. To satisfy the accuracy requirement of each IoT device, we can analyze the influence of data set $D_{n,t}$ on the computation resource requirements for each IoT device. Then, through the communication model, the relationship between the data set and the allocated bandwidth resources can be obtained. Thus, we can derive the impact of bandwidth allocation decisions on the computation resource requirements for each IoT device.

By analyzing the relationship among $\xi_{n,t}$, $C_{n,t}$ and $D_{n,t}$, we have the following results.

Lemma 1: The accuracy $\xi_{n,t}$ obtained by the edge server is an increasing and concave function with respect to the computation resources $C_{n,t}$ when the data set $D_{n,t}$ is given.

Proof: According to Eq. (4), we can derive the first and second derivatives of $\xi_{n,t}$ with respect to $C_{n,t}$ as follows:

$$\frac{\partial \xi_{n,t}}{\partial C_{n,t}} = \frac{1}{\ln 10} \frac{a}{\frac{C_{n,t}}{C_{\text{epoch}}} * \frac{D_{n,t}}{D_{\text{unit}}} + b} \frac{D_{n,t}}{C_{\text{epoch}} * D_{\text{unit}}}, \tag{9}$$

$$\frac{\partial^2 \xi_{n,t}}{\partial {C_{n,t}}^2} = \frac{1}{\ln 10} \left(\frac{D_{n,t}}{C_{\text{epoch}} * D_{\text{unit}}}\right)^2 \frac{-a}{\left(\frac{C_{n,t}}{C_{\text{epoch}}} * \frac{D_{n,t}}{D_{\text{unit}}} + b\right)^2}. \tag{10}$$

Since each item of Eq. (9) is positive, $(\partial \xi_{n,t}) / (\partial C_{n,t}) > 0$ holds. Since only $-a$ in Eq. (10) is negative, $(\partial^2 \xi_{n,t}) / (\partial {C_{n,t}}^2) < 0$ holds. Thus $\xi_{n,t}$ is an increasing and concave function of $C_{n,t}$.

Lemma 2: The accuracy $\xi_{n,t}$ obtained by the edge server is an increasing and concave function with respect to the data set $D_{n,t}$ when the computation resources $C_{n,t}$ is given.

The proof of Lemma 2 is similar to that of Lemma 1, so we omit it. We can also derive that $(\partial \xi_{n,t}) / (\partial D_{n,t}) > 0$ and $(\partial^2 \xi_{n,t}) / (\partial {D_{n,t}}^2) < 0$. Thus $\xi_{n,t}$ is an increasing and concave function of $D_{n,t}$.

Theorem 1: Accuracy $\xi_{n,t}$ is an increasing and concave function with respect to both the computation resources $C_{n,t}$ and the data set $D_{n,t}$.

Proof: According to Lemma 1 and Lemma 2, $\xi_{n,t}$ is an increasing and concave function with respect to $C_{n,t}$ or $D_{n,t}$ when the other variable is given. Furthermore, since $C_{n,t}$ and $D_{n,t}$ are independent, according to Ref. [28], it can be proved that $\xi_{n,t}$ is an increasing and concave function with respect to $C_{n,t}$ and $D_{n,t}$.

Based on Theorem 1, we have the following theorem for the optimal solution to **P1**.

Theorem 2: The optimal solution to **P1** should satisfy $\{\xi_{n,t} = \underline{\xi}_{n,t}, \forall n\}$ and $\sum_n w_{n,t} = \bar{W}$.

Proof: According to Theorem 1, for a specific IoT device n, $\xi_{n,t}$ is an increasing function of $C_{n,t}$ and $D_{n,t}$. First, we can prove that $\sum_n w_{n,t} = \bar{W}$ is a necessary condition for the optimal solution by contradiction as follows.

Assuming that there exists an optimal solution, denoted by $\{w'_{n,t}, \forall n\}$, satisfying $\sum_n w'_{n,t} < \bar{W}$ and $\xi_{n,t} = \underline{\xi}_{n,t}$, we can increase any $w'_{n,t}$ by δ_n, $0 < \delta_n \leqslant \bar{W} - \sum_n w'_{n,t}$, and find a smaller $C'_{n,t}$ satisfying $C'_{n,t} < C_{n,t}$ to make $\xi_{n,t} = \underline{\xi}_{n,t}$. This contradicts the objective function. Thus, $\sum_n w_{n,t} = \bar{W}$ always holds for the optimal solution to **P1**.

Then, we can prove that $\{\xi_{n,t} = \underline{\xi}_{n,t}, \forall n\}$ is a necessary condition for the optimal solution by contradiction. If there exists an optimal solution, denoted by $\xi'_{n,t}$, satisfying $\xi'_{n,t} > \underline{\xi}_{n,t}$ and $\sum_n w'_{n,t} < \bar{W}$. According to Theorem 1, we can decrease $C_{n,t}$ to

make $\xi_{n,t}=\underline{\xi_{n,t}}$ and keep $\sum_n w'_{n,t}=\bar{W}$. This contradicts the objective function. Thus, $\{\xi_{n,t}=\underline{\xi_{n,t}}, \forall n\}$ is another necessary condition for the optimal solution to P1.

According to Theorem 2, we have the relationship among $\xi_{n,t}$, $C_{n,t}$ and $D_{n,t}$ as follows:

$$\xi_{n,t}=\underline{\xi_{n,t}}=a\log_{10}\left(\frac{C_{n,t}}{C_{\text{epoch}}}*\frac{D_{n,t}}{D_{\text{unit}}}+b\right)+\delta_n. \tag{11}$$

Therefore, we can obtain the expression of $C_{n,t}$ about $D_{n,t}$ as follows:

$$C_{n,t}=C_{\text{epoch}}D_{\text{unit}}\left(10^{\frac{\underline{\xi_{n,t}}-\delta_n}{a}}-b\right)\frac{1}{D_{n,t}}. \tag{12}$$

Based on Eq. (12), we have the following property:

Lemma 3: The optimal computation resource $C_{n,t}$ is a decreasing and convex function of the data set $D_{n,t}$.

Proof: Based on Eq. (11), we can calculate the derivative of $C_{n,t}$ with respect to $D_{n,t}$ as follows:

$$\frac{\partial C_{n,t}}{\partial D_{n,t}}=C_{\text{epoch}}D_{\text{unit}}\left(10^{\frac{\underline{\xi_{n,t}}-\delta_n}{a}}-b\right)\frac{-1}{{D_{n,t}}^2}, \tag{13}$$

$$\frac{\partial^2 C_{n,t}}{\partial {D_{n,t}}^2}=C_{\text{epoch}}D_{\text{unit}}\left(10^{\frac{\underline{\xi_{n,t}}-\delta_n}{a}}-b\right)\frac{2}{{D_{n,t}}^3}. \tag{14}$$

It can be found that $(\partial\xi_{n,t})/(\partial C_{n,t})>0$ and $(\partial^2\xi_{n,t})/(\partial {C_{n,t}}^2)<0$, which means that $C_{n,t}$ is a decreasing and convex function of $D_{n,t}$.

According to the definition of data transmission rate $R_{n,t}$ in Eq. (1) and the data set $D_{n,t}$ in Eq. (3), it can be found that $D_{n,t}$ is a linear function of the bandwidth allocation $w_{n,t}$. Thus, we have the following lemma:

Lemma 4: The optimal computation resource $C_{n,t}$ is a decreasing and convex function of the bandwidth allocation $w_{n,t}$.

Proof: According to Lemma 3, $C_{n,t}$ is a decreasing and convex function of $D_{n,t}$. Thus, we have $(\partial\xi_{n,t})/(\partial C_{n,t})>0$ and $(\partial^2\xi_{n,t})/(\partial {C_{n,t}}^2)<0$. Since $D_{n,t}$ is a linear function of $w_{n,t}$, according to the chain rule of derivation, $(\partial\xi_{n,t})/(\partial C_{n,t})>0$ and $(\partial^2\xi_{n,t})/(\partial {C_{n,t}}^2)<0$ hold. Thus, $C_{n,t}$ is a decreasing and convex function of $w_{n,t}$.

Theorem 3: There exists a unique optimal solution $\{w_{n,t}, \forall n,t\}$ for P1.

Proof: According to Lemma 4, the objective function of P1 is a decrease and convex function of the bandwidth allocation $w_{n,t}$. It can be found that the first and second constraints are linear constraints of $w_{n,t}$. Hence, P1 is a convex optimization problem with respect to $w_{n,t}$. According to the properties of the convex optimization problem in Ref. [28], there exists a unique optimal bandwidth allocation $\{w_{n,t}, \forall n,t\}$ for P1.

Since P1 is a convex optimization problem, based on its KKT conditions, the optimal solution can be achieved by the following theorem.

Theorem 4: The optimal solution to P1 is

$$w^*_{n,t}=\frac{\bar{W}\sqrt{\frac{\beta^1_{n,t}}{\beta^2_{n,t}}}}{\sum_{n'}\sqrt{\frac{\beta_{1,n'}}{\beta_{2,n'}}}}. \tag{15}$$

Proof: Generally, since P1 for one time segment is independent with the other time segments, we can solve P1 for each time segment t.

Let v_t be the Lagrange multiplier associated with the constraint $\sum_n w_{n,t}\leqslant\bar{W}$. The Lagrangian of P1 is

$$L\left(w_{n,t},v_t\right)=\sum_n C_{n,t}+v_t\left(\sum_n w_{n,t}-\bar{W}\right)=$$
$$-\bar{W}*v_t+\sum_n\left(C_{n,t}+v_t*w_{n,t}\right). \tag{16}$$

It can be found that the above equation is separable. Thus, the dual function is

$$g\left(v_t\right)=-\bar{W}*v_t+\min_{\vec{w}}\sum_n\left(C_{n,t}+v_t*w_{n,t}\right). \tag{17}$$

According to Lemma 4, $C_{n,t}$ is a decreasing and convex function of $w_{n,t}$. Thus, we can get the minimal value of $\sum_n(C_{n,t}+v_t*w_{n,t})$ when $[\partial C_{n,t}+v_t*w_{n,t}]/\left(\partial w_{n,t}\right)=0$, which means

$$w^*_{n,t}=\sqrt{\frac{\beta^1_{n,t}}{\beta^2_{n,t}}}*\frac{1}{\sqrt{v_t}}, \tag{18}$$

where $\beta^1_{n,t}=C_{\text{epoch}}D_{\text{unit}}(10^{\frac{\underline{\xi_{n,t}}-\delta_n}{a}}-b)$ and $\beta^2_{n,t}=\log_2\left(1+\frac{P_{n,t}\left|h_{n,t}\right|^2}{(d_{n,t})^{\alpha}\sigma^2}\right)$. Thus we can rewrite Eq. (17) as

$$g\left(v_t\right)=-\bar{W}*v_t+2\sqrt{v_t}\sum_n\sqrt{\frac{\beta^1_{n,t}}{\beta^2_{n,t}}}, \tag{19}$$

and the dual problem is

$$\min_{v_t}\quad g\left(v_t\right),$$
$$\text{s.t.}\quad v_t\geqslant 0. \tag{20}$$

Since $[\partial^2 g(v_t)]/(\partial v_t^2) = -\frac{1}{2} v_t^{-\frac{3}{2}} \sum_n \sqrt{\beta_{n,t}^1 / \beta_{n,t}^2} < 0$, the dual function $g(v_t)$ is a concave function. According to the convex optimization theorem, the optimal v_t^* should satisfy $[\partial g(v_t)]/(\partial v_t) = 0$. Thus, the optimal v_t^* can be calculated by

$$v_t^* = \left(\frac{1}{\bar{W}} \sum_n \sqrt{\frac{\beta_{n,t}^1}{\beta_{n,t}^2}} \right)^2, \tag{21}$$

and substituting v_t^* into Eq. (18), we can obtain Eq. (15).

Therefore, according to the location information and accuracy requirements of all IoT devices, we design an efficient bandwidth allocation and computation configuration algorithm, named EBACC, which can solve P1 and get the optimal decision of bandwidth allocation and computation configuration.

Algorithm 1. Efficient Bandwidth Allocation and Computation Configuration Algorithm (EBACC)

1: **for** each time segment t, $t \in [1,T]$, **do**
2: **Input:** $\{d_{n,t}, P_{n,t}, \underline{\xi}_{n,t}, \forall n\}$.
3: According to Eq. (1) in the communication model, calculate $\beta_{n,t}^1$ of each IoT device.
4: According to Eq. (13) in the computation model, calculate $\beta_{n,t}^2$ of each IoT device.
5: According to Eq. (15), calculate the decision of bandwidth allocation $\vec{\boldsymbol{w}}_t$ by using $\beta_{n,t}^1$ and $\beta_{n,t}^2$.
6: According to Eqs. (1) and (4), calculate the decision of computation configuration $\vec{\boldsymbol{C}}_t$ based on $\vec{\boldsymbol{w}}_t$.
7: **Output:** optimal $\vec{\boldsymbol{w}}_t$ and $\vec{\boldsymbol{C}}_t$.
8: **end for**

4 Simulation

In this section, numerical experiments have been conducted to verify the correctness of the lemmas and performance of the proposed algorithm EBACC. We first consider a scenario where the AP has a coverage range of 200 m and there are $N = 60$ randomly scattered IoT devices within the coverage region. We randomly generate the distance $d_{n,t}$ between each IoT device and AP within $[10\text{ m}, 200\text{ m}]$. In the communication model, we assume that the upper bound of total bandwidth resources of the AP is $\bar{W} = 200$ MHz. And the reference signal-to-noise ratio (SNR) at the transmission distance $d_0 = 10$ m is set to $\gamma_0 = \left[P_{n,t} \left| h_{n,t} \right|^2 \right] / (d_{n,t})^{\alpha} \sigma^2 = 80$ dB. The propagation distance can be converted to $d'_{n,t} = d_{n,t}/d_0$, which is within $[1, 20]$, and the path loss exponent is set to $\alpha = 3$. Meanwhile, we randomly generate the accuracy requirement of each device within $[0.8, 0.95]$.

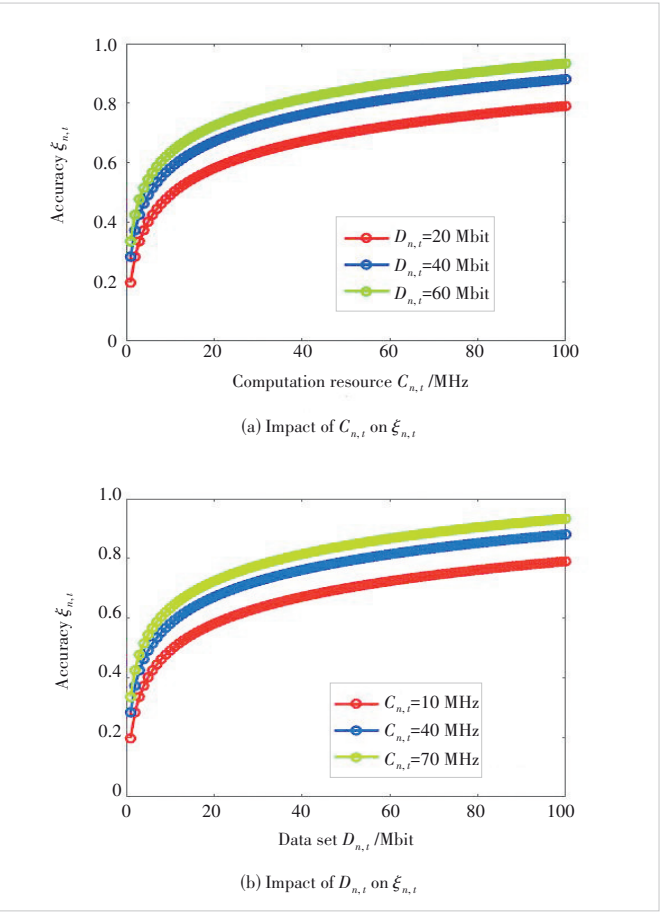
In the following subsection, we firstly explore the relationship between variables in the computation and communication models. Then we verify the correctness of the lemmas in Section 3. Last, we evaluate the performance of the proposed algorithm EBACC, which can get the optimal bandwidth resource allocation to minimize the total computation resources while satisfying the accuracy requirements of IoT devices.

4.1 Impact of Computation Resources and Data Set on Accuracy of Training Results

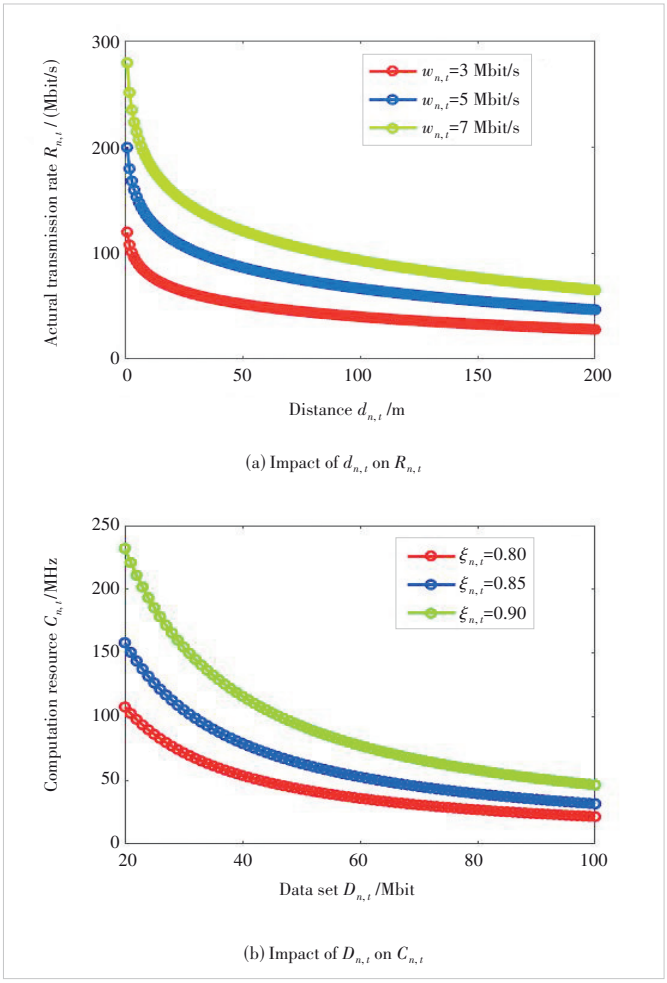
As shown in Figs. 2(a) and 2(b), the accuracy of training results shows a growing trend with the increase of data set size or computing resources, and the growth rate will be gradually slowed down, which verifies the conclusion of Lemma 1 that the accuracy $\xi_{n,t}$ is an increasing concave function with respect to $C_{n,t}$ and $D_{n,t}$. Thus, we can configure more computing resources or upload more data to improve the accuracy of training results.

4.2 Impact of Distance from IoT devices to AP on Actual Transmission Rate

As shown in Fig. 3(a), it is obvious that the actual transmission rate $R_{n,t}$ is a decreasing and convex function of distance



▲ **Figure 2. Impact of computation resources and data set on the accuracy of training results**



(a) Impact of $d_{n,t}$ on $R_{n,t}$

(b) Impact of $D_{n,t}$ on $C_{n,t}$

▲ **Figure 3. Impact of distance from IoT devices to AP on the actual transmission rate and that of data set on computation resources**

$d_{n,t}$ from IoT devices to the AP. The closer the IoT device is to the AP, the higher the actual transmission rate will be. Thus, we can allocate more bandwidth resources to the farther IoT devices, which can reduce the impact of distance to get smaller computing resource requirements.

4.3 Impact of Data Set on Computation Resources Requirement

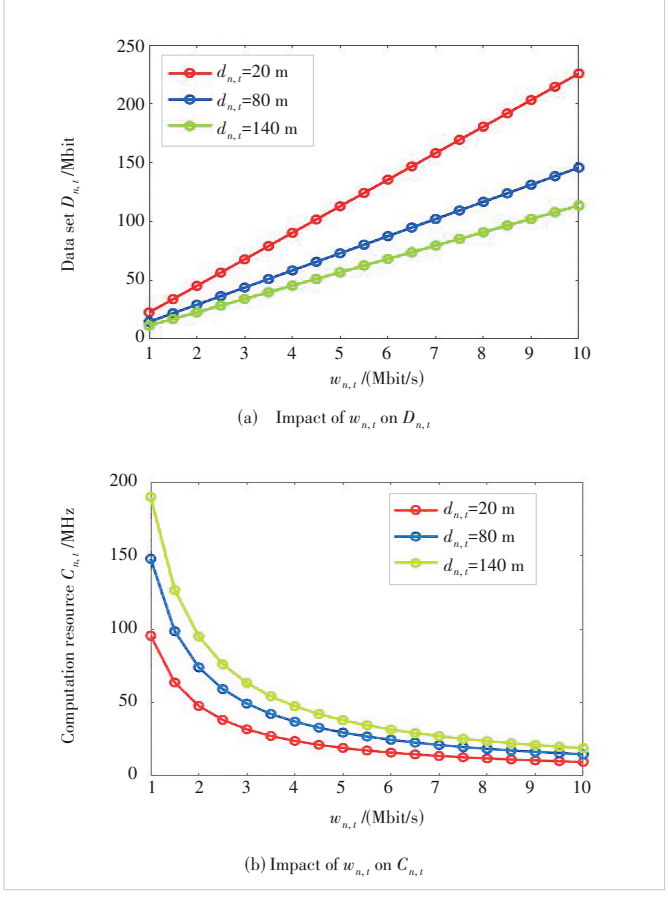
As shown in Fig. 3(b), when the accuracy $\xi_{n,t}$ is given, the computation resource requirement $C_{n,t}$ by the IoT device is a decreasing and convex function of the data set $D_{n,t}$, which has been proved by Lemma 3. It means when the uploaded data set is larger, the computing resources required by the model will be reduced. In addition, it can be found that, with the improvement of the accuracy $\xi_{n,t}$ of model requirements, the computation resources $C_{n,t}$ will become larger. Therefore, when the accuracy of model requirement is given, we can make a trade-off between the number of uploaded data and computing resources.

4.4 Impact of Bandwidth Resources Allocation on Data Set and Computation Resources Requirement

As shown in Fig. 4(a), during one time segment, data set $D_{n,t}$ that the IoT device can upload to the edge server is a linear and increasing function of the allocated bandwidth resources $w_{n,t}$. It can be found that the IoT device closer to the AP has a higher positive slope. Meanwhile, as shown in Fig. 4(b), for an IoT device, the computation resources requirement $C_{n,t}$ is a decreasing and convex function of the bandwidth resources allocated to it, which has proved the correctness of Lemma 4. We also find that the IoT device, which is farther away from the AP, will need more computation resources to satisfy the accuracy requirements when the bandwidth resource is given. Thus, if we want to minimize the total computation resources, we need to allocate bandwidth resources reasonably. In this way, the IoT device farther away from the AP should be allocated with more bandwidth resources.

4.5 Optimal Bandwidth Resources Allocation

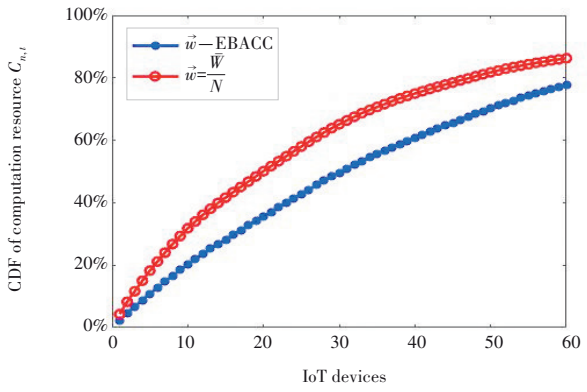
We compare two strategies of bandwidth allocation: 1) the optimal bandwidth resources allocation decided by EBACC; 2) allocating bandwidth resources equally to all IoT devices. As shown in Fig. 5(a), when the total computation resource



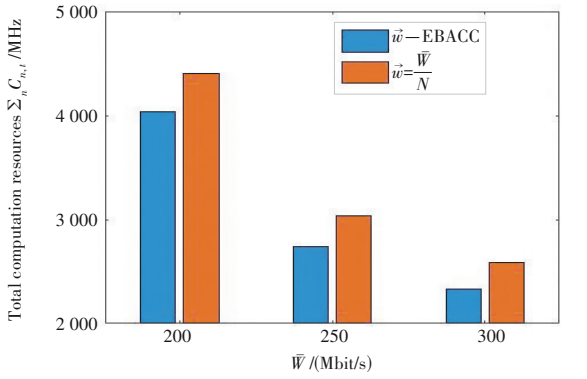
(a) Impact of $w_{n,t}$ on $D_{n,t}$

(b) Impact of $w_{n,t}$ on $C_{n,t}$

▲ **Figure 4. Impact of bandwidth resources allocated to IoT device on data set and computation resources**



(a) CDF of $C_{n,t}$



(b) Total computation resources requirement

CDF: Cumulative Distribution Function
EBACC: Efficient Bandwidth Allocation and Computation Configuration Algorithm
IoT: Internet of Things

▲Figure 5. CDF of computation resource requirement of each IoT device and total computation resources requirement under two situations: 1) optimal bandwidth resources allocation decided by EBACC; 2) allocating bandwidth resources equally to all IoT devices.

available is 3 000 MHz, the first strategy can use 77.85% of the total computation resource to satisfy the accuracy requirement of all IoT devices, but the second strategy needs 86.28%. It means that the proposed algorithm can significantly improve the efficiency of computing and bandwidth resources. Meanwhile, as shown in Fig. 5(b), the optimal bandwidth resource allocation can significantly reduce the demand of total computation resources of all IoT devices. Specifically, when the total bandwidth resource is $\bar{W}$ = 300 Mbit/s, the optimal bandwidth resource allocation can reduce the total computation resource requirement from 2 588.1 MHz to 2 335.4 MHz.

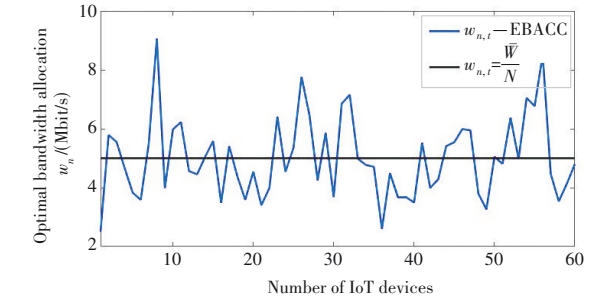
4.6 Relationship Between Optimal Bandwidth Allocation and Distance of IoT Devices

As shown in Figs. 6(a) and 6(b), we explore the relationship between the optimal bandwidth allocation decision $\boldsymbol{w}_t = \{\hat{w}_{1,t}, \hat{w}_{2,t}, \cdots, \hat{w}_{N,t}\}$ and the distances of all IoT devices $\boldsymbol{d}_t = \{d_{1,t}, d_{2,t}, \cdots, d_{N,t}\}$. Compared with the average allocation strategy, the optimal bandwidth allocation decision will be obviously affected by the accuracy requirement of IoT devices and the distance between the device and the AP. And it can be found that more bandwidth resources will be allocated to the IoT device farther away from the AP or with higher accuracy requirements.

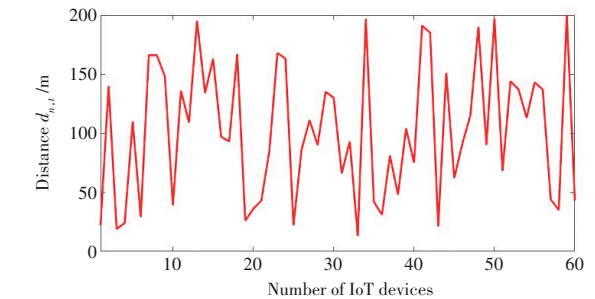
5 Conclusions

In this paper, we focus on the bandwidth allocation of AP and the computation resource management of the edge server to ensure the system accuracy can meet the industrial requirement. We formulate the bandwidth allocation and computation resource management problem for the industrial IoT as a cost minimization problem with a given accuracy requirement. Then, we analyze the relationship among the transmitted data, computation resources and system accuracy and then design an efficient algorithm to obtain the optimal computation resource allocation and communication resource management. Numerical experiment results demonstrate that the proposed algorithm EBACC can significantly reduce the number of total computation resources while satisfying the accuracy requirements of the industrial IoT.

For future work, we are going to consider the more general cases where IoT devices can choose different APs and edge servers to process their data and obtain a high-accuracy sys-



(a) Optimal bandwidth resource allocation decision $\vec{w}_t$



(b) Distances of IoT devices

EBACC: Efficient Bandwidth Allocation and Computation Configuration Algorithm
IoT: Internet of Things

▲Figure 6. Relationship between the optimal bandwidth allocation decision and distances of IoT devices

tem model. We will focus on the bandwidth allocation between multiple APs and multiple IoT devices, which would be more technically challenging.

References

[1] ADI E, ANWAR A, BAIG Z, et al. Machine learning and data analytics for the IoT [J]. Neural computing and applications, 2020, 32(20): 16205 – 16233. DOI: 10.1007/s00521-020-04874-y
[2] LEE J, STANLEY M, SPANIAS A, et al. Integrating machine learning in embedded sensor systems for Internet-of-Things applications [C]//IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 2016: 290 – 294. DOI: 10.1109/ISSPIT.2016.7886051
[3] LIU Y K, CANDELL R, KASHEF M, et al. Dimensioning wireless use cases in Industrial Internet of Things [C]//14th IEEE International Workshop on Factory Communication Systems (WFCS). IEEE, 2018: 1 – 4
[4] LUO Y, DUAN Y, LI W F, et al. A novel mobile and hierarchical data transmission architecture for smart factories [J]. IEEE transactions on industrial informatics, 2018, 14(8): 3534 – 3546. DOI: 10.1109/TII.2018.2824324
[5] LIU Y K, KASHEF M, LEE K B, et al. Wireless network design for emerging IIoT applications: reference framework and use cases [J]. Proceedings of the IEEE, 2019, 107(6): 1166 – 1192. DOI: 10.1109/JPROC.2019.2905423
[6] SAVAZZI S, KIANOUSH S, RAMPA V, et al. A joint decentralized federated learning and communications framework for industrial networks [C]//IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD). IEEE, 2020: 1 – 7. DOI: 10.1109/CAMAD50429.2020.9209305
[7] LONG N B, TRAN-DANG H, KIM D S. Energy-aware real-time routing for large-scale industrial Internet of Things [J]. IEEE Internet of Things journal, 2018, 5(3): 2190 – 2199. DOI: 10.1109/JIOT.2018.2827050
[8] JAGANNATH J, POLOSKY N, JAGANNATH A, et al. Machine learning for wireless communications in the Internet of Things: a comprehensive survey [J]. Ad hoc networks, 2019, 93: 101913. DOI: 10.1016/j.adhoc.2019.101913
[9] DING Z M, SHEN L F, CHEN H Y, et al. Energy-efficient relay-selection-based dynamic routing algorithm for IoT-oriented software-defined WSNs [J]. IEEE Internet of Things journal, 2020, 7(9): 9050 – 9065. DOI: 10.1109/JIOT.2020.3002233
[10] ZHAO R, WANG X J, XIA J J, et al. Deep reinforcement learning based mobile edge computing for intelligent Internet of Things [J]. Physical communication, 2020, 43: 101184. DOI: 10.1016/j.phycom.2020.101184
[11] KAUR K, GARG S, AUJLA G S, et al. Edge computing in the industrial Internet of Things environment: software-defined-networks-based edge-cloud interplay [J]. IEEE communications magazine, 2018, 56(2): 44 – 51. DOI: 10.1109/MCOM.2018.1700622
[12] ZHANG K, MAO Y M, LENG S P, et al. Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks [J]. IEEE access, 2016, 4: 5896 – 5907. DOI: 10.1109/access.2016.2597169
[13] HONG Z C, CHEN W H, HUANG H W, et al. Multi-hop cooperative computation offloading for industrial IoT-edge-cloud computing environments [J]. IEEE transactions on parallel and distributed systems, 2019, 30(12): 2759 – 2774. DOI: 10.1109/TPDS.2019.2926979
[14] GAO G J, XIAO M J, WU J, et al. Auction-based VM allocation for deadline-sensitive tasks in distributed edge cloud [J]. IEEE transactions on services computing, 2021, 14(6): 1702 – 1716. DOI: 10.1109/TSC.2019.2902549
[15] MA X, WANG S G, ZHANG S, et al. Cost-efficient resource provisioning for dynamic requests in cloud assisted mobile edge computing [J]. IEEE transactions on cloud computing, 2021, 9(3): 968 – 980. DOI: 10.1109/TCC.2019.2903240
[16] YANG B, CAO X L, LI X F, et al. Mobile-edge-computing-based hierarchical machine learning tasks distribution for IIoT [J]. IEEE Internet of Things journal, 2020, 7(3): 2169 – 2180. DOI: 10.1109/JIOT.2019.2959035
[17] SUN C, SHRIVASTAVA A, SINGH S, et al. Revisiting unreasonable effectiveness of data in deep learning era [C]//IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 843 – 852. DOI: 10.1109/ICCV.2017.97
[18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770 – 778. DOI: 10.1109/CVPR.2016.90
[19] HUANG J, RATHOD V, SUN C, et al. Speed/accuracy trade-offs for modern convolutional object detectors [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 3296 – 3297. DOI: 10.1109/CVPR.2017.351
[20] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for deep learning in NLP [C]//57th Annual Meeting of the Association for Computational Linguistics. ACL, 2019: 3645 – 3650
[21] QU Y B, LIU J J. Computation offloading for mobile edge computing with accuracy guarantee [C]//ACM Turing Celebration Conference. ACM, 2019: 1 – 5. DOI: 10.1145/3321408.3321582
[22] LIN J, CHEN W M, LIN Y J, et al. MCUNet: tiny deep learning on IoT devices [C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. NIPSF, 2020: 11711 – 11722
[23] CHEN X, JIAO L, LI W Z, et al. Efficient multi-user computation offloading for mobile-edge cloud computing [J]. IEEE/ACM transactions on networking, 2016, 24(5): 2795 – 2808. DOI: 10.1109/TNET.2015.2487344
[24] CHIANG M, HANDE P, LAN T, et al. Power control in wireless cellular networks [J]. Foundations and trends in networking, 2008, 2(4): 381 – 533. DOI: 10.1561/1300000009
[25] XIAO M B, SHROFF N B, CHONG E K P. A utility-based power-control scheme in wireless cellular systems [J]. IEEE/ACM transactions on networking, 2003, 11(2): 210 – 221. DOI: 10.1109/TNET.2003.810314
[26] MIECH A, ZHUKOV D, ALAYRAC J B, et al. HowTo100M: learning a text-video embedding by watching hundred million narrated video clips [C]//IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 2630 – 2640. DOI: 10.1109/ICCV.2019.00272
[27] HUANG J W, BERRY R A, HONIG M L. Distributed interference compensation for wireless networks [J]. IEEE journal on selected areas in communications, 2006, 24(5): 1074 – 1084. DOI: 10.1109/JSAC.2006.872889
[28] BOYD S, VANDENBERGHE L. Convex Optimization [M]. Cambridge: UK: Cambridge University Press, 2004. DOI: 10.1017/cbo9780511804441

Biographies

HUANG Rui received his BS degree in computer science from Wuhan University of Technology, China. He is currently pursuing his master's degree with the School of Computer Science and Engineering, Central South University, China. His research interests include mobile edge computing and network optimization.

LI Huilin received his BS degree in mechanical design manufacture and automation from Shandong University, China. He is currently pursuing his master's degree with the School of Computer Science and Engineering, Central South University, China. His research interests include mobile edge computing and federal learning.

ZHANG Yongmin (zhangyongmin@csu.edu.cn) received his PhD degree in control science and engineering from Zhejiang University, China in 2015. From 2015 to 2019, he was a post-doctoral research fellow at the Department of Electrical and Computer Engineering, University of Victoria, Canada. He is currently a professor with the School of Computer Science and Engineering, Central South University, China. His research interests include resource management and optimization in wireless networks, smart grid, and mobile computing. He won the Best Paper Award of the IEEE PIMRC'12 and the IEEE Asia-Pacific Outstanding Paper Award 2018.