IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received May 8, 2018, accepted June 20, 2018, date of publication June 28, 2018, date of current version July 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2851392

Resource Modeling and Scheduling for Mobile Edge Computing: A Service Provider's Perspective

SHUAISHUAI GUO[1], (Member, IEEE), DALEI WU[2], (Member, IEEE), HAIXIA ZHANG[1,3], (Senior Member, IEEE), AND DONGFENG YUAN[1], (Senior Member, IEEE)

[1]Shandong Provincial Key Laboratory of Wireless Communication Technologies, Shandong University, Qingdao 266237, China
[2]Department of Computer Science and Engineering, The University of Tennessee at Chattanooga, Chattanooga, TN 37403, USA
[3] School of Control Science and Engineering, Shandong University, Jinan 250061, China

Corresponding author: Haixia Zhang (haixia.zhang@sdu.edu.cn)

This work was supported in part by the National Science Fund of China for Excellent Young Scholars under Grant 61622111, in part by the National Natural Science Foundation of China under Grant 61671278, in part by the China Postdoctoral Science Foundation under Grant 2017M622202, and in part by the Shandong Natural Science Foundation under Grant ZR2018BF033. The work of D. Wu was supported in part by the National Science Fund of America under Grant 1647175 and in part by UTC CEACSE. T.

ABSTRACT This paper investigates resource modeling and management for a base station (BS) providing mobile edge computing (MEC) service. In the proposed modeling, BS is recognized as a queueing network consisting of multiple multi-type servers. The uplink transmission users, downlink transmission users, and MEC users with different priority levels are jointly considered. It is assumed that their service-requests arrive dynamically and are also served dynamically. With such a general resource modeling, the interaction among these users can be analyzed based on the queueing network theory. The average delay of each service-type with different priority levels is derived. Based on the derived results, two resource management optimization problems are formulated and solved from the perspective of a service provider. The revenue brought by MEC services is first maximized by doing user admission control while provisioning the quality-of-service (QoS) of all admitted users with the given amount of communication and computation resources. Then, the capital expenditure of resource deployment is minimized by satisfying the QoS of all users. It is formulated as an integer programming problem. An algorithm is developed to solve it, which can help service providers to determine the optimal amount of communication and computation resources to be placed in a BS to guarantee QoS for all users at a minimal total capital expenditure. Computer simulations are done to validate all analysis and comparisons are made with BS serving multi-type users of single priority level. Through comparison, an insight is gained that service providers can obtain more revenue or spare less capital expenditure by differentiating the user priority levels.

INDEX TERMS Mobile edge computing (MEC), queueing network model, admission control, resource management, quality-of-service (QoS), latency.

I. INTRODUCTION

Mobile networks provide worldwide coverage and mobility support. Now they are pursuing not only higher transmission rate but also lower transmission latency [1], [2]. In the evolution from the second generation (2G) to the fifth generation (5G), much effort has been focused on the development of transmission technologies, such as, multiple-input multiple-output (MIMO), millimeter wave (mmWave) transmission, etc [2]–[5], to obtain higher transmission rate. In some future communication scenarios, some data-sensitive delay-sensitive applications such as virtual reality, real-time control, etc, require that the communication latency should be no longer than 1 millisecond (ms). It is impossible to support those applications by using network communications between user devices, i.e., the source of data, and remote cloud computing severs due to the large communication delay resulting from long transmission distances. To overcome this issue, cloud-computing capabilities should be brought in close proximity to end devices. This motivates the deployment of mobile edge computing (MEC) centers [6]–[8].

Unlike conventional cloud computing centers having plenty of computation resource, a MEC center typically has limited communication and computation resource. To facilitate energy-efficient and low-latency MEC for multiple users, a plenty of work [9]–[21] has been done on the computation offloading design, the joint communication and computation resource management as well as the admission control. For instance, Chen *et al.* [9] proposed a game-theoretic computation offloading method in multi-user MEC systems, and this study was extended to multi-cell settings in [10]. In these works, the service request and network are assumed to be deterministic. With both stochastic characteristics and network dynamics considered, Lyapunov optimization based computation offloading approaches were proposed for MEC systems in [13]–[16]. Instead of controlling the offloading workload, [22] investigated the joint computation and transmission resource allocation to reduce the sum energy consumption of all mobile users. Different from most of work focusing on energy-delay trade off, [21] investigated the optimal resource allocation to maximize the revenue of service providers under the constraints of quality of service (QoS) for all mobile users. These works have gained a lot of insights on the offloading decision making and the resource management, and have been well summarized in a comprehensive survey paper [8]. However, all these works overlooked the interact among MEC service, traditional cellular uplink transmission (UT) service (e.g., file or message uploading), and traditional cellular downlink transmission (DT) service (e.g., file downloading or media streaming). It is almost certain that UT users, DT users and MEC user will interact on each other since they compete for the scarce and precious wireless transmission resources.

To analyze the interact, this paper investigates a queueing network model [23] for a base station (BS) providing MEC services, where the BS is treated as multi-type servers including UT servers, DT servers, and computation servers. It is assumed that each service request is represented by a packet which is stochastically generated by a user and has to wait in a queue if the corresponding server is busy. To differentiate QoS requirements, each waiting queue is assumed to be a multi-class non-preemptive priority queue such that users with higher QoS (lower latency) requirements are assigned with higher non-preemptive priorities. In other words, the delay-sensitive service-request-packets are always put in the head of the line (HOL) for service. To characterize the dynamic characteristics in the serving process, the service time of each server is also modeled to be stochastic. As MEC users consume both UT and DT transmission resources to offload the computation tasks to the nearest BS and receive the final computation results from the same BS, the admission control of MEC users becomes critically important. In addition, BS operators or service providers prefer not to deploy overmuch computation resources (i.e., edge servers) in order to avoid possible underutilization and high capital expenditure. Thus, deciding the amount of transmission and computation resource to be deployed at a BS to support all the dynamic multi-priority-level multi-type service is also a key issue for service providers to implement MEC.

The contributions of this paper in addressing the aforementioned issues can be summarized as follows.

- A queueing network model for a BS that offers MEC service, pure UT and DT service is investigated and the average delay of each class of each service type is analyzed, where users in a same user class are with the same priority level.
- Given the limited transmission and computation resources, the admission control of multi-class MEC service is formulated. Considering each class of MEC service has its own price, the problem aims to maximize the total revenue of service providers subject to the constraints of the average delay requirements of all admitted users. And the solution is also discussed.
- Given the service request distribution, the optimization of the resource deployment at a BS is also considered and an optimal resource deployment algorithm is developed. The algorithm can help service providers to determine the optimal amount of communication and computation resources to be placed at a BS to minimize the total capital expenditure subject to the required QoS from each type of service.

The rest of this paper is organized as follows. Section II introduces the proposed queueing network model. Based on the proposed model, the average delay of each class of each service type is analyzed in Section III. The optimization problems of admission control and resource deployment are investigated in Section IV and V, respectively. Section VI discusses the simulation results. At last, Section VII concludes the paper and points out the challenges and directions of future research.

II. QUEUEING NETWORK MODEL

A queueing network model in a BS simultaneously offering pure UT, pure DT and MEC service is illustrated in Fig. 1. Three individual user groups request for pure UT, pure DT and MEC service, respectively. The BS serving these user groups can be treated as multi-type servers including UT servers, DT servers and MEC computation servers. Inside the BS, the number of available UT servers, DT servers and computation servers are denoted by n_u, n_d and n_c, respectively.

A. MULTI-CLASS MULTI-TYPE SERVICES

Without loss of generality, we assume that there are multi-class users among each service type. As aforementioned, users in a same user class are with the same priority level. The numbers of user classes for pure UT, pure DT and MEC service are denoted as J_u, J_d and J_c, respectively. For any service type $x \in \{u, d, c\}$, class-i_x has higher priority over class-j_x if $1 \leq i_x < j_x \leq J_x$. Each class has its individual average delay QoS requirements. For class-j_x of service type x, the maximum tolerable average delay threshold is set to be $T_{x-th}^{(j_x)}$. Usually, $T_{x-th}^{(j_x)}$ increases as the priority goes lower, i.e., $T_{x-th}^{(1)} \leq T_{x-th}^{(2)} \leq \cdots \leq T_{x-th}^{(J_x)}$. For different service types (e.g., pure UT service and MEC service), the priorities of users can be assigned manually. For any class j_x users of

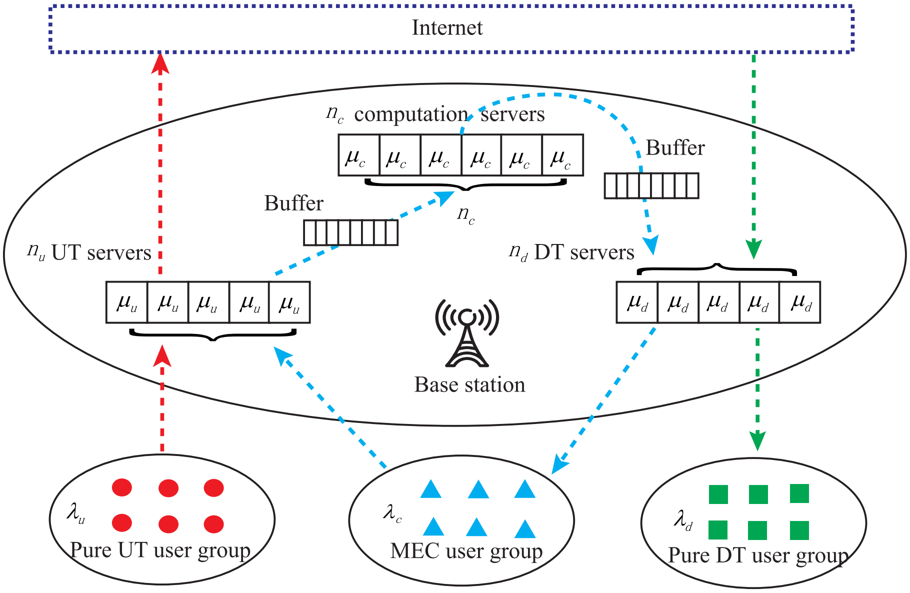


FIGURE 1. Base station model.

service type x, $j_{y \succ j_x}$, $x, y \in \{u, d, c\}$, $x \neq y$ represents the number of user classes of service type y that are assigned with higher priorities than class j_x users of service type x. For example, $j_{u \succ j_c} = 5$ represents there are 5 pure UT user classes with higher priorities than class-j_c MEC users. Based on the definition, one can easily obtain that the value $j_{y \succ j_x}$ ranges from 0 to J_y and that it is a constant for class-j_x users once the priorities among all classes of all service types are settled. The priority system has three different disciplines: preemptive-resume, preemptive-repeat, and non-preemptive (or HOL). Under a preemptive discipline, the arrival of high-priority user interrupts the serving of low-priority user. Though high-priority users are not affected by low-priority users under preemptive discipline, this discipline will not be applied in real transmission or MEC systems because it not only reduces the system efficiency but also damages the QoS of low-priority users greatly. Therefore, a non-preemptive discipline is applied, making the arrival of high-priority users only affect the waiting queue. The service-request-packets generated from class-j_x user group arrive according to a Poisson process with rate $\lambda_x^{(j_x)}$. The sum access rates of pure UT, pure DT and MEC service-request-packets are defined as λ_u, λ_d and λ_c, respectively, and have the following relationship with the arrival rate of each class as

$$\lambda_u = \sum_{j_u=1}^{J_u} \lambda_u^{(j_u)}, \tag{1}$$

$$\lambda_d = \sum_{j_c=1}^{J_d} \lambda_d^{(j_d)}, \tag{2}$$

and

$$\lambda_c = \sum_{j_d=1}^{J_c} \lambda_c^{(j)}. \tag{3}$$

Different types of users are served by different types of servers. As shown in Fig. 1, pure UT or DT users only requires service at one queue consisting of n_u UT servers or n_d DT servers, while MEC users requires service at three queues, the front queue consisting of n_u UT servers, the middle queue consisting of n_c computation servers and the last queue consisting of n_d DT servers. Without loss of generality, the following assumptions are made. Each request for transmission or computation service is only served by one transmission or computation server at a time. Each type of servers is homogeneous, consumes the same amount of transmission or computation resources and only serves a user at a time. The length of buffer is assumed to be infinite and the waiting time for users can be infinite.

The service time at a UT or DT server is related to the power of transmitter, the spectrum bandwidth, the channel condition, the size of the packet to be transmitted, the signal processing speed, etc. The distributions of the service time of all UT and DT servers with the same amount of resource are assumed to be identical and independent. Thus $M/G/n/\infty$ queue model would be suitable to model the UT and DT service queues. However, it is challenging to determine the exact distribution of the service time. For simplicity, the service time of UT and DT servers are assumed to follow exponential distributions with rate μ_u and rate μ_d, respectively. Note that the mean service rates for UT and DT are different because the parameters affecting the corresponding service time distribution are different. For example, the transmit power for UT provided by power-limited mobile devices is much smaller than that for DT offered by the powerful BS. The service time at a computation server is typically related to the amount of computation resource per server and the computation task specified in each packet. Similarly for simplicity, the service time at a computation server is assumed to follow an exponential distribution with rate μ_c.

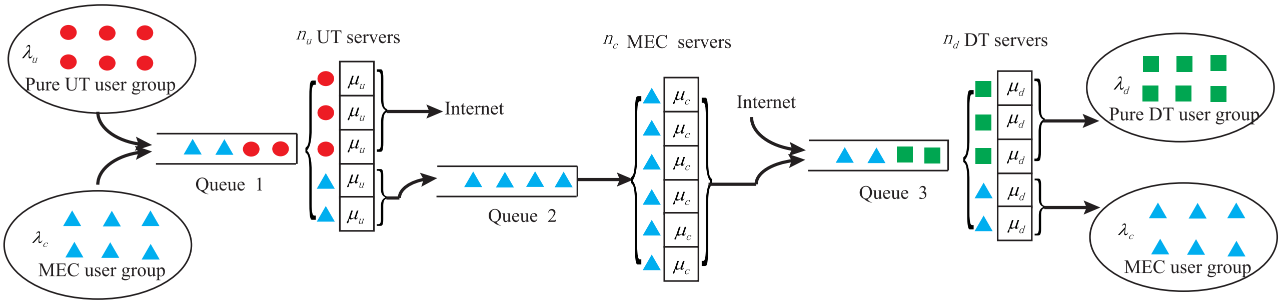


FIGURE 2. **Queuing network model for a base station providing MEC service.**

B. QUEUEING NETWORK MODELING AND STABILITY CONDITION

Based on the above assumptions, the BS providing MEC service is modeled as a queueing network as shown in Fig. 2.

- The UT service is modeled as an $M/M/n_u/\infty$ non-preemptive priority queue with $(J_u + J_c)$ priority classes shown as Queue 1 in Fig. 2. The utilization factor is defined as

$$\phi_u = \frac{\sum_{j_u=1}^{J_u} \lambda_u^{(j_u)} + \sum_{j_c=1}^{J_c} \lambda_c^{(j_c)}}{n_u \mu_u}. \tag{4}$$

- The DT service is modeled as an $M/M/n_d/\infty$ non-preemptive priority queue with $(J_d + J_c)$ priority classes shown as Queue 3 in Fig. 2. The utilization factor is defined as

$$\phi_d = \frac{\sum_{j_d=1}^{J_d} \lambda_d^{(j_d)} + \sum_{j_c=1}^{J_c} \beta_{j_c} \lambda_c^{(j_c)}}{n_d \mu_d}. \tag{5}$$

  where β_{j_c} represents the output-input ratio of class-j_c MEC service, and it depends on the service type.
- The MEC service is modeled as a sequence of queues. The front queue is the UT service queue. The middle queue is the computation queue, which can be modeled as an $M/M/n_c/\infty$ non-preemptive priority one with J_c priority classes shown as Queue 2 in Fig. 2. The utilization factor of the middle queue defined as

$$\phi_c = \frac{\sum_{j_c=1}^{J_c} \lambda_c^{(j_c)}}{n_c \mu_c} \tag{6}$$

  And the last queue is the DT service queue.

Proposition 1: Based on the queueing theory, the stability or ergodic conditions for the queueing network consisting of three queues are

$$\phi_u < 1, \tag{7}$$

$$\phi_d < 1, \tag{8}$$

and

$$\phi_c < 1. \tag{9}$$

The proof of the proposition can be found in [24].

III. AVERAGE DELAY ANALYSIS

The average delay performance of each user class of each service type needs to be taken into account in the admission control of MEC users and resource deployment. This section will derive the average delay of each user class of each service type. To that end, we make the following definitions based on Eqs. (4)-(6).

$$\Phi_u(j_u, j_c) \triangleq \frac{\sum_{j_u=1}^{j_u} \lambda_u^{(j_u)} + \sum_{j_c=1}^{j_c} \lambda_c^{(j_c)}}{n_u \mu_u}, \quad 0 \le j_u \le J_u, \quad 0 \le j_c \le J_c. \tag{10}$$

$$\Phi_d(j_d, j_c) \triangleq \frac{\sum_{j_d=1}^{j_d} \lambda_d^{(j_d)} + \sum_{j_c=1}^{j_c} \beta_{j_c} \lambda_c^{(j_c)}}{n_d \mu_d}, \quad 0 \le j_u \le J_u, \quad 0 \le j_c \le J_c. \tag{11}$$

$$\Phi_c(j_c) \triangleq \frac{\sum_{j_c=1}^{j_c} \lambda_c^{(j_c)}}{n_c \mu_c}, \quad 0 \le j_c \le J_c. \tag{12}$$

Thus, ϕ_u,ϕ_d and ϕ_c in Eqs. (4)-(6) can be rewritten as $\phi_u = \Phi_u(J_u, J_c)$, $\phi_d = \Phi_d(J_d, J_c)$ and $\phi_c = \Phi_c(J_c)$. Let us define $F(k, n, \phi)$, $G(n, \phi, \mu)$ and $R(n, \phi, \mu)$ as

$$F(k, n, \phi) \triangleq \frac{(n\phi)^k}{k!}, \tag{13}$$

$$G(n, \phi, \mu) \triangleq n\mu \left[(1 - \phi) \sum_{k=0}^{k=n-1} F(k, n, \phi) + F(n, n, \phi) \right], \tag{14}$$

$$R(n, \phi, \mu) \triangleq \frac{F(n, n, \phi)}{G(n, \phi, \mu)}. \tag{15}$$

$R(n, \phi, \mu)$ is also referred as the mean residential service time function [25]. Based on these definitions, one can obtain the average delay of each user class of each service type as follows.

Proposition 2: For pure UT service, the average delay including the waiting time and the service time of class-j_u $(1 \le j_u \le J_u)$ pure UT users is

$$T_u^{(j_u)} = \frac{1}{\mu_u} + \frac{R(n_u, \phi_u, \mu_u)}{\left[1 - \Phi_u(j_u - 1, j_{c \succ j_u})\right]\left[1 - \Phi_u(j_u, j_{c \succ j_u})\right]}. \tag{16}$$

Similarly, for pure DT service, the average delay including the waiting time and the service time of class-j_d $(1 \le j_d \le J_d)$

pure DT users is

$$T_d^{(j_d)} = \frac{1}{\mu_d} + \frac{R(n_d, \phi_d, \mu_d)}{\left[1-\Phi_d(j_d-1, j_{c\succ j_d})\right]\left[1-\Phi_d(j_d, j_{c\succ j_d})\right]} \tag{17}$$

The proof of Proposition 2 is given in appendix.

Proposition 3: For MEC service, the average delay including waiting time and the service time of class-j_c $(1 \le j_c \le J_c)$ MEC users contains three parts and can be expressed

$$T_c^{(j_c)} = T_{cu}^{(j_c)} + T_{cd}^{(j_c)} + T_{cc}^{(j_c)}. \tag{18}$$

$T_{cu}^{(j_c)}$ represents the average delay at the front queue consisting of n_u UT servers and can be given by

$$T_{cu}^{(j_c)} == \frac{1}{\mu_u} + \frac{R(n_u, \phi_u, \mu_u)}{\left[1-\Phi_u(j_{u\succ j_c}, j_c-1)\right]\left[1-\Phi_u(j_{u\succ j_c}, j_c)\right]}. \tag{19}$$

$T_{cd}^{(j_c)}$ represents the average delay at the last queue consisting of n_d DT servers and can be given by

$$T_{cd}^{(j_c)} == \frac{1}{\mu_d} + \frac{R(n_d, \phi_d, \mu_d)}{\left[1-\Phi_d(j_{d\succ j_c}, j_c-1)\right]\left[1-\Phi_d(j_{d\succ j_c}, j_c)\right]}. \tag{20}$$

$T_{cc}^{(j_c)}$ represents the average delay at the middle queue consisting of n_c computation servers and can be given by

$$T_{cc}^{(j_c)} = \frac{1}{\mu_c} + \frac{R(n_c, \phi_c, \mu_c)}{[1-\Phi_c(j_c-1)]\,[1-\Phi_c(j_c)]}. \tag{21}$$

Proposition 3 can be proved similarly as Proposition 2. As the MEC service is a tandem queueing network and the waiting room between queues is assumed to be infinite, the delay in the three queues can be treated as that in three individual and independent queues [26]. Therefore, by separately proving the correctness of the expressions in (19), (20) and (21) by using the method presented in appendix VII, Proposition 3 is proved.

IV. ADMISSION CONTROL

This section will discuss the admission control of multi-class MEC users given the limited transmission and computation resources (i.e., given the number of UT, DT and computation servers). An optimization problem will be formulated and analyzed.

A. PROBLEM FORMULATION

The known parameters include the priority order of each user class of each service type, the unit price of each class of MEC service $p_c^{(j_c)}$, $1 \le j_c \le J_c$, the arrival rates of multi-class pure UT and DT users $\lambda_u^{(j_u)}$, $1 \le j_u \le J_u$, $\lambda_d^{(j_d)}$, $1 \le j_d \le J_d$, the service rates μ_u, μ_d, μ_c, the number of available servers n_u, n_d, n_c, and the delay QoS thresholds $T_{u-th}^{(j_u)}$, $1 \le j_u \le J_u$, $T_{d-th}^{(j_d)}$, $1 \le j_d \le J_d$, $T_{c-th}^{(j_c)}$, $1 \le j_c \le J_c$.

The decision variables are the access rate of each class of MEC service $\lambda_c^{(j_c)}$, $1 \le j_c \le J_c$. The objective is to maximize the revenue of service providers for providing MEC service. The mean interest per unit time is treated as the objective function as

$$P(\lambda_c^{(1)}, \lambda_c^{(2)}, \cdots, \lambda_c^{(J_c)}) = \sum_{j_c=1}^{J_c} p_c^{(j_c)} \lambda_c^{(j_c)}. \tag{22}$$

The constraints are the average delay requirement of each user class of each service type specified by the maximum tolerable delay, i.e.,

$$T_u^{(j_u)} \le T_{u-th}^{(j_u)}, \quad 1 \le j_u \le J_u \tag{23}$$

$$T_d^{(j_d)} \le T_{d-th}^{(j_d)}, \quad 1 \le j_d \le J_d \tag{24}$$

$$T_c^{(j_c)} \le T_{c-th}^{(j_c)}, \quad 1 \le j_c \le J_c \tag{25}$$

and the stability conditions of the queueing network in (7), (8), and (9). Based on the above discussion, the optimization problem can be formulated as

$$\begin{aligned}
&\text{Find}: \lambda_c^{(j_c)}, \quad 1 \le j_c \le J_c \\
&\text{Maximize}: P(\lambda_c^{(1)}, \lambda_c^{(2)}, \cdots, \lambda_c^{(J_c)}) \\
&\text{subject to}: \lambda_c^{(j_c)} \ge 0, \quad 1 \le j_c \le J_c \\
&\qquad\qquad (7),\ (8),\ (9) \\
&\qquad\qquad (23),\ (24),\ (25)
\end{aligned} \tag{26}$$

In this problem, the object function, the nonnegative constraints and the stationary conditions are all linear. But the delay constraint of each user class of each service type is no longer linear. Through the following analysis, the delay constraints are found to be polynomials. Before introducing the reformulation of constraints, we denote a multivariate polynomial by

$$\begin{aligned}
Q(\overrightarrow{\lambda_c}) &\triangleq \sum_{\overrightarrow{\alpha} \in \mathbb{Z}^n} q_{\overrightarrow{\alpha}} \left[\lambda_c^{(1)}\right]^{\alpha_1} \left[\lambda_c^{(2)}\right]^{\alpha_2} \cdots \left[\lambda_c^{(J_c)}\right]^{\alpha_{J_c}} \\
&= \sum_{\overrightarrow{\alpha} \in \mathbb{Z}^n} q_{\overrightarrow{\alpha}} \left[\overrightarrow{\lambda_c}\right]^{\overrightarrow{\alpha}},
\end{aligned} \tag{27}$$

where $q_{\overrightarrow{\alpha}}$ represents the coefficient and the monomial

$$\left[\overrightarrow{\lambda_c}\right]^{\overrightarrow{\alpha}} \triangleq \left[\lambda_c^{(1)}\right]^{\alpha_1} \left[\lambda_c^{(2)}\right]^{\alpha_2} \cdots \left[\lambda_c^{(J_c)}\right]^{\alpha_{J_c}}. \tag{28}$$

The degree of the monomial $\left[\overrightarrow{\lambda_c}\right]^{\overrightarrow{\alpha}}$ is defined as

$$|\overrightarrow{\alpha}| = \sum_{j_c=1}^{J_c} \alpha_{j_c}, \tag{29}$$

and the degree of the polynomial is the maximum degree of a monomial $\left[\overrightarrow{\lambda_c}\right]^{\overrightarrow{\alpha}}$ for which $q_{\overrightarrow{\alpha}} \ne 0$.

B. REFORMULATION OF CONSTRAINTS

For any j_u, $1 \le j_u \le J_u$, based on Proposition 2, the constraint in (23) can be written as

$$\frac{R(n_u, \phi_u, \mu_u)}{\left[1-\Phi_u(j_u-1, j_{c\succ j_u})\right]\left[1-\Phi_u(j_u, j_{c\succ j_u})\right]} \le c_u^{(j_u)}, \tag{30}$$

where $c_u^{(j_u)} \triangleq T_u^{(j_u)} - \frac{1}{\mu_u}$ is a constant. Substituting the definition of function $R(n, \phi, \mu)$ in (15) into (30), we have

$$\frac{F(n_u, n_u, \phi_u)}{G(n_u, \phi_u, \mu_u)\left[1 - \Phi_u(j_u - 1, j_{c \succ j_u})\right]\left[1 - \Phi_u(j_u, j_{c \succ j_u})\right]} \leq c_u^{(j_u)}. \quad (31)$$

It is certain that the denominator is positive if the queueing network is stationary. Moving the denominator to the other side in (31) and swapping the inequalities, we can get

$$c_u^{(j_u)} G(n_u, \phi_u, \mu_u)\left[1 - \Phi_u(j_u - 1, j_{c \succ j_u})\right]\left[1 - \Phi_u(j_u, j_{c \succ j_u})\right] - F(n_u, n_u, \phi_u) \geq 0 \quad (32)$$

On the left side of the constraint reformulation in (32), all parts are either constants or polynomials with respect to the decision variables $\lambda_c^{(j_c)}$, $1 \leq j_c \leq J_c$. Specially, $c_u^{(j_u)}$ is a constant. $G(n_u, \phi_u, \mu_u)$ is a polynomial of ϕ_u with degree n_u according to (14) and ϕ_u is a linear combination of all decision variables (c.f. (4)). Therefore, $G(n_u, \phi_u, \mu_u)$ is a multivariate polynomial of decision variables with degree n_u. Based on the definition of $\Phi_u(J_u, J_c)$ given in (10), both $\Phi_u(j_u - 1, j_{c \succ j_u})$ and $\Phi_u(j_u, j_{c \succ j_u})$ are either constants if $j_{c \succ j_u} = 0$, or polynomials with degree 1 if $0 < j_{c \succ j_u} \leq J_c$. Based on the definition of $F(k, n, \phi)$ given in (13), $F(n_\mu, n_\mu, \phi_\mu)$ is a polynomial of ϕ_u with degree n_u and as mentioned ϕ_u is a linear combination of all decision variables. Therefore, $F(n_\mu, n_\mu, \phi_\mu)$ is also a multivariate polynomial with degree n_u.

With all the parts of the left side of (32), the delay constraint shown in (23) for any j_c, $1 \leq j_c \leq J_c$ can be reformulated to a multivariate polynomial either with degree n_u if $j_{c \succ j_u} = 0$ or with degree $(n_u + 2)$ if $0 < j_{c \succ j_u} \leq J_c$.

Similarly, for any j_d, $0 \leq j_d \leq J_d$, the delay constraint for class-j_d pure DT users (24) can be reformulated to a multivariate polynomial as

$$c_d^{(j_d)} G(n_d, \phi_d, \mu_d)\left[1 - \Phi_d(j_d - 1, j_{c \succ d})\right]\left[1 - \Phi_d(j_d, j_{c \succ d})\right] - F(n_d, n_d, \phi_d) \geq 0. \quad (33)$$

either with degree n_d if $j_{c \succ j_d} = 0$ or with degree $(n_d + 2)$ if $0 < j_{c \succ j_d} \leq J_c$, where constant $c_d^{(j_d)} \triangleq T_d^{(j_d)} - \frac{1}{\mu_d}$.

The reformulation of the constraint in (25) is a bit more complicated as it consists of three parts according to Proposition 3. For any j_c, $1 \leq j_c \leq J_c$, the delay constraint of class-j_c MEC users can be expressed as

$$\frac{R(n_u, \phi_u, \mu_u)}{\left[1 - \Phi_u(j_{u \succ j_c}, j_c - 1)\right]\left[1 - \Phi_u(j_{u \succ j_c}, j_c)\right]} + \frac{R(n_d, \phi_d, \mu_d)}{\left[1 - \Phi_d(j_{d \succ j_c}, j_c - 1)\right]\left[1 - \Phi_d(j_{d \succ j_c}, j_c)\right]} + \frac{R(n_c, \phi_c, \mu_c)}{[1 - \Phi_c(j_c - 1)][1 - \Phi_c(j_c)]} \leq c_c^{(j_c)}. \quad (34)$$

where constant $c_c^{(j_c)} \triangleq T_c^{(j_c)} - \frac{1}{\mu_u} - \frac{1}{\mu_d} - \frac{1}{\mu_c}$. Substituting the definition of function $R(n, \phi, \mu)$ in (15) into (34), one can obtain

$$\frac{F(n_u, n_u, \phi_u)}{G(n_u, \phi_u, \mu_u)\left[1 - \Phi_u(j_{u \succ j_c}, j_c - 1)\right]\left[1 - \Phi_u(j_{u \succ j_c}, j_c)\right]} + \frac{F(n_d, n_d, \phi_d)}{G(n_d, \phi_d, \mu_d)\left[1 - \Phi_d(j_{d \succ j_c}, j_c - 1)\right]\left[1 - \Phi_d(j_{d \succ j_c}, j_c)\right]} + \frac{F(n_c, n_c, \phi_c)}{G(n_c, \phi_c, \mu_c)[1 - \Phi_c(j_c - 1)][1 - \Phi_c(j_c)]} \leq c_c^{(j_c)}. \quad (35)$$

To simplify the constraint in (35), the numerators are defined as H_u, H_d and H_c functions as

$$H_u = G(n_u, \phi_u, \mu_u)\left[1 - \Phi_u(j_{u \succ j_c}, j_c - 1)\right] \times \left[1 - \Phi_u(j_{u \succ j_c}, j_c)\right], \quad (36)$$

$$H_d = G(n_d, \phi_d, \mu_d)\left[1 - \Phi_d(j_{d \succ j_c}, j_c - 1)\right] \times \left[1 - \Phi_d(j_{d \succ j_c}, j_c)\right], \quad (37)$$

and

$$H_c = G(n_c, \phi_c, \mu_c)[1 - \Phi_c(j_c - 1)][1 - \Phi_c(j_c)]. \quad (38)$$

It is easy to verify that H_u, H_d and H_c are all positive if the queueing network is stationary. Multiplying $H_u H_d H_c$ at the both sides of (35) and (35) can be written as

$$c_c^{(j_c)} H_u H_d H_c - H_d H_c F(n_u, n_u, \phi_u) - H_u H_c F(n_d, n_d, \phi_d) - H_u H_d F(n_c, n_c, \phi_c) \geq 0. \quad (39)$$

On the left side of the constraint reformulation in (39), all parts are also either constants or polynomials with respect to the decision variables $\lambda_c^{(j_c)}$, $1 \leq j_c \leq J_c$. In detail, $c_c^{(j_c)}$ is a constant. For H_u, H_d and H_c, there are two cases. Specially, for $j_c = 1$, $\Phi_u(j_{u \succ j_c}, j_c - 1)$, $\Phi_d(j_{d \succ j_c}, j_c - 1)$ and $\Phi_c(j_c - 1)$ are all constants, therefore based on the similar analysis as Section IV-B, H_u, H_d and H_c are polynomials with degrees $(n_u + 1)$, $(n_d + 1)$ and $(n_c + 1)$, respectively. Otherwise if $1 < j_c \leq J_c$, $\Phi_u(j_{u \succ j_c}, j_c - 1)$, $\Phi_d(j_{d \succ j_c}, j_c - 1)$ and $\Phi_c(j_c - 1)$ are all polynomials with degree 1, therefore based on the similar analysis as Section IV-B, H_u, H_d and H_c are polynomials with degrees $(n_u + 2)$, $(n_d + 2)$ and $(n_c + 2)$, respectively.

With all the parts of the left side of (39), the delay constraint shown in (25) for any j_c, $1 \leq j_c \leq J_c$ can be modified into a multivariate polynomial either with degree $(n_u + n_d + n_c + 3)$ if $j_c = 1$ or with degree $(n_u + n_d + n_c + 6)$ if $1 < j_c \leq J_c$.

C. GENERAL SOLUTION AND COMPLEXITY ANALYSIS

As mentioned, all the object function, the nonnegative constraints and the stationary conditions are linear. Since the linear equalities or inequalities can also be considered as polynomials with degree 1 and the delay constraints in (23), (24) and (25) are all polynomials according to the analysis in Section IV-B, the optimization problem in (26) is a general polynomial optimization problem. The problem is NP-hard, i.e., intractable. By using the moment-based convex linear matrix inequality (LMI) relaxations, the approximate solutions can be found [27], [28]. This approach for solving global optimization problems over polynomials has been embedded in the GloptiPoly solver [29]. Thus, the approximate solution can be obtained by applying GloptiPoly solver. According to [30], the complexity of the approach in terms of the number of LMI decision variables M_L and size of LMI N_L can be expressed as

$$M_L = \binom{n_v + 2\delta}{\delta} - 1, \tag{40}$$

and

$$N_L = \binom{n_v + \delta}{\delta} + m_c \binom{n_v + \delta - 1}{\delta - 1}. \tag{41}$$

where n_v is the number of polynomial variables, m_c denotes the number of constraints, $\delta = \lfloor (d+1)/2 \rfloor$ and d represents the overall polynomial degree. From (40) and (41), it is observed that M_L and N_L grow polynomially in $O(\delta^n)$ and in $O(m_c\delta^n)$, respectively. For the optimization problem in (26), $n_v = n_c$, $m_c = 3 + J_u + J_d + 2J_c$ and $d = n_u + n_d + n_c + 6$. Substituting these into (40) and (41), the complexity of solving this optimization problem can be derived.

V. OPTIMAL RESOURCE DEPLOYMENT

In this section, an optimal resource deployment strategy to provision the QoS and meanwhile to minimize the capital expenditure will be developed. Another optimization problem will be formulated and solved.

The known parameters are the priority order of each class of each service type, the cost of each server type $C_x, x \in \{u, c, d\}$, the access rates of multi-class pure UT, pure DT and MEC users $\lambda_u^{(j_u)}$, $1 \le j_u \le J_u$, $\lambda_d^{(j_d)}$, $1 \le j_d \le J_d$, $\lambda_c^{(j_c)}$, $1 \le j_c \le J_c$, the service rates μ_u, μ_d, μ_c, and the delay QoS thresholds $T_{u-th}^{(j_u)}$, $1 \le j_u \le J_u$, $T_{d-th}^{(j_d)}$, $1 \le j_d \le J_d$, $T_{c-th}^{(j_c)}$, $1 \le j_c \le J_c$.

The decision variables are the number of servers to be deployed, i.e., n_u, n_d, and n_c. The objective is to minimize the capital expenditure of servers. The total capital expenditure of deployed servers is treated as the cost function as

$$C(n_u, n_c, n_d) = n_u C_u + n_d C_d + n_c C_c. \tag{42}$$

The problem of minimizing the cost function can be given by

$$\begin{aligned} &\text{Find} : n_u,\ n_c,\ n_d \\ &\text{Minimize} : C(n_u, n_c, n_d) \\ &\text{subject to} : n_u,\ n_c,\ n_d \in \mathbb{Z}_+ \\ &\qquad\qquad (7),\ (8),\ (9) \\ &\qquad\qquad (23),\ (24),\ (25) \end{aligned} \tag{43}$$

A. OPTIMAL SOLUTION

The optimization problem in (43) is an integer programing problem. The queueing network stability conditions (7), (8) and (9) can be reformulated as

$$n_u > \frac{\sum_{j_u=1}^{J_u} \lambda_u^{(j_u)} + \sum_{j_c=1}^{J_c} \lambda_c^{(j_c)}}{\mu_u} \triangleq n_u^{\min}, \tag{44}$$

$$n_d > \frac{\sum_{j_d=1}^{J_d} \lambda_d^{(j_d)} + \sum_{j_c=1}^{J_c} \beta_{j_c} \lambda_c^{(j_c)}}{\mu_d} \triangleq n_d^{\min}, \tag{45}$$

and

$$n_c > \frac{\sum_{j_c=1}^{J_c} \lambda_c^{(j_c)}}{\mu_c} \triangleq n_c^{\min}. \tag{46}$$

The delay constraints in (23), (24) and (25) are polynomials with respect to $\lambda_c^{j_c}$, $1 \le j_c \le J_c$. However, in the resource deployment problem, the decision variables are n_u n_d and n_c. As shown in Propositions 2 and 3, the average delay of each class of each service type (the left sides of inequalities (23), (24) and (25)) is comprised of very complicated functions of decision variables n_u n_d and n_c, however, the following corollary can be obtained through analysis.

Corollary 1:

- The average delay of class-j_u, $1 \le j_u \le J_u$ pure UT users is only related to decision variable n_u and decreases along with the increase of n_u.
- The average delay of class-j_d, $1 \le j_d \le J_d$ pure DT users is only related to decision variable n_d and decreases along with the increase of n_d.
- The average delay of class-j_c, $1 \le j_c \le J_c$ MEC users is related to all decision variable n_u, n_d and n_c and decreases along with any increase of n_u, n_d or n_c.

Proof: Corollary 1 can be easily proved. From Proposition 2 and 3, the parameters that affect the average delay can be directly observed. And for the monotonicity, by thinking of each station as being comprised of a serving section and a waiting section, it is straightforward to conclude that increasing the number of servers definitely increases the service throughput and consequently decreases the waiting time. ■

Based on Corollary 1, the optimal solution of the optimization problem in (43) can be found by using the following steps. The first step is to find the minimum feasible number of UT servers $n_u^{\text{min feasible}}$ and that of DT servers $n_d^{\text{min feasible}}$ based on the delay constraints of pure UT and DT users in (23) and (24), respectively. Note that the minimum feasible number of computation servers is related to the available number of UT and DT servers n_u, n_c, so we let $n_c^{\text{min feasible}}(n_u, n_d)$ represent the minimum feasible number of computation servers given n_u and n_d.

Based on $n_u^{\text{min feasible}}$ and $n_d^{\text{min feasible}}$, the second step is to find the minimum feasible number of computation servers denoted by $N_c \triangleq n_c^{\text{min feasible}}\left(n_u^{\text{min feasible}}, n_d^{\text{min feasible}}\right)$. Then we can calculate the corresponding cost $C^{\text{benchmark}}$ based on (42) as

$$C^{\text{benchmark}} = n_u^{\text{min feasible}} C_u + n_d^{\text{min feasible}} C_d + N_c C_c \tag{47}$$

Based on this point $\left(n_u^{\text{min feasible}}, n_d^{\text{min feasible}}\right)$ we can search the area of (n_u, n_d) that is possible to have the sum cost less than $C^{\text{benchmark}}$. The search area can be given by

$$\mathbb{A} = \left\{ (n_u, n_d) \in \mathbb{Z}^+ \;\middle|\; \begin{array}{l} n_u \geq n_u^{\text{min feasible}} \\ n_d \geq n_d^{\text{min feasible}} \\ n_u C_u + n_d C_d \leq \\ C^{\text{benchmark}} - n_c^{\text{min feasible}}(n_u, n_d)\, C_c \end{array} \right\}. \tag{48}$$

Based on the reformulation of the stability condition in (46), it can be derived that $n_c^{\text{min feasible}}(n_u, n_d)$ should be greater than $n_c^{\min}$ and that the search area can be relaxed to a triangle as

$$\mathbb{A}^+ = \left\{ (n_u, n_d) \in \mathbb{Z}^+ \;\middle|\; \begin{array}{l} n_u \geq n_u^{\text{min feasible}} \\ n_d \geq n_d^{\text{min feasible}} \\ n_u C_u + n_d C_d \\ \quad < C^{\text{benchmark}} - n_c^{\min} C_c \end{array} \right\}. \tag{49}$$

We illustrate the triangle in Fig. 3 with three vertices being labeled as $\left(n_u^{\text{min feasible}}, n_d^{\text{min feasible}}\right)$, $\left(n_u^{\text{min feasible}}, N_d\right)$ and $\left(N_u, n_d^{\text{min feasible}}\right)$, where N_u and N_d are defined as $N_u \triangleq \frac{C^{\text{benchmark}} - n_c^{\min} C_c - n_d^{\text{min feasible}} C_d}{C_u}$ and $N_d \triangleq \frac{C^{\text{benchmark}} - n_c^{\min} C_c - n_u^{\text{min feasible}} C_u}{C_d}$.

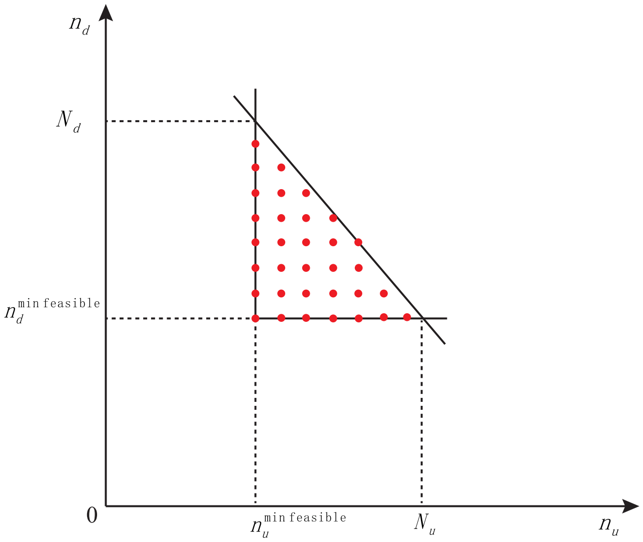


FIGURE 3. **The simplified search area for (n_u, n_d).**

The third step is to find the optimal solution in this triangle by exhaustively searching for all integer points in this area. For each point (n_u, n_c), $n_c^{\text{min feasible}}(n_u, n_d)$ can be found, the cost can be calculated, and the optimal solution can be obtained. To summarize the solution procedure, an optimal resource deployment algorithm as illustrated in Algorithm 1 is developed.

B. COMPLEXITY ANALYSIS

Algorithm 1 has three steps. For step 1, it requires $\left(n_u^{\text{min feasible}} - n_u^{\min}\right)$ calculations of all the J_u constraints

Algorithm 1 Optimal Resource Deployment Algorithm

```
Require: The priority order of each user class of each service
    type;
    The cost of each server type $C_x, x \in \{u, c, d\}$;
    The access rates $\lambda_u^{(j_u)},\ 1 \leq j_u \leq J_u$, $\lambda_d^{(j_d)},\ 1 \leq j_d \leq J_d$,
    $\lambda_c^{(j_c)},\ 1 \leq j_c \leq J_c$;
    The service rates $\mu_u, \mu_d, \mu_c$;
    The delay QoS thresholds $T_{u-th}^{(j_u)},\ 1 \leq j_u \leq J_u$, $T_{d-th}^{(j_d)},\ 1 \leq$
    $j_d \leq J_d$, $T_{c-th}^{(j_c)},\ 1 \leq j_c \leq J_c$.
Ensure: $n_u^{\text{opt}}, n_d^{\text{opt}}, n_c^{\text{opt}}$ and $C^{\min}$
    $n_u^{\min} = \frac{\sum_{j_u=1}^{J_u} \lambda_u^{(j_u)} + \sum_{j_c=1}^{J_c} \lambda_c^{(j_c)}}{\mu_u}$;
    $n_d^{\min} = \frac{\sum_{j_d=1}^{J_d} \lambda_d^{(j_d)} + \sum_{j_c=1}^{J_c} \beta_{j_c} \lambda_c^{(j_c)}}{\mu_d}$;
    $n_c^{\min} = \frac{\sum_{j_c=1}^{J_c} \lambda_c^{(j_c)}}{\mu_c}$;
    $n_u = n_u^{\min}$; $n_d = n_d^{\min}$; $n_c = n_c^{\min}$;
    Step 1: Find $n_u^{\text{min feasible}}$ and $n_d^{\text{min feasible}}$
    repeat
        $n_u = n_u + 1$;
    until all the constraints in (23) are satisfied.
    $n_u^{\text{min feasible}} = n_u$;
    repeat
        $n_d = n_d + 1$;
    until all the constraints in (24) are satisfied.
    $n_d^{\text{min feasible}} = n_d$;
    Step 2: Find $N_c \triangleq n_c^{\text{min feasible}}\left(n_u^{\text{min feasible}}, n_d^{\text{min feasible}}\right)$
    repeat
        $n_c = n_c + 1$;
    until all the constraints in (25) are satisfied.
    $N_c = n_c$;
    $C^{\text{benchmark}} = n_u^{\text{min feasible}} C_u + n_d^{\text{min feasible}} C_d + N_c C_c$;
    $N_u = \frac{C^{\text{benchmark}} - n_c^{\min} C_c - n_d^{\text{min feasible}} C_d}{C_u}$;
    $N_d = \frac{C^{\text{benchmark}} - n_c^{\min} C_c - n_u^{\text{min feasible}} C_u}{C_d}$;
    Step 3: Exhaustively search $(n_u, n_d)$ in $\mathbb{A}^+$ to find the
    optimal solution
    for $n_u = n_u^{\text{min feasible}} : 1 : N_u$ do
        for $n_d = n_d^{\text{min feasible}} : 1 : N_d$ do
            $n_c = n_c^{\min}$;
            repeat
                $n_c = n_c + 1$;
            until all the constraints in (25) are satisfied.
            $n_c^{\text{min feasible}}(n_u, n_d) = n_c$;
            $C = n_u C_u + n_d C_d + n_c^{\text{min feasible}}(n_u, n_d)\, C_c$;
            if $C < C^{\text{benchmark}}$ then
                $n_u^{\text{opt}} = n_u$;
                $n_d^{\text{opt}} = n_d$;
                $n_c^{\text{opt}} = n_c$;
                $C^{\text{opt}} = C$;
                $C^{\text{benchmark}} = C$;
            end if
        end for
    end for
```

in (23) and $\left(n_d^{\text{min feasible}} - n_d^{\min}\right)$ calculations of all the J_d constraints in (24). For step 2, $\left(N_c - n_c^{\min}\right)$ calculations of all the J_c constraints in (25) are required. For step 3, the complexity is determined by the number of points (n_u, n_d) inside the triangle and for each point the complexity lies in finding $n_c^{\text{min feasible}}(n_u, n_d)$. The number of points can be approximated by the area of the triangle which can be written as

$$|\mathbb{A}^+| = \frac{(N_u - n_u^{\text{min feasible}})(N_d - n_d^{\text{min feasible}})}{2}. \qquad (50)$$

and the complexity of finding $n_c^{\text{min feasible}}(n_u, n_d)$ for each point (n_u, n_d) is $\left(n_c^{\text{min feasible}}(n_u, n_d) - n_c^{\min}\right)$ calculations of all the J_c constraints in (25). Since $n_c^{\text{min feasible}}(n_u, n_d) < N_c$, the complexity is less than $\left(N_c - n_c^{\min}\right)$ and the total complexity of step 3 is less than $\left(\frac{(N_u - n_u^{\text{min feasible}})(N_d - n_d^{\text{min feasible}})(N_c - n_c^{\min})}{2}\right)$ calculations of all the J_c constraints in (25). Adding the complexities of three steps together, the total complexity of the algorithm can be obtained.

VI. SIMULATIONS

This section is provided to validate the theoretical analysis in Section III through a toy example as well as to show the effectiveness of the admission control scheme investigated in Section IV and the resource deployment strategy proposed in Section V. For comparison, the scheme without considering the priorities among user classes is also included as a benchmark. The simulation parameters with and without considering the priorities among users are given in table 1 and 2, respectively.

TABLE 1. Simulation parameters with considering priorities.

Parameters	Values
The number of pure UT user classes J_u	2
The number of pure DT user classes J_d	2
The number of MEC user classes J_c	2
Access rates of UT users $\{\lambda_u^{(1)}, \lambda_u^{(2)}\}$	$\{1,\ 2\}$/unit time
Access rates of DT users $\{\lambda_d^{(1)}, \lambda_d^{(2)}\}$	$\{3,\ 4\}$/unit time
Access rates of MEC users $\{\lambda_c^{(1)}, \lambda_c^{(2)}\}$	$\{3,\ 3\}$/unit time
Output-input ratio of MEC $\{\beta_1, \beta_2\}$	$\{1,\ 1\}$
Thresholds of UT $\{T_{u-th}^{(1)}, T_{u-th}^{(2)}\}$	$\{0.37,\ 0.39\}$ unit time
Thresholds of DT $\{T_{d-th}^{(1)}, T_{d-th}^{(2)}\}$	$\{0.27,\ 0.28\}$ unit time
Thresholds of MEC $\{T_{c-th}^{(1)}, T_{c-th}^{(2)}\}$	$\{1.14,\ 1.22\}$ unit time
UT service queue priorities	$j_u^{(1)} \succ j_c^{(1)} \succ j_u^{(2)} \succ j_c^{(2)}$
DT service queue priorities	$j_d^{(1)} \succ j_c^{(1)} \succ j_c^{(2)} \succ j_d^{(2)}$
MEC service queue priorities	$j_c^{(1)} \succ j_c^{(2)}$
Prices of an MEC service $\{p_c^{(1)}, p_c^{(2)}\}$	$\{0.02, 0.018\}$
The number of UT servers n_u	5
The number of DT servers n_d	6
The number of computation servers n_c	7
Service rate of UT servers μ_u	3/unit time
Service rate of DT servers μ_d	4/unit time
Service rate of computation servers μ_c	2/unit time
The cost per UT server C_u	15
The cost per DT server C_d	10
The cost per computation server C_c	40
Observation time	3000 unit time

TABLE 2. Simulation parameters without considering priorities.

Access rate of pure UT users λ_u	3/unit time
Access rate of pure DT users λ_d	7/unit time
Access rate of MEC users λ_c	6/unit time
Output-input ratio β	1
Delay threshold of pure UT T_{u-th}	0.37 unit time
Delay threshold of pure DT T_{d-th}	0.27 unit time
Delay threshold of MEC T_{c-th}	1.14 unit time
Prices of an MEC service request p_c	0.02
The number of UT servers n_u	5
The number of DT servers n_d	6
The number of computation servers n_c	7
Service rate of UT servers μ_u	3/unit time
Service rate of DT servers μ_d	4/unit time
Service rate of computation servers μ_c	2/unit time
The cost per UT server C_u	15
The cost per DT server C_d	10
The cost per computation server C_c	40
Observation time	3000 unit time

A. THE IMPACT OF THE ADMISSION OF MEC USERS ON PURE UT & DT SERVICE

Firstly, Monte Carlo simulations and numerical calculations are done by using the system setup in table 1 except that the arrival rate of the class-1 MEC users $\lambda_c^{(1)}$ is set ranging from 1 to 5. Along with the variation of $\lambda_c^{(1)}$, it can be calculated that the utilization factor of UT servers ϕ_u ranges from 0.4667 to 0.7333, and that of DT servers from 0.4583 to 0.6250. Both UT and DT service queues are always stable. Figs. 4 and 5 demonstrate the average delay of pure UT service and DT service, respectively. Both figures show that the delay of each user class of UT & DT service increases with the increased admission of MEC users. From Fig. 4, it is observed that the average delay of class-1 and class-2 pure UT users increases by 6.915% and 26.817%, respectively. From Fig. 5, it is observed that the average delay of class-1 and class-2 pure DT users increases by 2.9553% and 15.1491%, respectively. All these results indicate that with higher priority, less UT/DT service is affected. During the observation time, the simulation results of the average delay matches very well with

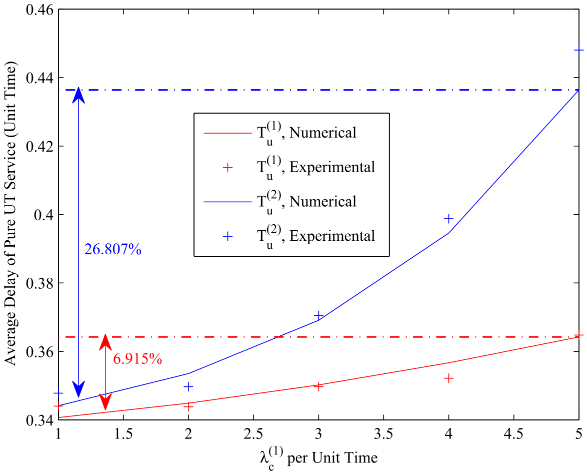


FIGURE 4. The impact of the access of MEC users on pure UT service.

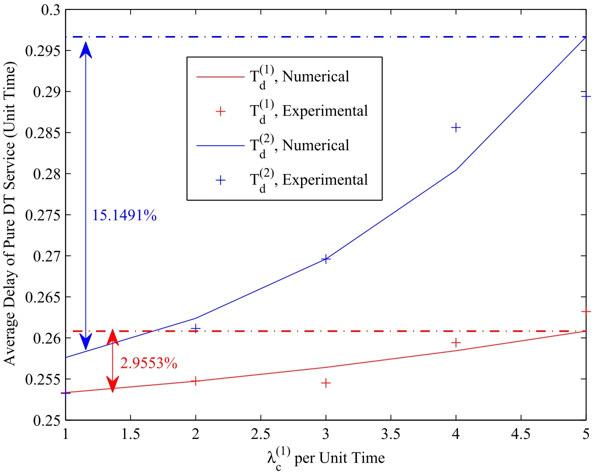


FIGURE 5. The impact of the access of MEC users on pure DT service.

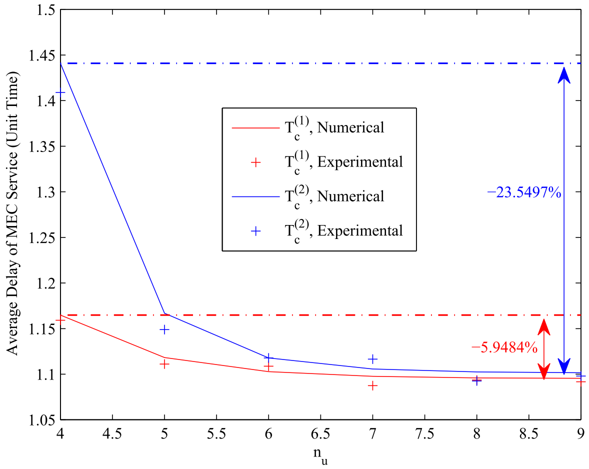


FIGURE 7. The impact of the number of UT servers on MEC service.

the numerical results. This validates the correctness of our theoretical analysis on average delay shown in Proposition 2. Above all, all these results demonstrate the significance of the admission control of MEC users.

B. THE IMPACT OF THE RESOURCE DEPLOYMENT

Secondly, Monte Carlo simulations and the numerical calculations are done by using the system setup in Tab. 1 except that the number of UT servers n_u is set to increase from 4 to 9. Along with the increment of n_u, it can be calculated that the utilization factor of UT servers ϕ_u decreases from 0.75 to 0.3333. The UT queue is always stable along the variation of n_u. Figs 6 and 7 demonstrate the average delay of pure UT service and MEC service, respectively. Both figures show that the delay of each user class of UT and MEC service decreases along with the increase of the number of UT servers. Fig. 4 demonstrates that the average delay of class-1 and class-2 pure UT users decreases by 12.1578% and 27.5959%, respectively. Fig. 5 demonstrates that the average delay of class-1 and class-2 MEC users decreases by 5.9484% and 23.5497%, respectively. These results indicate that with higher priority, less UT/MEC service is affected by the increment of resource. The reason is that the available resource is mainly provided to serve the higher-priority users. Therefore, increasing the resource is more helpful to the lower-priority service. Similarly, during the observation time of 3000, the simulation results match very well with the numerical results validating the analysis in Propositions 2 and 3. All these results imply the importance of the proposed resource deployment optimization method.

C. ADMISSION CONTROL OPTIMIZATION

Thirdly, based on the admission control optimization problem formulated in Section IV, Fig. 8 shows the feasible zone using the shadow area and demonstrates the injection point between the linear objective function and the feasible zone. From the illustrated constraints, it is observed that most of them are linear or approximately linear. This is because the

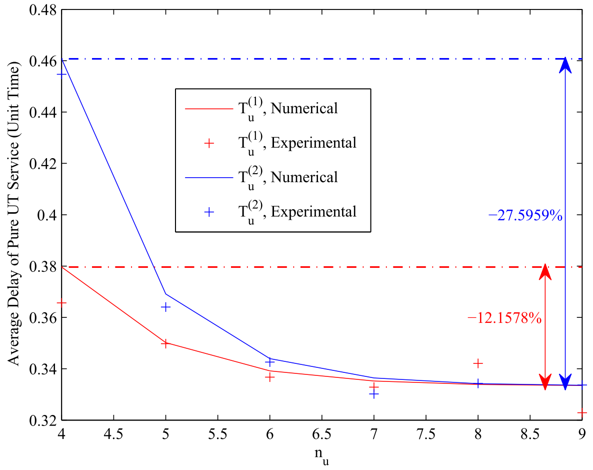


FIGURE 6. The impact of the number of UT servers on pure UT service.

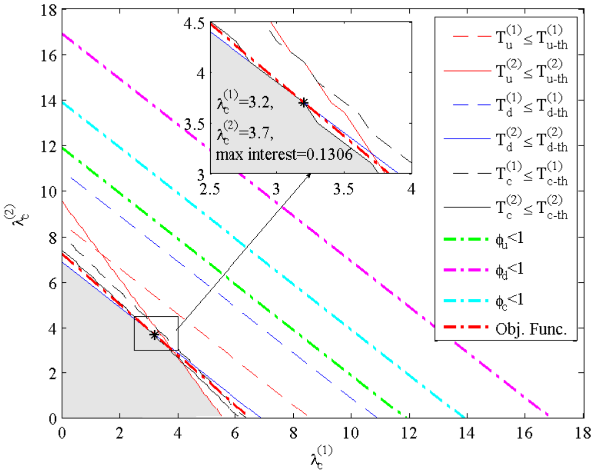


FIGURE 8. The feasible zone of the admission control optimization.

partial constraints in (23), (24) and (25) are only related to the linear combinations (e.g., ϕ_a, ϕ_d and ϕ_c) of all decision variables, and the delay constraints are obviously monotone decreasing functions of these linear combinations. Thus, these polynomial constraints can be replaced by linear ones. For those approximately linear constraints, they are not only related to the linear combinations of all decision variables but also affected by the partial combinations or single variables. Thus, they are not strictly linear, which is indicated in the zoomed figure in Fig. 8. Using either the GloptiPoly solver [29] or drawing method, the injection point between the linear objective function and the feasible zone is found to be (3.2, 3.7) and the corresponding maximum interest is calculated to be 0.1306 per unit time.

For comparison, Fig. 9 demonstrates the average delay variation along with the increase of access rate of MEC users without considering the priorities among users. For fair comparison, the simulation parameters set in Tab. 2 are same with that considering priorities in Tab. 1 except that the access rate of each service type is the sum of the access rates of all priority classes using this service type. The delay constraint of each service type is set as same as the high-priority users using this service type and so is the price. Under such simulation setup, it is seen from Fig. 9 that the average delay increases with more access rate of MEC users. It demonstrates that the access rate of MEC users should be less than 5.8 to satisfy the quality requirement of all service. Multiplying the maximum access rate with the highest price, the maximum interest is calculated to be 0.116 per unit time, 12.59% lower than that considering the priorities among users.

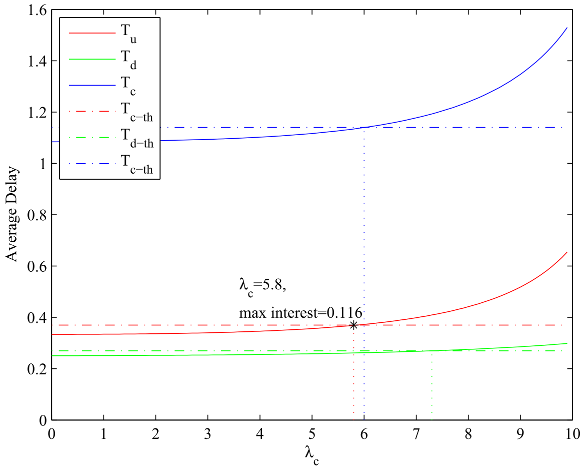


FIGURE 9. Average delay variation along with the increase of access rate of MEC users without considering the priorities among users.

D. RESOURCE DEPLOYMENT OPTIMIZATION

At last, the optimal resource deployment of the given system setup is analyzed by using the proposed optimal resource deployment algorithm in Algorithm 1. The results are illustrated in Fig.10. Note that the cost of a computation server is set to be higher than that of a UT or DT server as shown in Tables 1 and 2. The reason is that there already exist the transmission resources at a BS and upgrading these transmission servers is supposed to be more cheaper than adding new computation servers along with other accessory equipment. Under this consideration, the results in Fig.10 show that for systems considering the user priorities, 5 UT servers, 7 DT servers and 5 computation servers are enough and the minimum total cost is 360. For that without considering the user priorities, the minimum numbers of all type of servers are all 6 and the minimum total cost is 390, 8.33% higher than systems considering the user priorities.

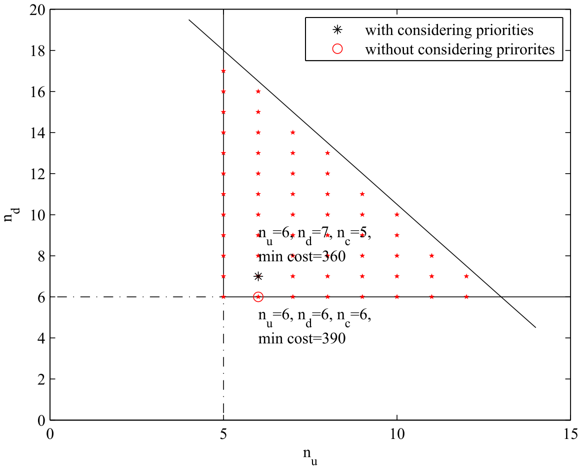


FIGURE 10. Optimal resource deployment of given system setup with and without considering priorities.

Through comparisons in Sections VI-B and VI-D, one can gain an insight into the future scenario that it is best for service providers to offer customers with differentiated service priorities either from the interest perspective or from the cost perspective.

VII. CONCLUSION AND FUTURE WORK

This paper investigated a queueing network model for a BS providing pure UT, DT and MEC service simultaneously. Based on the proposed model, the admission control optimization and the resource deployment optimization from the standpoint of service providers were investigated. An optimal resource deployment algorithm was developed. Simulations have been done to verify all the analysis. This work provided a basic and novel queueing model for MEC different from existing solutions. In the future, more realistic and complicated scenarios will be investigated. Several future research directions on this model and the challenges are listed as follows.

- The service time distribution of UT & DT servers needs to be modeled by a more generalized one. As mentioned in Section II-A, the $M/G/n/\infty$ queue model simplified the modeling of the UT & DT queue and thus accurate modeling of the real the exact distribution of service time of UT & DT servers is still a challenge.

Additionally, even modeling it as with the $M/G/n/\infty$ queue model, the queueing analysis considering multiple priority classes is still a very complicated problem as the mean residual service time is hard to be derived [25].

- A queueing network model and resource allocation can be jointly investigated. In the current model, different user classes are differentiated by the priorities and they can also be differentiated by allocating different resources. Combining them together could obtain the better scheduling result. One way to combine them is that assuming each class has its own service time distribution, where the distribution is related to the resource allocated to the user class. This approach is more realistic, but it still very challenging to analyze such a model [24].
- By considering the real characteristics such as retransmission, finite buffer length, finite waiting time, more complicated and realistic models of the queueing network can be developed.
- The discrete-time queueing network model [31] will be more suitable than the current used continuous-time model.
- The outage probability describes the percent of users that are not served. It might be better to use it as the optimization criteria or constraints.

APPENDIX
PROOF OF PROPOSITION 2

Proof: Proposition 2 can be proved by the following lemma in book [25].

Lemma 1: For an $M/M/m/\infty$ non-preemptive priority queue with K priority classes and all having exponentially distributed service times with common mean $1/\mu$, the average delay for all priority classes $k = 1, 2, \cdots, K$ including waiting time and service time is

$$T = \begin{cases} \frac{1}{\mu} + \frac{R}{(1-\phi_1)}, & k = 1, \\ \frac{1}{\mu} + \frac{R}{(1-\phi_1 \cdots - \phi_{k-1})(1-\phi_1 \cdots - \phi_k)}, & k > 1. \end{cases} \tag{51}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_K$ are the access rates of all classes,

$$\phi_k \triangleq \frac{\lambda_k}{m\mu}, \quad \phi \triangleq \sum_{k=1}^{K} \phi_k, \tag{52}$$

$$p_0 \triangleq \left[\sum_{n=0}^{n-1} \frac{(m\phi)^n}{n!} + \frac{(m\phi)^m}{m!(1-\phi)} \right]^{-1}, \tag{53}$$

$$P_Q \triangleq \frac{p_0 (m\phi)^m}{m!(1-\phi)}, \tag{54}$$

and the mean residual service time

$$R \triangleq \frac{P_Q}{m\mu}. \tag{55}$$

The UT service queue is an $M/M/n_u/\infty$ with $(J_u + J_c)$ priority classes. And in front of class-j_u users, there are $(j_u - 1 + j_{c>j_u})$ user classes. According to Lemma 1 and the definitions made in (10)-(15), the expression of the average delay of the class-j_u pure UT users can be obtained as (16) in Proposition 2. ■

REFERENCES

[1] A. Osseiran *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[2] C.-X. Wang *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[3] M. Cai, J. N. Laneman, and B. Hochwald, "Beamforming codebook compensation for beam squint with channel capacity constraint," in *Proc. IEEE ISIT*, Aachen, Germany, Jun. 2017, pp. 76–80.

[4] S. Guo, H. Zhang, P. Zhang, D. Wu, and D. Yuan, "Generalized 3-D constellation design for spatial modulation," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3316–3327, Aug. 2017.

[5] S. Guo, H. Zhang, P. Zhang, and D. Yuan, "Adaptive mapper design for spatial modulation with lightweight feedback overhead," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 8940–8950, Oct. 2017.

[6] M. T. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile edge computing: A taxonomy," in *Proc. AFIN*, 2014, pp. 1–7.

[7] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Proc. ISCO*, 2016, pp. 1–8.

[8] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[9] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

[10] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[11] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading decision and resource allocation for multi-user multi-task mobile cloud," in *Proc. IEEE ICC*, May 2016, pp. 1–6.

[12] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading and resource allocation for computation and communication in mobile cloud with computing access point," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[13] J. Kwak, Y. Kim, J. Lee, and S. Chong, "DREAM: Dynamic resource and task allocation for energy minimization in mobile cloud systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2510–2523, Dec. 2015.

[14] Z. Jiang and S. Mao, "Energy delay trade-off in cloud offloading for mutli-core mobile devices," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.

[15] Y. Mao, J. Zhang, S. Song, and K. B. Letaief, "Power-delay tradeoff in multi-user mobile-edge computing systems," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.

[16] Y. Mao, J. Zhang, Z. Chen, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[17] K. Wang, K. Yang, and C. Magurawalage, "Joint energy minimization and resource allocation in C-RAN with mobile cloud," *IEEE Trans. Cloud Comput.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/document/7393804/

[18] Y. Wen, W. Zhang, and H. Luo, "Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2716–2720.

[19] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017.

[20] P. Di Lorenzo, S. Barbarossa, and S. Sardellitti. (2016). "Joint optimization of radio resources and code partitioning in mobile edge computing." [Online]. Available: https://arxiv.org/abs/1307.3835

[21] D. T. Hoang, D. Niyato, and P. Wang, "Optimal admission control policy for mobile cloud computing hotspot with cloudlet," in *Proc. IEEE WCNC*, Apr. 2012, pp. 3145–3149.

[22] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Joint allocation of computation and communication resources in multiuser mobile cloud computing," in *Proc. IEEE 14th SPAWC*, Jun. 2013, pp. 26–30.

[23] S. Guo, D. Wu, H. Zhang, and D. Yuan, "Queueing network model and average delay analysis for mobile edge computing," in *Proc. IEEE ICNC*, Maui, HI, USA, Mar. 2018, pp. 1–6.

[24] H. R. Gail, S. L. Hantler, and B. A. Taylor, "Analysis of a non-preemptive priority multiserver queue," *Adv. Appl. Probab.*, vol. 20, no. 4, pp. 852–879, Dec. 1988.

[25] D. P. Bertsekas and R. G. Gallager, *Data Networks*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 1992.

[26] D. Gross, *Fundamentals of Queueing Theory*. Hoboken, NJ, USA: Wiley, 2008.

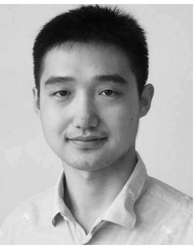
[27] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM J. Optim.*, vol. 11, no. 3, pp. 796–817, Sep. 2004.

[28] C. Feng, "Polynomial optimization based approaches to system design, analysis and identification," Ph.D. dissertation, Pennsylvania State Univ., State College, PA, USA, 2013.

[29] D. Henrion, J.-B. Lasserre, and J. Löfberg, "GloptiPoly 3: Moments, optimization and semidefinite programming," *Optim. Methods Softw.*, vol. 24, nos. 4–5, pp. 761–779, 2008.

[30] D. Henrion and J.-B. Lasserre, "Detecting global optimality and extracting solutions in GloptiPoly," in *Positive Polynomials in Control*. Berlin, Germany: Springer, 2005, pp. 293–310.

[31] I. Rubin and Z.-H. Tsai, "Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems," *IEEE Trans. Inf. Theory*, vol. 35, no. 3, pp. 637–647, May 1989.



SHUAISHUAI GUO (M'17) received the B.E. and Ph.D. degrees in communication and information systems from the School of Information Science and Engineering, Shandong University, Jinan, China, in 2011 and 2017, respectively. He visited The University of Tennessee at Chattanooga, USA, from 2016 to 2017, and the King Abdullah University of Science and Technology, Saudi Arabia, since 2017. He is currently a Post-Doctoral Research Fellow with Shandong University.

His research interests include wireless multiple-input multiple-output communications, signal processing for communications, optical wireless communication, and mobile edge computing.

DALEI WU (M'11) received the B.S. and M.Eng. degrees in electrical engineering from Shandong University, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer engineering from the University of Nebraska-Lincoln, Lincoln, NE, USA, in 2010.

From 2011 to 2014, he was a Post-Doctoral Research Associate with the Mechatronics Research Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Since 2014, he has been an Assistant Professor with the Department of Computer Science and Engineering, The University of Tennessee at Chattanooga. His research interests include wireless communications and networking, cyber-physical systems, and intelligent systems.

HAIXIA ZHANG (M'08–SM'11) received the B.E. degree from the Department of Communication and Information Engineering, Guilin University of Electronic Technology, Guilin, China, in 2001, and the M.Eng. and Ph.D. degrees in communication and information systems from the School of Information Science and Engineering, Shandong University, Jinan, China, in 2004 and 2008, respectively.

From 2006 to 2008, she was an Academic Assistant with the Institute for Circuit and Signal Processing, Munich University of Technology, Munich, Germany. She is currently a Full Professor with Shandong University. Her current research interests include cognitive radio systems, cooperative (relay) communications, cross-layer design of wireless communication networks, space-time process techniques, precoding/beamforming, and fifth-generation wireless communications.

Dr. Zhang has been actively participating in many academic events, serving as a Technical Program Committee Member and a Session Chair, giving invited talks for conferences, and serving as a reviewer for numerous journals. She is an Associate Editor of the *International Journal of Communication Systems.*

DONGFENG YUAN (SM'01) received the M.S. degree from the Department of Electrical Engineering, Shandong University, Jinan, China, in 1988, and the Ph.D. degree from the Department of Electrical Engineering, Tsinghua University, Beijing, China, in 2000.

From 1993 to 1994, he was with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada. He was with the Department of Electrical Engineering, University of Erlangen, Erlangen, Germany, from 1998 to 1999; with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, from 2001 to 2002; with the Department of Electrical Engineering, Munich University of Technology, Munich, Germany, in 2005; and with the Department of Electrical Engineering, Heriot-Watt University, Edinburgh, U.K., in 2006. He is currently a Full Professor with the School of Information Science and Engineering, Shandong University. His current research interests include cognitive radio systems, cooperative (relay) communications, and fifth-generation wireless communications.