This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

Optimal SIC Ordering and Computation Resource Allocation in MEC-aware NOMA NB-IoT Networks

Li Ping Qian, *Senior Member, IEEE*, Anqi Feng, Yupin Huang, Yuan Wu, *Senior Member, IEEE*
Bo Ji, *Senior Member, IEEE*, Zhiguo Shi, *Senior Member, IEEE*

***Abstract*—Non-orthogonal multiple access (NOMA) and mobile edge computing (MEC) have been emerging as promising techniques in narrowband Internet of Things (NB-IoT) systems to provide ubiquitously connected IoT devices with efficient transmission and computation. However, the successive interference cancellation (SIC) ordering of NOMA has become the bottleneck limiting the performance improvement for the uplink transmission, which is the dominant traffic flow of NB-IoT communications. Also, in order to guarantee the fairness of task execution latency across NB-IoT devices, the computation resource of MEC units has to be fairly allocated to tasks from IoT devices according to the task size. For these reasons, we investigate the joint optimization of SIC ordering and computation resource allocation in this paper. Specifically, we formulate a combinatorial optimization problem with the objective to minimize the maximum task execution latency required per task bit across NB-IoT devices under the limitation of computation resource. We prove the NP-hardness of this joint optimization problem. To tackle this challenging problem, we first propose an optimal algorithm to obtain the optimal SIC ordering and computation resource allocation in two stages: the convex computation resource allocation optimization followed by the combinatorial SIC ordering optimization. To reduce the computational complexity, we design an efficient heuristic algorithm for the SIC ordering optimization. As a good feature, the proposed low-complexity algorithm suffers a negligible performance degradation in comparison with the optimal algorithm. Simulation results demonstrate the benefits of NOMA in reducing the task execution latency.**

***Index Terms*—Non-orthogonal multiple access, Mobile edge computing, successive interference cancellation ordering, computation resource allocation, execution latency minimization**

I. Introduction

With the rapid development of Internet of Things (IoT) technology, various IoT applications have been emerging in recent years, including smart metering [1], smart manufacturing [2], smart home/city [3], automatic driving [4], health monitoring [5], and many others (see a recent survey in [6]). Driven by the requirements of these applications on low power consumption, low latency, and massive connectivity [7]–[9], the narrowband Internet of Things (NB-IoT) technology has been proposed in 3GPP as a promising low power wide area (LPWA) radio technology for IoT systems [10]. Specifically, NB-IoT systems operate in the underlying cellular network with very small bandwidth, and accommodate a massive number of low-data-rate machine-type connections with low power consumption in extreme coverage conditions. Along with the ever-growing popularity of NB-IoT devices, the unprecedent IoT traffic volumes are delivered to small-cell base stations (BSs) in the context of Long-Term Evolution (LTE) network. The critical-IoT traffics further require the reliable and ultra low-latency data computing. However, due to the limited spectrum-computation resource, current wireless cellular networks are becoming incapable to provide the massive connectivity and intensive computation for IoT traffics [11].

To cope with these challenges, non-orthogonal multiple access (NOMA) [12] and mobile edge computing (MEC) [13] are emerging as promising technologies for NB-IoT systems. Unlike the conventional orthogonal multiple access (OMA) in which the frequency/time/code resource is orthogonally allocated to multiple users at the same time, the NOMA serves multiple users on the same frequency-time resource simultaneously through the non-orthogonal resource allocation in the power domain [14] or code domain [15]. Therefore, the NOMA can meet various demands of NB-IoT systems on superior spectral-energy efficiency, massive connections, and ultra low transmission latency. Unlike the conventional cloud computing operated in the remote cloud that suffers severe transmission latency via the Internet, MEC offers cloud computing capabilities at the edge of radio access network (e.g., at small-cell BSs) in close proximity to NB-IoT devices. Through bringing intensive computation tasks from NB-IoT devices to MEC units, the low-latency as well as reliable computing services can be implemented for NB-IoT devices. As a result, it is envisioned that integrating NOMA and MEC into NB-IoT systems can bring enormous potential benefits to various IoT applications.

In this paper, we apply the power-domain NOMA to MEC-aware NB-IoT networks. For simplicity, the word "NOMA" in the rest of the paper means the power-domain NOMA. In the implementation of NOMA, multiple NB-IoT devices with distinct channel conditions transmit their data to the small-cell BS via superposition coding (SC), and the small-cell BS decodes the data from each NB-IoT device sequentially via successive interference cancellation (SIC). Despite the superior benefits of NOMA, the joint radio and computation resource allocation has to be considered when the NOMA is applied

Manuscript received July 27, 2018; revised September 14, 2018; accepted October 06, 2018. This work was supported in part by the National Natural Science Foundation of China under Project 61379122 and Project 61572440, in part by the Zhejiang Provincial Natural Science Foundation of China under Project LR16F010003 and Project LR17F010002, and in part by the NSF under Grants CNS-1651947. (*Corresponding author: Yuan Wu.*)

L. P. Qian, A. Feng, Y. Huang, and Y. Wu are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China (email: {lpqian, iewuy}@zjut.edu.cn).

B. Ji is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122 USA (e-mail: boji@temple.edu).

Z. Shi is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310008, China (email: shizg@zju.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

to MEC-aware NB-IoT networks. However, for the features of NB-IoT (or machine-type) communications, such as low power, latency-sensitivity, and massive connectivity, how to allocate limited radio-computation resource among NB-IoT devices becomes a challenging issue.

While existing works such as [16] and [33] have studied the joint allocation of radio and computation resources, none of these works consider a joint optimization technique for MEC-aware NOMA NB-IoT networks with taking into account the SIC ordering. It is worth noting that exiting works on NOMA assume that the SIC ordering follows the ascending order of channel conditions [17]–[21], in which the user with the worst channel condition will be decoded first. This doing may result in poor performance/fairness for users with poor channel conditions, although the radio resource allocation has been considered. To this end, the current study aims to minimize the maximum task execution latency[1] of computing 1-bit task across NB-IoT devices through a joint optimization of SIC ordering and computation resource allocation for MEC-aware NOMA NB-IoT networks. We focus on the uplink transmission, since it is the dominant traffic flow of NB-IoT communications. Our contributions in this paper can be summarized as follows.

- We propose a novel NOMA-based MEC model for NB-IoT networks that captures the gains of uplink MEC-aware NOMA in the task execution latency. Specifically, we present a joint optimization framework that minimizes the maximum task execution latency required per task bit across NB-IoT devices through jointly optimizing the SIC ordering of NB-IoT devices and computation resource allocation.
- The proposed optimization problem is a combinatorial optimization problem involving the SIC ordering and computation resource allocation, and we prove its NP-hardness. We then propose an optimal algorithm to obtain the optimal SIC ordering and computation resource allocation in two stages. In the first stage, given the SIC ordering, we obtain the optimal computation resource allocation in the closed-form expression in the two-device case, and by the bisection searching in the multi-device case. In the second stage, we obtain the optimal SIC ordering based on the min-max execution latency obtained for all possible SIC ordering in the first stage through exhaustive search.
- To reduce the computational complexity, we design an efficient heuristic algorithm for SIC ordering. Specifically, we assign the SIC order of each device sequentially one by one, where the optimal SIC order is determined by the min-max execution latency across all assigned devices. In doing so, the computational complexity can be reduced to $O(N^2)$ for SIC ordering, where N is the number of all NB-IoT devices existing in the network. We compare the performance of our proposed heuristic algorithm and optimal algorithm, and the simulation result shows that the heuristic algorithm suffers a negligible degradation in performance. We also evaluate the performance of our proposed heuristic algorithm via extensive simulations, in which we show that benefits of NOMA in reducing the task execution latency.

The rest of this paper is organized as follows. In Section II, we review the related studies in the existing literature. In Section III, we introduce the NOMA-argumented MEC model and present the problem formulation. In Section IV, we propose the optimal algorithm and low-complexity heuristic algorithm to solve the proposed optimization problem. Finally, Sections V and VI present numerical results and conclude this paper, respectively.

II. Related Work

In this paper, we introduce two techniques, e.g., NOMA and MEC, to NB-IoT networks to meet various demands on low power consumption, low latency, and massive connectivity. Therefore, we elaborate on the related studies on NOMA and MEC in this section.

Thanks to the superior benefits of NOMA, researchers have spent significant amount of research efforts on this topic recently [20]–[26]. Islam *et al.* [23] summarized the potentials and challenges of NOMA applied to the fifth-generation networking system. With the goal of interference mitigation, user pairing (or clustering) algorithms were proposed for NOMA networks in [20]–[22]. Regarding the communication resource allocation, the joint optimization of sub-carrier assignment and power allocation were studied in [24] for maximizing the sum utility of NOMA network. Qian *et al.* [25] proposed a coalitional game based algorithm to maximize the system-wide utility and minimize the total power consumption through jointly optimizing the user association and power allocation for NOMA-enabled small-cell networks. In [26], the joint optimization problem of power allocation, user pair selection, and time-frequency resource allocation were studied for multi-cell NOMA. With the explosive growth of IoT applications, the studies on NOMA have been extended to IoT networks [27]–[30]. Ding *et al.* [27] designed precoding and power allocation coefficients for the MIMO-NOMA IoT network with two users categorized by their quality-of-service requirements, not by their channel conditions. Mostafa *et al.* [28] proposed a joint sub-carrier and transmission power allocation algorithm to maximize the connection density of NB-IoT devices for uplink NOMA NB-IoT networks. Wu *et al.* [29] investigated the spectral efficiency maximization problem for wireless powered NOMA IoT networks. Zhai *et al.* [30] proposed a joint user scheduling and power allocation algorithm to minimize the long-term power consumption based on the stochastic optimization theory for NOMA IoT networks.

As a common feature, all previous work in [20]–[30] make an assumption that all users are decoded by SIC based on the levels of channel conditions. This assumption is plausible in the downlink NOMA, since the data signals of users in the strong channel conditions must be decoded after the data signals of users in the weak channel conditions are subtracted. However, in the uplink NOMA where the small-call BS works

[1]In this paper, the task execution latency includes the task transmission latency from the NB-IoT device to the small-cell BS, and the task computation delay at MEC units equipped at the small-cell BS.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

as a common receiver for all served users, all SIC ordering can be performed for the same sum data rate in the decoding process. It implies that it is important to improve the quality-of-service of individual users via a proper SIC ordering in the uplink NOMA. On the other hand, these existing work mainly focus on the efficient transmission for NB-IoT networks.

As discussed earlier, NB-IoT networks are expected to provide the efficient transmission and computation for ubiquitously connected IoT devices. Therefore, MEC has also attracted great attentions of academia and industry recently in the networking context of IoT [31]–[37]. Chiang *et al.* [31] summarized the opportunities and challenges of edge computing in IoT applications. Sun *et al.* [32] proposed a hierarchical mobile edge computing architecture to provide flexible and scalable computation resource provisioning for IoT networks. Kiani *et al.* [33] proposed an NOMA-based optimization framework for 5G networks, which aims at minimizing the energy consumption of MEC users through optimizing the user clustering, computation and communication resource allocation, and transmit power. Lyu *et al.* [34] proposed an asymptotically optimal scheduling scheme for MEC-aware IoT networks, in which the transmission time, energy intake, and data admission of all IoT devices were optimized in each time slot for maximizing a time-average network utility. Amjad *et al.* [35] proposed a cognitive edge-computing based framework for the efficient utilization of computation resource in IoT networks. Rodrigues *et al.* [36] proposed a technique for minimizing service delay in edge cloud computing through virtual machine migration and transmission power control. Wu *et al.* [37] proposed a secrecy-based resource management framework for computation offloading. Although the works in [32]–[37] have studied the joint communication and computation resource allocation problem for MEC-aware IoT networks, they usually consider the conventional multiple access techniques, such as NOMA with fixed SIC ordering [33] and TDMA [34]. Also, these work mainly focus on energy consumption minimization and system-utility maximization, and thus the task execution latency of all users cannot be strictly guaranteed, although it is a rigorous requirement on NB-IoT systems. To minimize the maximum task execution latency of computing 1-bit task across IoT devices, the joint SIC ordering and computation resource allocation should be carefully considered for MEC-aware NOMA NB-IoT networks, which motivates the work of this paper.

III. System Model and Problem Formulation

A. System Model

We consider a uplink MEC-aware NOMA NB-IoT network with a set $\mathcal{N} = \{1, \cdots, N\}$ of NB-IoT devices and a small-cell BS equipped with MEC units to execute the data computing for NB-IoT devices, as illustrated in Fig. 1. NB-IoT device n transmits a B_n-bit task data to the small-cell BS[2], where NOMA is exploited as the multiple access scheme. We define one time slot as the duration in which the set of NB-IoT devices transmitting task data to the small-cell BS keeps constant. We use $\mathcal{F}_1$ to indicate the SIC ordering used at the small-cell BS in the first time slot. Specifically, we have $\mathcal{F}_1 = d_1 \rightarrow \cdots \rightarrow d_N$, where d_n means the index of the nth device in the SIC ordering. Assume that a subset $\mathcal{N}_{t-1}$ of NB-IoT devices finish the task data transmission until time slot $t-1$, and the SIC ordering in the time slot t satisfies $\mathcal{F}_t = \mathcal{F}_1 \backslash \mathcal{N}_{t-1}$.

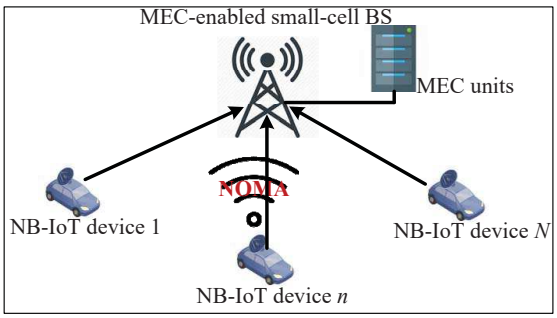

Fig. 1. A uplink MEC-aware NOMA NB-IoT network with N NB-IoT devices and a small-cell BS equipped with MEC units .

We use $\Delta B_{t,d_n}$ to indicate the backlog of task data that the NB-IoT device d_n have not transmitted up to time slot t. Obviously, if the NB-IoT device d_n has sent the whole task data to the small-cell BS (i.e., $\Delta B_{t,d_n} = 0$), it will terminate the task data transmission in time slot t, and thus its data rate is equal to zero. Let h_{d_n} denote the channel gain between the NB-IoT device d_n and the small-cell BS. Considering the NOMA scheme, the data rate of device d_n in time slot t can be expressed as

$$R_{t,d_n}(\mathcal{F}_1) = \begin{cases} 0, \text{if } \Delta B_{t,d_n} = 0 \\ W \log\left(1 + \frac{h_{d_n} p}{\sum\limits_{\substack{\forall j | n+1 < j \leq N \\ \text{and } \Delta B_{t,d_j} \neq 0}} h_{d_j} p + N_0}\right), \text{otherwise.} \end{cases} \tag{1}$$

Here, N_0 denotes the power of additive Gaussian noise at the small-cell BS, p denotes the transmit power used by each device, and W denotes the bandwidth occupied by N devices. By the definition of time slot, a new time slot begins when one NB-IoT completes its task data transmission at least. Thus, the length of time slot t can be expressed as

$$L_t(\mathcal{F}_1) = \min_{\forall n | \Delta B_{t,d_n} \neq 0} \frac{\Delta B_{t,d_n}}{R_{t,d_n}}, \tag{2}$$

and we have

$$\Delta B_{t+1,d_n} = \Delta B_{t,d_n} - L_t(\mathcal{F}_1) R_{t,d_n}. \tag{3}$$

Since the NB-IoT device d_n completes the task data transmission when $\Delta B_{t,d_n} = 0$, its transmission latency, denoted by $Tt_{d_n}(\mathcal{F}_1)$, is expressed as

$$Tt_{d_n}(\mathcal{F}_1) = \sum_{k=1}^{\underset{t}{\text{argmin}}\{t | \Delta B_{t,d_n} = 0\}} L_k(\mathcal{F}_1) \tag{4}$$

After the small-cell BS receives the whole task data of NB-IoT device d_n, its equipped MEC units will compute this task.

[2]In the calculation of task transmission latency, we do not take into account the packet header size of network protocols, and assume that the transmission data size is equal to the task data size.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

Let $C_{d_n}(\mathcal{F}_1)$ denote the computation resource allocated to computing the task from NB-IoT device d_n when the SIC ordering is $\mathcal{F}_1$. Thus, the time that the MEC units spend in computing the task from device d_n is expressed as

$$Tc_{d_n}(\mathcal{F}_1) = \frac{B_{d_n}}{C_{d_n}(\mathcal{F}_1)}, \forall n \in \mathcal{N}. \tag{5}$$

We consider a practical constraint that the MEC units have the computational capacity of C Mbps, i.e.,

$$\sum_{n=1}^{N} C_{d_n}(\mathcal{F}_1) \leq C. \tag{6}$$

Note that the computational capacity is evaluated as the ratio between the computational frequency (i.e., the number of CPU cycles per second) and the computational efficiency (i.e., the number of CPU cycles required for computing 1-bit task) of MEC units [38].

B. Problem Formulation

Since the size of task data that N NB-IoT devices transmit to the small-cell BS is different, we adopt the maximum task execution latency required for computing 1-bit task data across these NB-IoT devices to evaluate the computational efficiency of MEC-aware NOMA NB-IoT network. In this paper, we aim to minimize this latency through jointly optimizing the SIC ordering and computation resource allocation. Mathematically, we formulate the optimization problem in the following form

$$\begin{aligned} \textbf{P1:} \quad & \min_{\mathcal{F}_1, C_{d_n}(\mathcal{F}_1)'s} \max_n \frac{Tt_{d_n}(\mathcal{F}_1) + Tc_{d_n}(\mathcal{F}_1)}{B_{d_n}} \\ & s.t. \quad \sum_{n=1}^{N} C_{d_n}(\mathcal{F}_1) \leq C, \\ & \quad C_{d_n}(\mathcal{F}_1) \geq 0, \forall n \in \mathcal{N} \\ & \quad \mathcal{F}_1 \in \mathcal{P}, \end{aligned} \tag{7}$$

where $\mathcal{P}$ is the set of permutations of the set $\{1, \cdots, N\}$. It is worth noting that due to the fact that the computational time $Tc_{d_n}(\mathcal{F}_1)$ decreases with $C_{d_n}(\mathcal{F}_1)$, the first constraint in (7) is active when the optimal solution is obtained.

The following theorem shows the hardness of the optimization problem (7).

Theorem 1. The optimization problem in (7) is NP-hard.

Proof: The optimization problem in (7) involves the SIC ordering and computation resource allocation. Given the SIC ordering, the optimization of computation resource allocation is reduced to finding the root of polynomial equations, and thus it is convex in obtaining the corresponding maximum task execution latency per task bit. It implies that the optimization problem in (7) is NP-hard, if the optimization of SIC ordering is NP-hard. We next prove that the optimization of SIC ordering is reduced to the job-shop scheduling problem, for which the NP-hardness is proved in [39]. We consider an NB-IoT network with N NB-IoT devices. The mathematical statement of the SIC ordering optimization problem is to find an assignment of the device set $\{1, 2, \cdots, N\}$ to the ordering site set $\{d_1, d_2, \cdots, d_N\}$ such that the maximum task execution latency per task bit is a minimum. The device set and the ordering site set can be reducible to the job set and machine set in the job-shop scheduling problem, respectively. The maximum task execution latency per task bit corresponds to the cost function in the job-shop scheduling problem. Since that job-shop scheduling problem is to find an assignment of jobs to machines such that the cost function is minimized, the SIC ordering optimization problem is equivalent to the job-shop scheduling problem. Together with the convexity of obtaining the maximum task execution latency per task bit, it follows that the optimization problem in (7) is NP-hard. ■

IV. Algorithm Design

In this section, we first focus on obtaining the optimal solution to (7). To reduce the computational complexity, we then design an efficient heuristic algorithm for SIC ordering.

Due to the min-max objective function, given $\mathcal{F}_1$, the optimal solution $(C^*_{d_n}(\mathcal{F}_1))_{\forall d_n \in \mathcal{N}}$ to (7) is the root satisfying

$$\begin{cases} \frac{Tt_{d_1}(\mathcal{F}_1)}{B_{d_1}} + \frac{1}{C_{d_1}} = \cdots = \frac{Tt_{d_n}(\mathcal{F}_1)}{B_{d_n}} + \frac{1}{C_{d_n}} = \cdots \\ = \frac{Tt_{d_N}(\mathcal{F}_1)}{B_{d_N}} + \frac{1}{C_{d_N}} \\ \sum_{n=1}^{N} C_{d_n}(\mathcal{F}_1) = C, \\ C_{d_n}(\mathcal{F}_1) \geq 0, \forall n \in \mathcal{N}. \end{cases} \tag{8}$$

For the notational brevity, we let $\frac{Tt_{d_n}(\mathcal{F}_1)}{B_{d_n}} = \alpha_{d_n}$. By introducing a new variable β such that $\beta = \alpha_{d_n} + \frac{1}{C_{d_n}(\mathcal{F}_1)}$, we can obtain the root of (8) through solving

$$\begin{cases} \sum_{n=1}^{N} \frac{1}{\beta - \alpha_{d_n}} = C \\ \beta \geq \max_n \alpha_{d_n}. \end{cases} \tag{9}$$

The inequality in (9) implies that the computation resource allocation $C_{d_n}(\mathcal{F}_1)$ is non-negative for all n's. Note that although (9) is a polynomial equation of β, it is difficult to obtain the solution β in the closed-form expression when $N > 2$. Therefore, in the following, we obtain the optimal solution of (8) through solving (9) in the two-device case (i.e., $N = 2$) and multi-device case (i.e., $N > 2$), respectively.

A. Two-Device Case

By calculating (9) and $C_{d_n}(\mathcal{F}_1) = \frac{1}{\beta - \alpha_{d_n}}$, we have

$$\begin{aligned} C^*_{d_1}(\mathcal{F}_1) &= \begin{cases} \frac{C}{2} - \frac{1}{A_{\mathcal{F}_1}} + \sqrt{\frac{C^2}{4} + \frac{1}{A^2_{\mathcal{F}_1}}}, & \text{if } A_{\mathcal{F}_1} > 0 \\ \frac{C}{2} - \frac{1}{A_{\mathcal{F}_1}} - \sqrt{\frac{C^2}{4} + \frac{1}{A^2_{\mathcal{F}_1}}}, & \text{otherwise} \end{cases} \\ C^*_{d_2}(\mathcal{F}_1) &= \begin{cases} \frac{C}{2} + \frac{1}{A_{\mathcal{F}_1}} - \sqrt{\frac{C^2}{4} + \frac{1}{A^2_{\mathcal{F}_1}}}, & \text{if } A_{\mathcal{F}_1} > 0 \\ \frac{C}{2} + \frac{1}{A_{\mathcal{F}_1}} + \sqrt{\frac{C^2}{4} + \frac{1}{A^2_{\mathcal{F}_1}}}, & \text{otherwise} \end{cases} \end{aligned} \tag{10}$$

where $A_{\mathcal{F}_1}$ is equal to

$$A_{\mathcal{F}_1} = \frac{Tt_{d_1}(\mathcal{F}_1)}{B_{d_1}} - \frac{Tt_{d_2}(\mathcal{F}_1)}{B_{d_2}}. \tag{11}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

Since there are two NB-IoT devices, the number of time slots is two at most. Thus, we have

$$\begin{aligned} Tt_{d_1}(\mathcal{F}_1) =& \frac{L_1(\mathcal{F}_1)}{B_{d_1}} + \frac{B_{d_1} - R_{1,d_1}(\mathcal{F}_1)L_1(\mathcal{F}_1)}{R_{2,d_1}(\mathcal{F}_1)B_{d_1}} \\ Tt_{d_2}(\mathcal{F}_1) =& \frac{L_1(\mathcal{F}_1)}{B_{d_2}} + \frac{B_{d_2} - R_{1,d_2}(\mathcal{F}_1)L_1(\mathcal{F}_1)}{R_{2,d_2}(\mathcal{F}_1)B_{d_2}}. \end{aligned} \tag{12}$$

Together with the fact that the SIC ordering $\mathcal{F}_1$ is any one permutation of $\{1, 2\}$, it follows that the optimal SIC ordering $\mathcal{F}_1^*$ is obtained by

$$\mathcal{F}_1^* = \underset{\mathcal{F}_1}{\text{argmin}} \frac{L_1(\mathcal{F}_1)}{B_{d_1}} + \frac{B_{d_1} - R_{1,d_1}(\mathcal{F}_1)L_1(\mathcal{F}_1)}{R_{2,d_1}(\mathcal{F}_1)B_{d_1}} + \frac{1}{C_{d_1}^*(\mathcal{F}_1)}. \tag{13}$$

Accordingly, the optimal computation resource allocation is $C_{d_1}^*(\mathcal{F}_1^*)$ and $C_{d_2}^*(\mathcal{F}_1^*)$ by (10).

Next, we focus on analyzing the property of the optimal solution to (7) in the two-device case.

Theorem 2. When the size of task data that two NB-IoT devices transmit to the small-cell BS is equal, the optimal SIC ordering in (7) follows the descending order of channel gains.

Proof: Without loss of generality, we assume that $h_1 \geq h_2$, and $B_1 = B_2 = B$. We set $\mathcal{F}_{11} = 1 \rightarrow 2$ and $\mathcal{F}_{12} = 2 \rightarrow 1$. Given $\mathcal{F}_{11}$, if the NB-IoT device 2 completes the task data transmission before the NB-IoT device 1 (i.e., $\frac{h_1}{h_2} \leq \frac{ph_2}{N_0} + 1$), then the worst-case end-to-end delay can be expressed as

$$\begin{aligned} D_{\mathcal{F}_{11}} &= \frac{1}{W\log(1+\frac{ph_2}{N_0})} + \frac{1}{C_2^*(\mathcal{F}_{11})} \\ &= \frac{1}{W\log(1+\frac{ph_2}{N_0})} + A_{\mathcal{F}_{11}} + \frac{1}{C_1^*(\mathcal{F}_{11})} \end{aligned} \tag{14}$$

by (8), where $A_{\mathcal{F}_{11}}$ satisfies

$$A_{\mathcal{F}_{11}} = \frac{1}{W\log(1+\frac{ph_1}{N_0})} - \frac{\log(1+\frac{ph_1}{ph_2+N_0})}{W\log(1+\frac{ph_1}{N_0})\log(1+\frac{ph_2}{N_0})}. \tag{15}$$

When we adopt $\mathcal{F}_{12}$ as the SIC ordering, the worst-case end-to-end delay can be expressed as

$$\begin{aligned} D_{\mathcal{F}_{12}} &= \frac{1}{W\log(1+\frac{ph_1}{N_0})} + \frac{1}{C_1^*(\mathcal{F}_{12})} \\ &= \frac{1}{W\log(1+\frac{ph_1}{N_0})} + A_{\mathcal{F}_{12}} + \frac{1}{C_2^*(\mathcal{F}_{12})}, \end{aligned} \tag{16}$$

where $A_{\mathcal{F}_{12}}$ satisfies

$$A_{\mathcal{F}_{12}} = \frac{1}{W\log(1+\frac{ph_2}{N_0})} - \frac{\log(1+\frac{ph_2}{ph_1+N_o})}{W\log(1+\frac{ph_1}{N_0})\log(1+\frac{ph_2}{N_0})}. \tag{17}$$

Thus, we have

$$\begin{aligned} &2D_{\mathcal{F}_{11}} - 2D_{\mathcal{F}_{12}} \\ &= \frac{1}{C_2^*(\mathcal{F}_{11})} + \frac{1}{C_1^*(\mathcal{F}_{11})} - \frac{1}{C_1^*(\mathcal{F}_{12})} - \frac{1}{C_2^*(\mathcal{F}_{12})} \\ &= \frac{C}{C_1^*(\mathcal{F}_{11})C_2^*(\mathcal{F}_{11})} - \frac{C}{C_1^*(\mathcal{F}_{12})C_2^*(\mathcal{F}_{12})}. \end{aligned} \tag{18}$$

Due to the fact that $0 < A_{\mathcal{F}_{11}} \leq A_{\mathcal{F}_{12}}$, we have

$$C_1^*(\mathcal{F}_{11})C_2^*(\mathcal{F}_{11}) \geq C_1^*(\mathcal{F}_{12})C_2^*(\mathcal{F}_{12}) \tag{19}$$

by (10). It follows that $2D_{\mathcal{F}_{11}} - 2D_{\mathcal{F}_{12}} \leq 0$, and thus $D_{\mathcal{F}_{11}} \leq D_{\mathcal{F}_{12}}$. It implies that $\mathcal{F}_{11}$ is the optimal SIC ordering when $1 \leq \frac{h_1}{h_2} \leq \frac{ph_2}{N_0} + 1$.

On the other hand, given $\mathcal{F}_{11}$, if the NB-IoT device 1 completes the task data transmission before the NB-IoT device 2 (i.e., $\frac{h_1}{h_2} > \frac{ph_2}{N_0} + 1$), we have

$$\begin{aligned} &2D_{\mathcal{F}_{11}} - 2D_{\mathcal{F}_{12}} \\ &= \frac{C}{C_1^*(\mathcal{F}_{11})C_2^*(\mathcal{F}_{11})} - \frac{C}{C_1^*(\mathcal{F}_{12})C_2^*(\mathcal{F}_{12})} \\ &+ \frac{\log(\frac{ph_1+N_0}{N_0}\frac{ph_2+N_0}{ph_1+ph_2+N_0})}{W\log(1+\frac{ph_1}{N_0})} \\ &\times \underbrace{(\frac{1}{\log(1+\frac{ph_1}{ph_2+N_0})} - \frac{1}{\log(1+\frac{ph_2}{N_0})})}_{<0}. \end{aligned} \tag{20}$$

Because of $\frac{h_1}{h_2} > \frac{ph_2}{N_0} + 1$, we further have

$$A_{\mathcal{F}_{11}} = \frac{1}{W\log(1+\frac{ph_1}{ph_2+N_0})} - \frac{1}{W\log(1+\frac{ph_2}{N_0})} \leq 0. \tag{21}$$

Since we have $0 \leq |A_{\mathcal{F}_{11}}| \leq A_{\mathcal{F}_{12}}$, we have

$$\frac{C}{C_1^*(\mathcal{F}_{11})C_2^*(\mathcal{F}_{11})} - \frac{C}{C_1^*(\mathcal{F}_{12})C_2^*(\mathcal{F}_{12})} < 0. \tag{22}$$

by (10). Together with (20), we have $D_{\mathcal{F}_{11}} < D_{\mathcal{F}_{12}}$ when $\frac{h_1}{h_2} > \frac{ph_2}{N_0} + 1$.

Until now, we have proved that $D_{\mathcal{F}_{11}} < D_{\mathcal{F}_{12}}$ when $h_1 \geq h_2$. Therefore, Theorem 2 follows. ■

Based on the basic operations in (10)-(13) and Theorem 2, we present the procedure of obtaining the optimal solution to (7) in Algorithm 1 for two-device case.

It is worth noting that by Theorem 2, the optimal SIC ordering is determined according to the channel gains when two NB-IoT devices transmit equal-size task data to the small-cell BS. However, this conclusion cannot be extended to the multi-device case due to the NP-hardness of the min-max task execution latency optimization. In this regard, we need to resort to the numerical algorithm to obtain the optimal SIC ordering, which will be introduced latter.

Algorithm 1 The procedure of obtaining the optimal solution to (7) in the two-device case

1: **if** $B_1 = B_2$ **then**
2: $\quad \mathcal{F}_1^* = \{\underset{i=1,2}{\text{argmax}}\ h_i, \underset{i=1,2}{\text{argmin}}\ h_i\}$.
3: **else**
4: $\quad$ **for all** $\mathcal{F}_1 \in \{1 \rightarrow 2, 2 \rightarrow 1\}$ **do**
5: $\quad\quad$ Obtain $C_{d_1}^*(\mathcal{F}_1)$ and $C_{d_2}^*(\mathcal{F}_1)$ by (10).
6: $\quad$ **end for**
7: $\quad$ Obtain the optimal SIC ordering $\mathcal{F}_1^*$ by (13).
8: **end if**
9: Obtain the optimal computation resource allocation $C_{d_1}^*(\mathcal{F}_1^*)$ and $C_{d_2}^*(\mathcal{F}_1^*)$ by (10).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

B. Multi-Device Case

A close look at (9) finds that the number of fractional terms is equal to the number of NB-IoT devices. Thus, it is impossible to obtain the root of (9) in the closed-form expression. It implies that it is of practical importance to design an efficient numerical algorithm to obtain the root of (9), and thus the optimal solution to (7) in the multi-device case. In what follows, we focus on deriving the optimal solution. Furthermore, to reduce the computational complexity, a sub-optimal solution is proposed.

1) Optimal Solution

For the notational simplicity, we let $f(\beta) = \sum_{n=1}^{N} \frac{1}{\beta - \alpha_{d_n}}$. We have the first-order derivative of $f(\beta)$, denoted by $f'(\beta)$, satisfying

$$f'(\beta) = -\sum_{n=1}^{N} (\beta - \alpha_{d_n})^{-2}. \tag{23}$$

Due to the fact that $f'(\beta) < 0$, the function $f(\beta)$ is decreasing with β. Therefore, we can obtain the root of (9) based on the idea of bisection searching. The following Lemma 1 shows the range of the root of (9), which can be used to decide the initial upper bound and lower bound of β when running the bisection searching.

Lemma 1. The root of (9) resides in the rang of $\left[\max\left\{\max_n \alpha_{d_n}, \frac{N}{C} + \min_n \alpha_{d_n}\right\}, \frac{N}{C} + \max_n \alpha_{d_n}\right]$.

Proof: From the equation in (9), the set $\mathcal{N}$ can be divided into two disjoint subsets $\mathcal{N}_1$ and $\mathcal{N}_2$, which satisfy

$$\frac{1}{\beta(\mathcal{F}_1) - \alpha_{d_n}} \begin{cases} \geq \frac{N}{C}, & \text{if } n \in \mathcal{N}_1, \\ \leq \frac{N}{C}, & \text{if } n \in \mathcal{N}_2. \end{cases} \tag{24}$$

It follows that

$$\begin{aligned} \beta(\mathcal{F}_1) &\geq \frac{N}{C} + \max_{n \in \mathcal{N}_1} \alpha_{d_n} \geq \frac{N}{C} + \max_{n \in \mathcal{N}} \alpha_{d_n} \\ \beta(\mathcal{F}_1) &\leq \frac{N}{C} + \min_{n \in \mathcal{N}_2} \alpha_{d_n} \leq \frac{N}{C} + \min_{n \in \mathcal{N}} \alpha_{d_n}. \end{aligned} \tag{25}$$

Together with the fact that $\beta(\mathcal{F}_1) \geq \max_{n \in \mathcal{N}} \alpha_{d_n}$ in (9), we can get

$$\beta(\mathcal{F}_1) \in \left[\max\left\{\max_n \alpha_{d_n}, \frac{N}{C} + \min_n \alpha_{d_n}\right\}, \frac{N}{C} + \max_n \alpha_{d_n}\right], \tag{26}$$

and Lemma 1 follows. ■

In particular, the optimal computation resource allocation and SIC ordering algorithm works as follows. First, by Lemma 1, we can set the initial upper bound and lower bound of β. Second, we obtain the optimal computation resource allocation and worst-case end-to-end computational delay (i.e., the root of (9)) based on the idea of bisection search for any given SIC ordering. Finally, based on the root of (9) (say $\beta(\mathcal{F}_1)$), we can obtain the optimal SIC ordering $\mathcal{F}_1^*$ of (7) by the following formula.

$$\mathcal{F}_1^* = \underset{\mathcal{F}_1 \in \mathcal{P}}{\operatorname{argmin}} \frac{1}{\beta(\mathcal{F}_1) + \alpha_{d_1}}. \tag{27}$$

Algorithm 2 The optimal computation resource allocation and SIC ordering algorithm in the multi-device case

1: **for all** $\mathcal{F}_1 \in \mathcal{P}$ **do**
2: **Initialization:** Set the lower bound $\beta^L = \max\left\{\max_n \alpha_{d_n}, \frac{N}{C} + \min_n \alpha_{d_n}\right\}$, the upper bound $\beta^U = \frac{N}{C} + \max_n \alpha_{d_n}$, and the stopping tolerance ϵ.
3: **if** $|f(\beta^{(k-1)}) - C| \geq \epsilon$ **then**
4: Set $\beta = \frac{\beta^L + \beta^U}{2}$.
5: **if** $f(\beta^{(k-1)}) - C > 0$ **then**
6: Set $\beta^L = \beta$.
7: **else**
8: Set $\beta^U = \beta$.
9: **end if**
10: **else**
11: Terminate the calculation and obtain the root of (9) (say $\beta(\mathcal{F}_1)$) as β.
12: **end if**
13: **end for**
14: Obtain the optimal SIC ordering $\mathcal{F}_1^*$ by (27).
15: Obtain the optimal computation resource allocation, i.e., $C_{d_n}^*(\mathcal{F}_1^*) = \frac{1}{\beta(\mathcal{F}_1^*) - \alpha_{d_n}}$ for all $n \in \mathcal{N}$.

In details, the optimal computation resource allocation and SIC ordering (i.e., the optimal solution to (7)) can be obtained by Algorithm 2 in the multi-device case.

It is worth noting from Algorithm 2 that although the optimal computation resource allocation can be efficiently obtained with the SIC ordering $\mathcal{F}_1$, the optimal SIC ordering is obtained based on the emulation over $\mathcal{P}$. Due to the NP-hardness, the computational complexity of seeking the optimal SIC ordering is exponentially increasing with the number of NB-IoT devices. In the practical implementation, we want to obtain a sub-optimal SIC ordering with the low computational complexity. Therefore, we next focus on the derivation of sub-optimal SIC ordering.

2) Sub-optimal Solution

To reduce the complexity of seeking the optimal SIC ordering, we propose to sort the SIC order of all NB-IoT devices one device by one device based on the greedy meta-scheduling idea [40]. The rationale that serves as motivation for this proposal is that the SIC ordering is processed in arrival order of all NB-IoT devices, and each newly arriving NB-IoT device is assigned to the SIC order with the optimal min-max execution latency. The key idea is as follows. Assume that we have set the SIC ordering of NB-IoT devices $1, \cdots, n$, say $\mathcal{F}_1(n)$. We then decide the SIC order of NB-IoT device $n+1$ through comparing the optimal β over all $n+1$ SIC orderings of NB-IoT devices $1, \cdots, n+1$, where each SIC ordering is obtained by inserting NB-IoT device $n+1$ in the SIC ordering $\mathcal{F}_1(n)$. Specifically, the SIC ordering with the minimum optimal β is set as the optimal SIC ordering of NB-IoT devices $1, \cdots, n+1$, i.e., $\mathcal{F}_1(n+1)$. In doing this, we can obtain the sub-optimal SIC ordering of N NB-IoT devices as $\mathcal{F}_1(N)$. For example of a 5-device network, the procedure of obtaining $\mathcal{F}_1(5)$ is illustrated in Fig. 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

The details of low-complexity computation resource allocation and SIC ordering algorithm is presented in Algorithm 3. In the practical implementation, every newly arriving NB-IoT device first informs its channel state and task data size to the small-cell BS. Second, the small-cell BS determines the SIC order of this newly arriving NB-IoT device by inserting it in the SIC ordering of all existing NB-IoT devices. Third, the small-cell BS updates the optimal SIC ordering to all NB-IoT devices, and predicts the transmission data rate and the optimal computation resource allocation of all NB-IoT devices accordingly. Finally, the small-cell informs the transmission data rate to each NB-IoT device, and all NB-IoT devices transmits their task data simultaneously to the small-cell BS for the computation in MEC units.

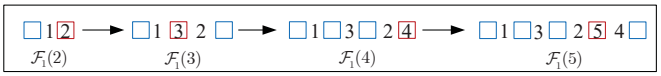

Fig. 2. An illustration of obtaining $\mathcal{F}_1(5)$ in a 5-device network. Here, all squares mean the possible ordering positions of decoding NB-IoT device $n+1$ and the red square means the best ordering position of decoding NB-IoT device $n+1$. At the nth step, the best ordering of NB-IoT devices $1, \cdots, n+1$ (i.e., $\mathcal{F}_1(n+1)$) is obtained through comparing the optimal worst-case end-to-end computational delay over $n+1$ possible ordering positions.

Algorithm 3 The low-complexity computation resource allocation and SIC ordering algorithm in the multi-device case

for all NB-IoT device $n = 1, \cdots, N$ **do**
  if $n == 1$ **then**
    Set $\mathcal{F}_1(1) = d_1 = 1$.
  else
    for all Inserting site $i = 1, \cdots, n$ **do**
      Set $\mathcal{G}(i) = d_1 \rightarrow \cdots \rightarrow d_{i-1} \rightarrow n \rightarrow d_i \rightarrow \cdots \rightarrow d_{n-1}$
      Obtain the optimal β over $\mathcal{G}(i)$ by running Steps 1-11 in Algorithm 2, say $\beta(i)$.
    end for
    Obtain the best site of inserting NB-IoT device n, $i^* = \underset{i}{\text{argmax}}\beta(i)$, and set $\mathcal{F}_1(n) = \mathcal{G}(i^*)$.
  end if
end for
Set $\mathcal{F}_1(N)$ as the sub-optimal SIC ordering, and set $\beta(i^*)$ as the sub-optimal worst-case end-to-end computational delay.
Obtain the optimal computation resource allocation, i.e., $C_{d_n}(\mathcal{F}_1(N)) = \frac{1}{\beta(i^*)-\alpha_{d_n}}$ for all $n \in \mathcal{N}$.

The following Theorem 3 shows the computational complexity of Algorithm 3.

Theorem 3. The computational complexity of the proposed low-complexity computation resource allocation and SIC ordering algorithm (i.e., Algorithm 3) is of $O(-N^2 \log(\epsilon))$ iterations.

Proof: When we decide the optimal SIC order of NB-IoT device n in Algorithm 3, we need to run the bisection searching n times. Thus, we run the bisection searching $\frac{N(N+1)}{2}$ times in total, when we obtain $\mathcal{F}_1(N)$. Since the computational complexity of bisection searching is of $O(-\log(\epsilon))$ iterations, it follows that the computational complexity of Algorithm 3 is of $O(-N^2 \log(\epsilon))$ iterations. Therefore, Theorem 3 follows. ■

TABLE I
SIMULATION PARAMETERS

Simulation parameters	Value settings
Carrier frequency	2000 MHz
Path loss model	$(38 + 30 \log_{10}(d))$ dB (d in meters)
Bandwidth W	15 kHz
Noise power spectral density	-174 dBm/Hz
Transmit power of NB-IoT devices	23 dBm
The task data size of NB-IoT device n	Uniformly distributed in $[0, 1]$ Mb
Stopping tolerance ϵ	10^{-4}

V. Numerical Results

In this section, we will evaluate the performance of the proposed algorithms. According to the NB-IoT parameters in [7], we set the simulation parameters in Table I. In the following simulations, we consider a set of uplink MEC-aware NB-IoT networks as shown in Fig. 3, where the MEC-enabled small-cell BS is placed in (200m, 200m), and N NB-IoT devices are uniformly deployed in a circle area with the radius of 200m and center of (200m, 200m).

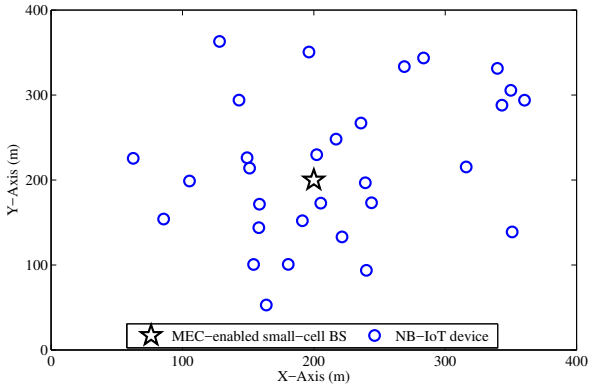

Fig. 3. The network topology used for simulations.

A. Performance Evaluation

Example 1: We start with verifying the optimality of Algorithm 1 in the two-device case. We place NB-IoT device 1 at the location of (100m, 200m), and move NB-IoT device 2 from (250m, 200m) to (300m, 200m) along the line. The computational capacity of MEC units is set to be 1 Mbps. For the verification target, we use Algorithm 2 as the benchmark. Table II shows the optimal SIC ordering and computation resource allocation obtained by Algorithm 1 and Algorithm 2. We can see that the optimal computation resource allocation in two-device case can be obtained in the closed-form expression

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

TABLE II
THE OPTIMALITY VERIFICATION IN THE TWO-DEVICE CASE

Location of NB-IoT device 2	Algorithm 1		Algorithm 2	
	$\mathcal{F}_1^*$	$\{C_1^*(\mathcal{F}_1^*), C_2^*(\mathcal{F}_1^*)\}$	$\mathcal{F}_1^*$	$\{C_1^*(\mathcal{F}_1^*), C_2^*(\mathcal{F}_1^*)\}$
(50m, 200m)	{2, 1}	{0.2583, 0.7417} Mbps	{2, 1}	{0.2583, 0.7417} Mbps
(70m, 200m)	{2, 1}	{0.2371, 0.7629} Mbps	{2, 1}	{0.2371, 0.7629} Mbps
(90m, 200m)	{2, 1}	{0.2229, 0.7771} Mbps	{2, 1}	{0.2229, 0.7771} Mbps
(110m, 200m)	{1, 2}	{0.7806, 0.2194} Mbps	{1, 2}	{0.7806, 0.2194} Mbps
(130m, 200m)	{1, 2}	{0.7712, 0.2288} Mbps	{1, 2}	{0.7712, 0.2288} Mbps
(150m, 200m)	{1, 2}	{0.7712, 0.2288} Mbps	{1, 2}	{0.7712, 0.2288} Mbps

given in Algorithm 1. Also, we can see that the optimal SIC ordering of two NB-IoT devices follows the descending order of channel gains when the task data size of these two NB-IoT devices is equal, which coincides with Theorem 2.

Example 2: In this simulation, we want to evaluate the performance of the proposed low-complexity algorithm (i.e., Algorithm 3). For the comparison target, we adopt the optimal algorithm (i.e., Algorithm 2) as the benchmark. In Fig. 4, each point is obtained by averaging over 100 different topologies of the same device density. From Fig. 4(a), we can see that Algorithm 3 can perform very close to the optimal min-max task execution latency, which is obtained by the optimal algorithm (i.e., Algorithm 2). For example, the largest performance degradation of Algorithm 3 is 0.24%, which is obtained when the number of NB-IoT devices is 8. However, Fig. 4(b) shows that the computational complexity of Algorithm 2 and Algorithm 3 is increasing exponentially and increasing quadratically with the increase of the number of NB-IoT devices, respectively. Due to the negligible performance degradation and low complexity, Algorithm 3 can be considered as the optimal computation resource allocation and SIC ordering algorithm in the practical implementation.

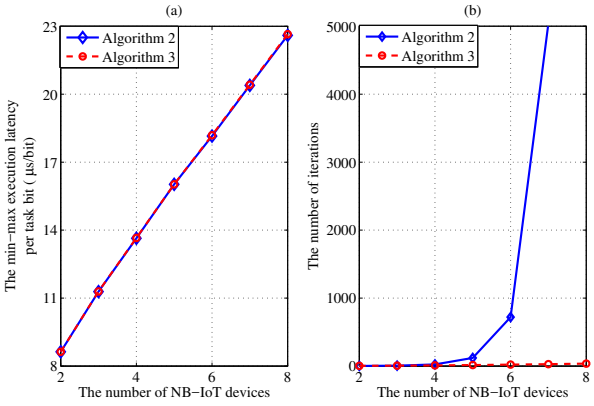

Fig. 4. The performance and computational complexity of Algorithm 2 and Algorithm 3.

B. Performance Comparison

To the best of our knowledge, there is no algorithm proposed for the same target through jointly optimizing SIC ordering and computation resource allocation in the literature. For the comparison with our proposed computation resource allocation and SIC ordering algorithm (i.e., Algorithm 3), we therefore introduce three baseline schemes, i.e., Scheme 1, Scheme 2, and Scheme 3. In these schemes, the SIC ordering follows the descending order of channel gains, the ascending order of channel gains, and the ascending order of task size, respectively, and then the optimal computation resource allocation is performed by the bisection searching. In the following simulations, we compare the min-max task execution latency obtained by these four algorithms.

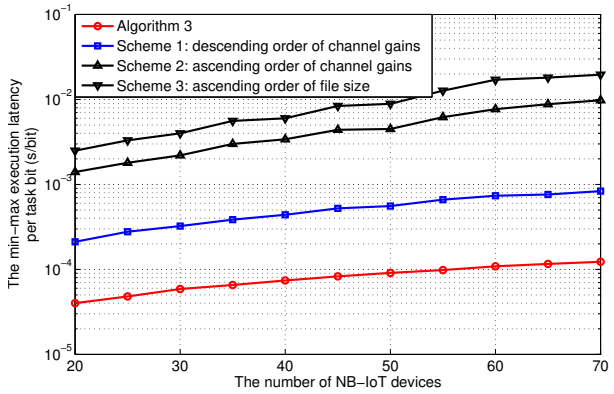

Fig. 5. The min-max execution latency per task bit obtained by four algorithms at different densities of NB-IoT devices.

Example 3 (Performance comparison at different densities of NB-IoT devices): We consider a set of uplink MEC-aware NB-IoT networks, as illustrated in Fig. 3. We vary the number of NB-IoT devices from 20 to 70. The computational capacity of MEC units is set to be 10 Mbps. Each point in Fig. 5 is obtained by averaging over 100 different topologies with the same density of NB-IoT devices.

Fig. 5 shows that as the number of NB-IoT devices increases, the obtained min-max task execution latency increases for all four algorithms. It can also be seen that compared with Algorithm 3, Scheme 1, Scheme 2, and Scheme 3 increase the min-max task execution latency by 5 times, 50 times, and 100 times on average, respectively. This observation implies that the SIC ordering would have a measurable impact on the min-max task execution latency, when we adopt the power-domain NOMA in uplink MEC-aware NB-IoT networks. Therefore, it is of practical meaning to optimize the SIC ordering, instead of following the order of channel gains or task size.

Example 4 (Performance comparison at different computational capacity): We consider a set of uplink MEC-aware NB-IoT networks with 50 NB-IoT devices, as illustrated in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

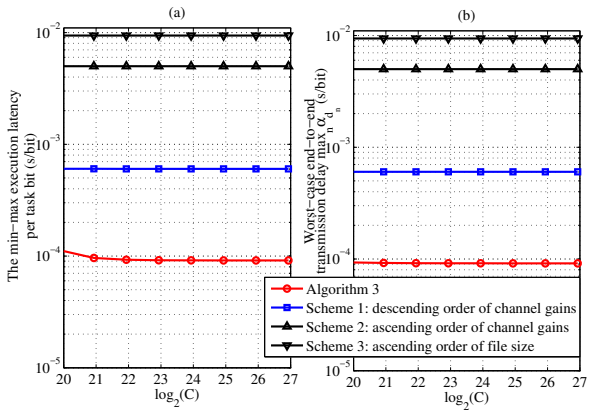

Fig. 6. The min-max execution latency per task bit obtained by four algorithms under different computational capacity.

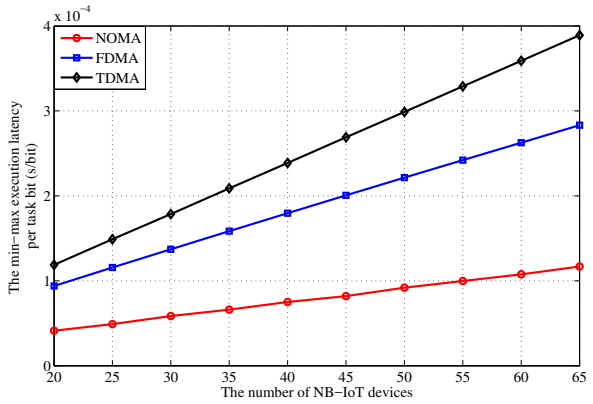

Fig. 7. The min-max execution latency per task bit obtained by NOMA, FDMA, and TDMA at different densities of NB-IoT devices.

Fig. 3. We vary the setting of computational capacity of MEC units from 1 Mbps to 128 Mbps. Each point in Fig. 6 is obtained by averaging over 100 different topologies with the same computational capacity.

Fig. 6(a) shows that with the increase of computational capacity of MEC units, the obtained min-max task execution latency slowly decreases only for the proposed Algorithm 3. On the contrary, the obtained min-max task execution latency almost keeps constant for all other three algorithms. It can be seen from Fig. 6(b) that the min-max task execution latency (i.e., $\max_n \alpha_{d_n}$) is the main bottleneck of decreasing the min-max task execution latency due to the low data rate in the NB-IoT network. Therefore, it is difficult to decrease the min-max task execution latency through increasing the computational capacity of MEC units for the MEC-aware NB-IoT network. Nevertheless, we see from Fig. 6 that we can effectively decrease the min-max task execution latency through decreasing the min-max task execution latency with the optimal SIC ordering. In particular, compared with Scheme 1, Scheme 2, and Scheme 3, Algorithm 3 can reduce the min-max task execution latency by 10 times, 50 times, and 100 times, respectively. This observation implies that the SIC ordering of NB-IoT devices should be optimized to minimize the min-max task execution latency according to their channel gains and task size in the MEC-aware NB-IoT network.

Example 5 (Performance comparison with other multiple access techniques): In the following simulations, we want to compare the min-max task execution latency obtained by NOMA, FDMA and TDMA. Specifically, the min-max task execution latency is obtained by Algorithm 3 for NOMA, the min-max task execution latency is obtained by equal bandwidth allocation among NB-IoT devices for FDMA, and the min-max task execution latency is obtained by equal time allocation among NB-IoT devices for TDMA.

We set the computational capacity of MEC units to be 10 Mbps. Fig. 7 shows the min-max task execution latency obtained under the different densities of NB-IoT devices from 20 to 65, when NOMA, FDMA and TDMA are applied to the network topology in Fig. 3. Each point in Fig. 7 is an average over 100 different topologies with the same density of NB-IoT devices. From Fig. 7, we see that the min-max task execution latency almost linearly increases with the increase of the number of NB-IoT devices for all three multiple access techniques. This is because that given the transmission resource (e.g., the bandwidth for FDMA, the transmission time for TDMA, and the SINR for NOMA) allocation and the computation resource allocation of MEC units, the task transmission latency and task computation latency of each NB-IoT device almost linearly increases with the increase of the number of NB-IoT devices. Together with the fact that the task execution latency is the sum of the task transmission latency and task computation latency, we have the observation in Fig. 7. Also, we can see that the NOMA always outperforms the FDMA and TDMA in terms of min-max task execution latency. Compared with FDMA and TDMA, the NOMA can reduce the task execution latency by 58.8% and 69.2% on average, respectively.

We vary the computational capacity of MEC units from 1 Mbps to 128 Mbps, and set the number of NB-IoT devices to be 50. Fig. 8 shows the min-max task execution latency obtained by NOMA, FDMA and TDMA for the network topology in Fig. 3 under different computational capacity. Each point in Fig. 8 is an average over 100 different topologies with the same computational capacity. We see that the min-max task execution latency decreases with the increase of computational capacity for all three multiple access technologies. This is because that the increase of computational capacity can reduce the task execution latency at MEC units. Also, we can see that the NOMA always outperforms the FDMA and TDMA in terms of min-max task execution latency. Compared with FDMA and TDMA, the NOMA can reduce the task execution latency by 58.2% and 68.5% on average, respectively. The observations from Figs. 7 and 8 reveal that considering the limited bandwidth in NB-IoT networks, the NOMA can be exploited as a promising multiple access technology for NB-IoT networks.

VI. Conclusions

In this paper, we have investigated the minimization of maximum task execution latency per task bit across devices for uplink MEC-aware NOMA NB-IoT networks by jointly considering SIC ordering and computation resource allocation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2875046, IEEE Internet of Things Journal

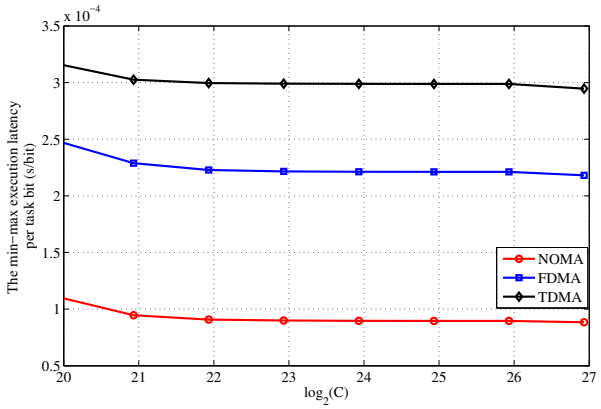

Fig. 8. The min-max execution latency per task bit obtained by NOMA, FDMA, and TDMA under different computational capacity.

Particularly, the problem has been proved to be reducible to the job-shop scheduling problem, and it is NP-hard. To obtain the optimal solution, we have exploited the decomposition optimization technique to solve the computation resource allocation and SIC ordering sequentially. Further, we have exploited the greedy meta-scheduling technique to devise a low-complexity and easy-implemented SIC ordering algorithm. Only according to the SIC ordering of existing NB-IoT devices, the SIC order of a newly arriving device can be determined with a negligible performance degradation. Finally, simulation results have verified the effectiveness of the proposed algorithm by comparing it with other multiple access schemes. In the future work, we will study the dynamic optimization of device scheduling in which all NB-IoT devices should decide which devices to transmit together, and how long to transmit together. To this end, the device scheduling can be optimized through solving the user grouping and time allocation sequentially.

REFERENCES

[1] L. P. Qian, Y. J. Zhang, J. Huang, and Y. Wu, "Demand Response Management via Real-Time Electricity Price Control in Smart Grids," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1268-1280, Jul. 2013.

[2] C. Wang, Z. Bi, and L. D. Xu, "IoT and cloud computing in automation of assembly modeling systems," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1426-1434, May 2014.

[3] A. Zanella, N. Bui, A. Castellani *et al.*, "Internet of things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22-32, Feb. 2014.

[4] L. Kong, M. K. Khan, F. Wu, G. Chen, and P. Zeng, "Millimeter-wave wireless communications for IoT-cloud supported autonomous vehicles: overview, design, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 62-68, Jan. 2017.

[5] M. Chen, Y. Ma, Y. Li *et al.*, "Wearable 2.0: enabling human-cloud integration in next generation healthcare systems," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 54-61, Jan. 2017.

[6] S. Verma, Y. Kawamoto, Z. M. Fadlullah *et al.*, "A survey on network methodologies for real-time analytics of massive IoT data and open research issues," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1457-1477, Third Quarter 2017.

[7] J. Chen, K. Hu, Q. Wang, Y. Sun, Z. Shi, and S. He, "Narrowband internet of things: implementations and applications," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2309-2314, Dec. 2017.

[8] F. Tang, B. Mao, Z. M. Fadlullah, and N. Kato, "On a novel deep learning based intelligent partially overlapping channel assignment in SDN-IoT," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 80-86, Sept. 2018.

[9] X. Yang, X. Wang, Y. Wu *et al.*, "Small-cell assisted secure traffic offloading for Narrowband Internet of Thing (NB-IoT) systems," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1516-1526, Jun. 2018.

[10] A. Hoglund, X. Lin, O. Liberg *et al.*, "Overview of 3GPP Release 14 enhanced NB-IoT," *IEEE Network*, vol. 31, no. 6, pp. 16-22, Nov./Dec. 2017.

[11] F. Tang, Z. M. Fadlullah, B. Mao, and N. Kato, "An intelligent traffic load prediction based adaptive channel assignment algorithm in SDN-IoT: a deep learning approach," *IEEE Internet Things J.*, pp. 1-1. DOI: 10.1109/JIOT.2018.2838574, 2018.

[12] L. P. Qian, Y. Wu, H. Zhou, and X. S. Shen, "Dynamic cell association for non-orthogonal multiple-access V2S networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2342-2356, Oct. 2017.

[13] H. Guo, J. Liu, J. Zhang, W. Sun, and N. Kato, "Mobile-edge computation offloading for ultra-dense IoT networks," *IEEE Internet Things J.*, pp. 1-1, DOI: 10.1109/JIOT.2018.2838584, 2018.

[14] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2771-2784, Dec. 2017.

[15] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438-4454, Jun. 2016.

[16] T. G. Rodrigues, K. Suto, H. Nishiyama *et al.*, "Cloudlets activation scheme for scalable mobile edge computing with transmission power control and virtual machine migration," *IEEE Trans. Computers*, vol. 67, no. 9, pp. 1287-1300, Sept. 2018.

[17] Y. Wu, L. Qian, H. Mao, X. Yang, and X. S. Shen, "Optimal Power Allocation and Scheduling for Non-Orthogonal Multiple Access Relay-Assisted Networks," *IEEE Trans. Mobile Computing*, pp. 1-1, DOI:10.1109/TMC.2018.2812722, 2018.

[18] M.Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55-61, Sept. 2017.

[19] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667-2680, Dec. 2017.

[20] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319-5332, Dec. 2017.

[21] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191-5202, Oct. 2017.

[22] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1465-1468, Jul. 2016.

[23] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721-742, Second Quarter 2017.

[24] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686-7698, Nov. 2016.

[25] L. P. Qian, Y. Wu, H. Zhou, and X. S. Shen, "Joint uplink base station association and power control for small-cell networks with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5567-5582, Sept. 2017.

[26] L. You, D. Yuan, L. Lei *et al.*, "Resource optimization with load coupling in multi-cell NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4735-4749, Jul. 2018.

[27] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393-1405, 2016.

[28] A. E. Mostafa, Y. Zhou, and V. W. S. Wong, "Connectivity maximization for narrowband IoT systems with NOMA," in *Proc. of IEEE ICC 2017*, pp. 1-6, May 2017.

[29] Q. Wu, W. Chen, D. W. K. Ng, and R. Schober, "Spectral and energy efficient eireless powered IoT networks: NOMA or TDMA?," *IEEE Trans. Veh. Tech.*, vol. 67, no. 7, pp. 1-5, Jul. 2018.

[30] D. Zhai, R. Zhang, L. Cai, B. Li, and Y. Jiang, "Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive IoT devices," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1299-1306, Jun. 2018.

[31] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854-864, Dec. 2016.

[32] X. Sun and N. Ansari, "EdgeIoT: mobile edge computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 22-29, Dec. 2017.

[33] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299-1306, Apr. 2018.

[34] X. Lyu, W. Ni, H. Tian *et al.*, "Optimal schedule of mobile edge computing for Internet of Things using partial information," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2342-2356, Nov. 2017.
[35] A. Amjad, F. Raaby, S. Sadia *et al.*, "Cognitive edge computing based resource allocation framework for Internet of Things," in *Proc. of FMEC 2017*, pp. 1-7, May 2017.
[36] T. G. Rodrigues, K. Suto, H. Nishiyama, and N. Kato, "Hybrid method for minimizing service delay in edge cloud computing through VM migration and transmission power control," *IEEE Trans. Computers*, vol. 66, no. 5, pp. 810-819, May 2017.
[37] Y. Wu, L. Qian, H. Mao *et al.*, "Secrecy-driven resource management for vehicular computation-offloading networks," *IEEE Network*, vol. 32, no. 3, pp. 84-91, Jun. 2018
[38] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784-1797, Mar. 2018.
[39] Y. N. Sotskov and N. V. Shakhlevich, "NP-hardness of shop-scheduling problems with three jobs," *Discrete Applied Math.*, vol. 59, no. 3, pp. 237-266, May 1995.
[40] G. Sabin, R. Kettimuthu, A. Rajan, and P. Sadayappan, "Scheduling of parallel jobs in a heterogenous multi-site enviroments," *Job Scheduling Strategies for Parallel Processing.* vol. 2862, pp. 87-104, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2003.

Li Ping Qian (M'10-SM'16) received the PhD degree in Information Engineering from the Chinese University of Hong Hong, Hong Kong, in 2010. She worked as a postdoctoral research associate at the Chinese University of Hong Kong, Hong Kong, during 2010-2011. Since 2011, she has been with College of Information Engineering, Zhejiang University of Technology, China, where she is currently a full Professor. From 2016 to 2017, she was a visiting scholar with the Broadband Communications Research Group, ECE Department, University of Waterloo. Her research interests include wireless communication and networking, resource management in wireless networks, massive IoTs, mobile edge computing, emerging multiple access techniques, and machine learning oriented towards wireless communications. She was a co-recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2011, the Best Paper Award from IEEE ICC 2016, and the Best Paper Award from IEEE Communication Society GCCTC 2017. She is currently on the Editorial Board of IET Communications.

Anqi Feng received the B.E. degree in Control Science and Engineering from Quzhou University, Quzhou, China, in 2017. She is currently working towards the M. S. degree in Control Science and Engineering in the Zhejiang University of Technology, Hangzhou, China. Her research interests lie in the areas of networking and intelligent system.

Yupin Huang received the B.E. degree in Control Science and Engineering from Renai College, Tianjin University ,Tianjin ,china, in 2017. He is currently working towards the M. S. degree in Control Science and Engineering in the Zhejiang University of Technology, Hangzhou, China. His research interests lie in the areas of vehicular networking systems, including traffic control and vehicle speed prediction.

Yuan Wu (M'10-SM'16) received the Ph.D degree in Electronic and Computer Engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2010. He is currently an Associate Professor in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. During 2010-2011, he was the Postdoctoral Research Associate at the Hong Kong University of Science and Technology. During 2016-2017, he was with the Broadband Communications Research (BBCR) group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests focus on resource management for wireless communications and networks, and smart grid. He is the recipient of the Best Paper Award from IEEE ICC in 2016 and the Best Paper Award from IEEE Communication Society GCCTC 2017.

Bo Ji (M'12-SM'18) received his B.E. and M.E. degrees in Information Science and Electronic Engineering from Zhejiang University, Hangzhou, China, in 2004 and 2006, respectively, and his Ph.D. degree in Electrical and Computer Engineering from The Ohio State University, Columbus, OH, USA, in 2012. Dr. Ji joined Department of Computer and Information Sciences (CIS) at Temple University in July 2014, where he is currently an assistant professor. He is also a faculty member of the Center for Networked Computing (CNC) at Temple University. Prior to joining Temple University, he was a Senior Member of the Technical Staff with AT&T Labs, San Ramon, CA, from January 2013 to June 2014. His research interests are in the modeling, analysis, control, optimization, and learning of computer and networking systems, such as communication networks, information-update systems, cloud/datacenter networks, and cyber-physical systems. Dr. Ji is a senior member of the IEEE and a member of the ACM. He is a National Science Foundation (NSF) CAREER awardee (2017) and an NSF CISE Research Initiation Initiative (CRII) awardee (2017).

Zhiguo Shi (M'09-SM'15) received his B.S. and Ph.D. degrees in electronic engineering from Zhejiang University, Hangzhou, China, in 2001 and 2006, respectively. Since 2006, he has been a faculty member with the College of Information Science and Electronic Engineering, Zhejiang University, where he is currently a full Professor. He was visiting broadband communications research group at University of Waterloo from September 2011 to November 2013. Now he is serving as an Editor for IEEE Network, and IET Communications. He is also a General Co-Chair of the 2020 IEEE Sensor Array and Multichannel Signal Processing Workshop.