# A Multi-User Mobile Computation Offloading and Transmission Scheduling Mechanism for Delay-Sensitive Applications

Changyan Yi, *Member, IEEE*, Jun Cai, *Senior Member, IEEE*, and Zhou Su

**Abstract**—In this paper, a mobile edge computing framework with multi-user computation offloading and transmission scheduling for delay-sensitive applications is studied. In the considered model, computation tasks are generated randomly at mobile users along the time. For each task, the mobile user can choose to either process it locally or offload it via the uplink transmission to the edge for cloud computing. To efficiently manage the system, the network regulator is required to employ a network-wide optimal scheme for computation offloading and transmission scheduling while guaranteeing that all mobile users would like to follow (as they may naturally behave strategically for benefiting themselves). By considering tradeoffs between local and edge computing, wireless features and noncooperative game interactions among mobile users, we formulate a mechanism design problem to jointly determine *a computation offloading scheme*, *a transmission scheduling discipline*, and *a pricing rule*. A queueing model is built to analytically describe the packet-level network dynamics. Based on this, we propose a novel mechanism, which can maximize the network social welfare (i.e., the network-wide performance), while achieving a game equilibrium among strategic mobile users. Theoretical and simulation results examine the performance of our proposed mechanism, and demonstrate its superiority over the counterparts.

**Index Terms**—Mobile dge computing, computation offloading, transmission scheduling, delay sensitive, mechanism design, game

## 1 Introduction

Due to the increasing popularity of smart mobile devices (e.g., smart phones and tablets), a variety of novel mobile applications, such as natural language processing, face recognition, interactive gaming, augmented reality and healthcare monitoring, are developed and recently attracting great interests [1]. These mobile applications commonly consume intensive computation resources and tremendously high energy, which prevent them from being executed locally by most resource-constrained mobile devices. To overcome these computation and power limitations on mobile devices and at the same time achieve high network efficiency, mobile cloud computing is envisioned as a promising paradigm, which allows mobile devices to fully/partially offload their computation tasks to resource-rich cloud infrastructures (such as Amazon EC2, Google Compute Engine and Microsoft Azure) [2]. However, since traditional public clouds are usually located far away from mobile devices/users, the data exchange via the wide area network may result in a considerably long latency. To further overcome this drawback, mobile edge computing has been proposed as an alternative solution. By enabling cloud computing capabilities at the edges of ubiquitous radio access networks (e.g., 3G/4G macro-cell or small-cell base stations) that are close to mobile users, mobile edge computing can provide pervasive, prompt and agile computation services at anytime and anywhere [3], [4].

For mobile edge computing, offloading computation tasks must involve wireless communications between mobile users and the edge cloud, and thus its performance highly depends on the wireless access efficiency [5]. Due to inherently limited radio resources [6], if the wireless access among multiple mobile users for computation offloading is not well coordinated, the wireless network capacity may be quickly strained by the overwhelmed offloading tasks, leading to low transmission efficiency (e.g., a long data transmission delay) and eventually resulting in dissatisfactions on mobile edge computing services. This prompted some recent research efforts [7], [8], [9], [10] in studying both computation offloading and radio resource allocation for mobile edge computing. Nevertheless, there are still some critical issues which are closely related to practical implementations and have not been well addressed:

i) In practice, mobile applications, such as live streaming applications, may generate a stream of computation tasks, which arrive randomly at mobile users along the time [11], [12]. Such potential network uncertainties requests a dynamic management of the system with long-term performance guarantees.
ii) Naturally, different mobile users may run computation-intensive applications with different delay sensitivities. For instance, augmented reality services

- *C. Yi and J. Cai are with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada. E-mail: {changyan.yi, jun.cai}@umanitoba.ca.*
- *Z. Su is with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China. E-mail: zhousu@ieee.org.*

*Manuscript received 29 Jan. 2018; revised 7 Dec. 2018; accepted 29 Dec. 2018. Date of publication 9 Jan. 2019; date of current version 3 Dec. 2019. (Corresponding author: Jun Cai.)
Digital Object Identifier no. 10.1109/TMC.2019.2891736*

normally require delays less than 100 ms, while 4K live videos can tolerate up to 500 ms delays [13]. Thus, a proper delay-sensitive scheduling mechanism should be adopted to meet the heterogeneous quality-of-service (QoS) requirements of mobile users in computation offloading.

iii) Most importantly, human/device intelligence allows mobile users to behave strategically and selfishly, even though they are rational [14], [15], [16]. As a result, mobile users may pursue different self-interests by selecting their own computation offloading strategies [17], [18]. Obviously, such strategic behaviors may affect the wireless traffic volume in computation offloading and the overall system performance [19]. Since transmission resources are limited and shared among mobile users, this naturally leads to a noncooperative competitions at user ends [20] and motivates the need of considering a game-theoretic decision making process [7], [21], [22], [23]. Therefore, in order to guarantee efficiency and robustness, mobile edge computing system should be designed for not only maximizing the overall network performance, but also regulating the strategic and noncooperative interactions (i.e., the game behaviors) of smart mobile users in a desired manner, i.e., making the outcome of the network-wide optimal solution achieve an equilibrium such that all mobile users are satisfied.

However, addressing all these features raises new challenges in the management of mobile edge computing systems. Specifically,

a. It is very difficult to explicitly analyze the complicated relationship between dynamics of the control system and QoS of mobile users' delay-sensitive computation tasks, and thus a packet-level queueing modeling is required.
b. Since both computation offloading decisions and wireless transmission scheduling contribute to the overall performance of mobile edge computing systems, they have to be jointly determined with the consideration of their mutual effects, which results in a complicated joint optimization problem.
c. The designed mechanism should be capable of inducing mobile users to voluntarily participate in the system following the network-wide optimal management. However, by considering the fact that mobile users may strategically deviate the optimal solution unilaterally for potentially maximizing their own interests, this requests us to integrate the regulation of mobile users' noncooperative game behaviors into the network-wide optimization. Obviously, more constraints/requirements related to individuals have to be introduced and included, which further increase the complexity of the original optimization problem.

To tackle these difficulties, in this paper, we propose a novel mechanism which can efficiently manage multi-user computation offloading and transmission scheduling for delay-sensitive applications in mobile edge computing. In this work, we consider random arrivals of heterogeneous delay-sensitive computation tasks at different mobile users. Upon receiving a task, the associated mobile user can strategically decide to either process it locally or offload it via the uplink cellular transmission to the edge (i.e., the wireless access point) for cloud computing. Since offloading will incur delay costs and cloud service charges, while running computations locally may consume a large amount of energy, there always exists a tradeoff among energy consumption, delay performance and payment cost for mobile users. The packet level operation of such management framework is first modeled as a dynamic priority queue, where the heterogeneity of mobile users in terms of their delay sensitivities in running different applications is taken into account. Based on this formulated queueing model, we then design a mechanism to jointly determine the computation offloading scheme, the transmission scheduling discipline and the pricing rule. Theoretical analyses show that our proposed mechanism can reach an equilibrium which not only maximize the network social welfare,[1] but also prevent individual mobile users from deviating the network-wide optimal solution unilaterally.

The main contributions of this paper are in the following.

- Joint computation offloading and wireless transmission scheduling for edge computing with delay-sensitive applications are formulated as a mechanism design problem upon a dynamic queueing model.
- A delay-dependent prioritized transmission scheduling discipline is developed to minimize the total delay cost of all mobile users for any given computation offloading decision.
- With the objective of maximizing the network social welfare, an optimal computation offloading scheme is derived through the convex optimization.
- An appropriate pricing rule is designed to make the network-wide optimal computation offloading and transmission scheduling scheme achieve an equilibrium of the noncooperative strategy making process among individual mobile users.
- Theoretical analyses and simulation results examine the performance of our proposed mechanism, and demonstrate its feasibility and superiority compared to the counterparts.

The rest of this paper is organized as follows: Section 2 gives a brief review of related works and highlights the novelties of this work. Section 3 describes the considered system model and the problem formulation. In Section 4, an efficient multi-user computation offloading and transmission scheduling mechanism for delay-sensitive applications in mobile edge computing is proposed and analyzed. Simulation results are presented in Section 5, followed by the discussions of conclusion and future work in Section 6.

## 2 Related Work

Mobile edge computing, as a key enabling technology for Internet-of-things (IoT) and 5G, has drawn a lot of research

[1] From the network point of view, this is a common objective, as it is equivalent to studying how network resource can be fully utilized and optimized to maximize the overall system performance.

TABLE 1
Important Notations in This Paper

| Symbol | Meaning |
|---|---|
| $\mathcal{N}$ | set of mobile users |
| $M$ | number of cellular channels |
| $\lambda_i$ | arrival rate of computation tasks at mobile user $i$ |
| $\lambda_i^{\text{local}}$ | local processing rate of mobile user $i$ |
| $\lambda_i^{\text{offload}}$ | computation offloading rate of mobile user $i$ |
| $S_i$ | size of mobile user $i$'s computation tasks |
| $R_i$ | uplink transmission rate of mobile user $i$ |
| $\mathbb{E}[U_i]$ | expected utility of mobile user $i$ |
| $C_i$ | expected cost of mobile user $i$ |
| $T_i$ | experienced delay of user $i$ in offloading each task |
| $1/\mu_i$ | mean transmission time for one task of mobile user $i$ |
| $\rho_i$ | offloading traffic intensity of mobile user $i$ |
| $L_i$ | workload of mobile user $i$ in uplink transmissions |
| $L$ | total workload in uplink transmissions |
| $a_i(\tau)$ | amount of tasks in mobile user $i$'s buffer at time $\tau$ |
| $W_i(\cdot)$ | delay cost function of mobile user $i$ |
| $V_i(\cdot)$ | utility gain of user $i$ in computation offloading |
| $Y(\cdot)$ | network operating cost |
| $\mathcal{SW}$ | network social welfare |
| $\beta_i$ | computation offloading ratio of mobile user $i$ |
| $\pi_i$ | service charge on mobile user $i$ |
| $\zeta$ | transmission scheduling discipline |
| $Q(\cdot)$ | formulated queueing management system |

attentions recently. For instance, Chen et al. in [23] formulated a decentralized computation offloading game, where each mobile user could choose to either fully offload its computing task to the edge cloud or process it locally for minimizing the overhead. In [24], Dinh et al. proposed an optimization approach for determining computation offloading and CPU frequency of a mobile user with the objective of minimizing both its task execution time and energy consumption. These works focused on the computation offloading process only, while ignoring the resource allocation in wireless access networks.

Some recent works have been dedicated in studying joint radio-and-computational resource management [8], [9], [10]. Specifically, Wang et al. in [8] presented a decentralized method to jointly optimize computation offloading, spectrum allocation and content caching in wireless cellular networks with the mobile edge computing capability. Sardellitti et al. in [9] developed an iterative algorithm based on the successive convex approximation for minimizing the total energy consumption of mobile users under delay and power budgets. You et al. in [10] introduced an optimal threshold-based policy for managing offloading data volumes and channel access opportunities in time-division multiple access (TDMA) mobile edge computing systems. However, all these works employed a common assumption that the network scenario was quasi-static, i.e., the set of computing tasks was assumed to remain unchanged regardless of the potential temporal dynamics.

A few works have started to analyze mobile edge computing in dynamic network scenarios with random task arrivals. For example, Kwak et al. in [12] investigated the energy-delay tradeoff for mobile edge computing with various types of tasks that were generated randomly at the user end. The minimization of long-term average task execution costs was discussed in [25] and [26], where the former optimized the computation offloading by using Markov decision process (MDP) and the latter jointly controlled the local CPU frequency, modulation scheme and data rate under a semi-MDP framework. Besides, using queueing theory to analyze the performance of mobile edge computing has been initiated in [27], [28]. Because of the complicated temporal correlation in dynamic operations, most of these works were restricted to a single-user case only.

Mechanism design integrating queueing scheduling is a novel technique for optimally managing dynamic systems with multi-user interactions [29], [30]. Its recent applications include power allocations for wireless networks with dynamic energy harvesting [31] and delay-sensitive transmission managements in remote healthcare networks [32], [33], [34]. However, mechanisms developed in these works cannot be applied for solving the problem in this paper, because none of them can capture the variable inputs of queueing systems (due to the potential computation offloading), and most of them relied on the assumptions of specific distributions (e.g., exponential and poisson) in the queueing modeling.

In summary, different from all existing works in the literature, this paper studies a joint optimization of computation offloading and transmission scheduling for mobile edge computing systems with delay-sensitive applications by considering generally distributed network dynamics and multiple heterogeneous strategic mobile users.

## 3 System Model and Problem Formulation

In this section, the network model of multi-user computation offloading and transmission scheduling for mobile edge computing with delay-sensitive applications is first described. Then, utility functions and strategies of mobile users are defined. After that, a mechanism design problem, which jointly considers computation offloading and transmissions scheduling, is formulated. For convenience, Table 1 lists some important notations used in this paper.

### 3.1 Network Model

Consider a group of mobile users, denoted by $\mathcal{N}$, who are running computation-intensive and delay-sensitive applications.[2] Each mobile user has a stream of computation tasks that are required to be completed as soon as possible. There exists a wireless access point through which mobile users can offload their computation tasks to the edge cloud. In practice, the wireless access point can be a 3G/4G macrocell or small-cell base station, which manages wireless transmissions for all associated mobile users over $M$ homogeneous cellular channels[3] by using OFDMA. Note that similar to [11], [36], we particularly focus on computation offloading and uplink transmission scheduling for mobile

[2] Delay-sensitive applications can be described by three potential features: i) there is a maximum tolerable delay requirement for each of its computation task; ii) its computation task suffers a cost (e.g., a loss of value) which increases with the delay before it has been completed (i.e., a delay cost); and iii) the combination of i) and ii). Practical examples include augmented reality services which normally require delays less than 100 ms [13], and mobile gaming where the quality of experience (QoE) commonly decreases with the increase of delay [35].

[3] This means that i) each channel has the same bandwidth; and ii) all channels experience the same fading effect.

users, and ignore the communication overhead for the edge cloud to send computation outcomes back to mobile users. This is because the size of computation outcomes is commonly negligible compared to that of computation tasks (each of which may consist of various mobile system settings, programs and input parameters).

For each mobile user $i, \forall i \in \mathcal{N}$, let the dynamic arrival of its computation tasks be a generally distributed random process with an average rate $\lambda_i$, and consider that each task has a random size of $S_i$ with a finite mean $\mathbb{E}[S_i]$ and requires a random number of CPU cycles $Z_i$ with a finite mean $\mathbb{E}[Z_i]$. Upon receiving each computation task, the mobile user can choose to either process it locally or offload it to the edge cloud via the uplink cellular transmission. The network regulator has to determine a long-term network-preferable scheme for joint computation offloading and uplink transmission scheduling while guaranteeing that all individual mobile users would like to follow. Denote $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)$ as the computation offloading scheme, where $\beta_i \in [0, 1]$ indicates the long-term offloading ratio (or equivalently the offloading probability for each computation task) of mobile user $i, \forall i \in \mathcal{N}$. Then, the average arrival rates of computation tasks for offloading and local processing can be calculated as $\lambda_i^{\text{offload}} = \beta_i \lambda_i$ and $\lambda_i^{\text{local}} = (1 - \beta_i)\lambda_i$, respectively. If a mobile user decides to offload a computation task to the edge cloud (i.e., not to process it locally) while cannot immediately deliver it due to the channel shortage in uplink cellular transmissions, this computation task will be temporarily stored in the buffer of the associated mobile user until it is scheduled for transmission. By considering the fact that the resources for wireless transmissions (i.e., radio channels) are much more constrained than those for higher-layer traffic management and computing [5], [23], we limit our focus on wireless transmission delays only and omit the detailed discussions on other potential latencies caused by task processing in local/edge cloud computing[4]. Furthermore, we do not consider buffer overflow.[5] Then, the management framework of computation offloading and transmission scheduling for mobile edge computing can be modeled as a queueing system, as shown in Fig. 1.

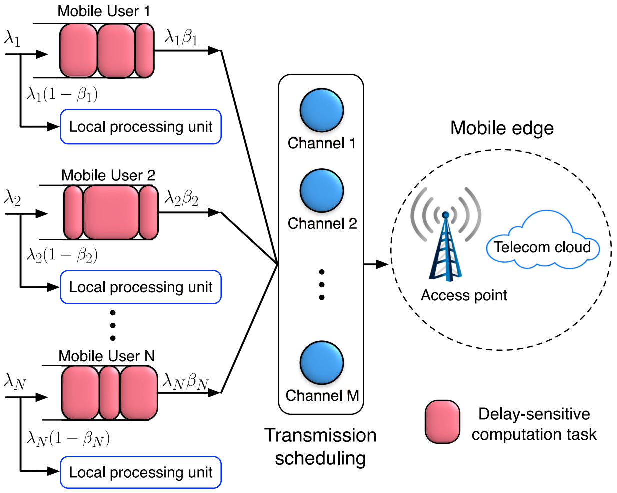


Fig. 1. An illustration of computation offloading and uplink transmission scheduling for mobile edge computing.

With a proper scheduling, the uplink transmission rate of each mobile user $i, \forall i \in \mathcal{N}$, on one cellular channel can be expressed as

$$r_i = B \cdot \log_2\left(1 + \frac{|h_i|^2 \cdot P_i \cdot d_i^{-\eta}}{\sigma^2}\right), \tag{1}$$

where $B$ and $\sigma^2$ denote the channel bandwidth and the variance of additive Gaussian noise, respectively; $P_i$ is the predetermined transmission power of mobile user $i$; $|h_i|^2$ captures the Rayleigh fading effect and follows an exponential distribution with a unity mean; $d_i^{-\eta}$ signifies the path loss effect, where $d_i$ specifies the distance from mobile user $i$ to the wireless access point and $\eta \geq 2$ is the path loss exponent. Consider that mobile users are roaming within the cell according to a random mobility pattern, e.g., following a random waypoint model [37]. Then, the distance between mobile user $i, \forall i \in \mathcal{N}$, and the wireless access point can be modeled as a random variable $D_i$. Due to the randomness of both $|h_i|^2$ and $d_i$, the uplink transmission rate of mobile user $i, \forall i \in \mathcal{N}$, can be represented by a generally distributed random variable $R_i$ with a probability density function (PDF) $f_{R_i}(\cdot)$ and a finite mean $\mathbb{E}[R_i]$. Similar definitions can also be found in [38] and [39].

### 3.2 Utility Functions and Strategies of Mobile Users

Since mobile users are rational and potentially selfish [7], [17], their individual utility functions and strategies have to be carefully taken into account in the management of computation offloading and uplink transmission scheduling. Given the computation offloading and transmission scheduling scheme, the long-term expected utility of each mobile user $i, \forall i \in \mathcal{N}$, can be represented as

$$\mathbb{E}[U_i] = \lambda_i \cdot \mathbb{E}[v_i] - C_i(\lambda_i^{\text{local}}, \lambda_i^{\text{offload}}), \tag{2}$$

where $\mathbb{E}[v_i]$ is the mean valuation of completing one computation task for mobile user $i$, so that $\lambda_i \mathbb{E}[v_i]$ indicates the long-term expected valuation gained by mobile user $i$ (including both local and edge cloud computing); $C_i(\lambda_i^{\text{local}}, \lambda_i^{\text{offload}})$ denotes the total expected cost of mobile user $i$, which consists of a processing cost for local computing $C_i^{\text{local}}$, a contract cost for edge cloud computing $C_i^{\text{cont}}$, an uplink transmission cost $C_i^{\text{tx}}$ and a delay cost $C_i^{\text{delay}}$. Following the similar definitions in [8], [24], we discuss all these cost terms in detail as follows.

For each mobile user $i, \forall i \in \mathcal{N}$, its computation tasks can be executed by either the local or the edge cloud processing units. Local processing will cause energy consumption on mobile user $i$, which can be computed as

$$C_i^{\text{local}} = \lambda_i^{\text{local}} \mathbb{E}[Z_i] \varphi_i = (1 - \beta_i)\lambda_i \mathbb{E}[Z_i] \varphi_i, \tag{3}$$

[4] It is worth noting that this does not mean that the task processing time is ignored in the proposed analytical framework. As will be shown in Section 3.2, the task processing time at local CPU is indirectly considered in (3) for calculating the local processing cost, and the task processing time at the cloud can be included by the formulation of the general delay cost function $W(\cdot)$. Therefore, the detailed study on the task processing can be easily integrated, and the assumption of neglecting the processing time of task computing is not necessary.

[5] The storages of current mobile devices are in hundred gigabytes so that the size of each user's buffer can be considerably large. Moreover, with the capability of local processing, the traffic flow can be well balanced, and thus buffer overflow barely exists.

where $\mathbb{E}[Z_i]$ denotes the average number of CPU cycles required by one computation task of mobile user $i$, and $\varphi_i$ specifies the average energy cost per CPU cycle, which can be obtained by existing measurement approaches [40]. Thus, $C_i^{\text{local}}$ denotes the mean cost per unit of time for mobile user $i$ in local processing. In contrast, for computation tasks that are offloaded by mobile user $i$ to the edge cloud, a contract cost exists (to compensate the edge in consuming cloud computing resources) and can be calculated as

$$C_i^{\text{cont}} = \lambda_i^{\text{offload}} \pi_i^{\text{cont}} = \beta_i \lambda_i \pi_i^{\text{cont}}, \tag{4}$$

where $\pi_i^{\text{cont}}$ is the average contract cost for one computation task, and $C_i^{\text{cont}}$ represents the mean contract cost per unit of time for mobile user $i$ in offloading computation tasks.

Besides, to offload computation tasks to the edge cloud, each mobile user $i, \forall i \in \mathcal{N}$, has to rely on the uplink cellular transmission, which results in another cost, denoted by $C_i^{\text{tx}}$. Intuitively, $C_i^{\text{tx}}$ should include transmission energy cost and transmission service charge, i.e.,

$$C_i^{\text{tx}} = \lambda_i^{\text{offload}} (\mathcal{E}_i^{\text{tx}} + \pi_i) = \beta_i \lambda_i (\mathcal{E}_i^{\text{tx}} + \pi_i^{\text{tx}}), \tag{5}$$

where $\mathcal{E}_i^{\text{tx}}$ and $\pi_i^{\text{tx}}$ stand for the mean transmission energy cost and transmission service charge for one computation task of mobile user $i$, respectively. To be more specific, $\mathcal{E}_i^{\text{tx}}$ is computed as

$$\mathcal{E}_i^{\text{tx}} = \mathbb{E}\left[\frac{S_i}{R_i}\right] P_i \psi_i = \left(\mathbb{E}[S_i] \int_{-\infty}^{\infty} \frac{1}{r} f_{R_i}(r) dr\right) P_i \psi_i, \tag{6}$$

where $\mathbb{E}[S_i/R_i]$, $P_i$ and $\psi_i$ are mobile user $i$'s average uplink transmission time for one computation task, pre-determined transmission power, and cost coefficient of transmission energy, respectively. By substituting (6) into (5), we have

$$C_i^{\text{tx}} = \beta_i \lambda_i \left( \left(\mathbb{E}[S_i] \int_{-\infty}^{\infty} \frac{1}{r} f_{R_i}(r) dr\right) P_i \psi_i + \pi_i^{\text{tx}} \right). \tag{7}$$

In addition, since mobile users are running different delay-sensitive applications, there exists a specific delay cost, i.e., $C_i^{\text{delay}}$, on each mobile user $i, \forall i \in \mathcal{N}$, due to the delay in uplink transmissions for computation offloading. Without loss of generality, we define

$$C_i^{\text{delay}} = \lambda_i^{\text{offload}} \mathbb{E}[W_i(T_i)] = \beta_i \lambda_i \mathbb{E}[W_i(T_i)], \tag{8}$$

where $T_i$ is a random variable which models the time duration from a computation task arriving at mobile user $i$ to being received by the wireless access point (i.e., $T_i$ describes the total delay that an offloaded computation task will experience in uplink transmissions). Obviously, $T_i$ is determined by the uplink transmission scheduling and the volumn of offloaded traffics $\boldsymbol{\lambda}^{\text{offload}} = (\lambda_1^{\text{offload}}, \lambda_2^{\text{offload}}, \ldots, \lambda_N^{\text{offload}})$ or equivalently the computation offloading scheme $\boldsymbol{\beta}$. $W_i(\cdot)$ denotes a general delay cost function which is defined to be increasing and convex with respect to $T_i$. Clearly, $W_i(\cdot)$ well depicts the intuition in practice that mobile users are commonly less sensitive to a small delay, but their concerns grow dramatically when the delay continuously increases and becomes very large [41], [42], e.g., approaching a certain deadline. The physical meaning of $C_i^{\text{delay}}$ can be interpreted as the cost resulted by mobile user $i$'s dissatisfaction on the quality of experience in running its delay-sensitive application and the energy cost in buffering the computation tasks in uplink offloading.

After defining all cost terms as in (3), (4), (5), (6), (7), and (8), the expected utility of each mobile user $i, \forall i \in \mathcal{N}$, can be written as

$$\mathbb{E}[U_i] = \lambda_i \mathbb{E}[v_i] - (C_i^{\text{local}} + C_i^{\text{cont}} + C_i^{\text{tx}} + C_i^{\text{delay}}). \tag{9}$$

To further characterize its properties and simplify equations, let us define

$$\begin{aligned} V_i(\beta_i) &= \lambda_i \mathbb{E}[v_i] - C_i^{\text{local}} - \beta_i \lambda_i \mathcal{E}_i^{\text{tx}} \\ &= \lambda_i \mathbb{E}[v_i] - (1-\beta_i)\lambda_i \mathbb{E}[Z_i]\varphi_i - \beta_i \lambda_i \mathcal{E}_i^{\text{tx}}, \ \forall i \in \mathcal{N}. \end{aligned} \tag{10}$$

Intuitively, we must have $\mathbb{E}[Z_i]\varphi_i > \mathcal{E}_i^{\text{tx}}$, i.e., the local processing cost is larger than the associated uplink transmission energy cost. Otherwise, mobile user $i$ will never offload its computation tasks because doing so will definitely result in a utility loss. In addition, there should be a lower bound on $\mathbb{E}[v_i]$ such that $\mathbb{E}[v_i] \geq \mathbb{E}[Z_i]\varphi_i$, i.e., the value of completing a computation task is larger or equal to its local processing cost. Otherwise, mobile user $i$ will not be willing to run such application because it is too computation-intensive and doing so may bring a high risk of suffering a utility loss. By ignoring these trivial cases, we can show that $V_i(\beta_i)$ must be an increasing function with respect to $\beta_i$.

With (10), the utility function (9) for each mobile user $i, \forall i \in \mathcal{N}$, can now be rewritten as

$$\begin{aligned} \mathbb{E}[U_i] &= V_i(\beta_i) - C_i^{\text{cont}} - \beta_i \lambda_i \pi_i^{\text{tx}} - C_i^{\text{delay}} \\ &= V_i(\beta_i) - \beta_i \lambda_i (\pi_i^{\text{cont}} + \pi_i^{\text{tx}}) - \beta_i \lambda_i \mathbb{E}[W_i(T_i)]. \end{aligned} \tag{11}$$

Note that the utility function shown above is derived based on the assumption that all mobile users will follow the designed computation offloading and transmission scheduling management. However, as indicated previously, since mobile users are rational and potentially selfish, though they are not able to manipulate the uplink transmission scheduling (because it is centrally controlled by the network regulator), they can strategically determine their own computation offloading ratios, i.e., $\tilde{\boldsymbol{\beta}} = \{\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_N\}$, which may be different from the network-wide optimal computation offloading scheme $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_N\}$, if and only if they can benefit from such behavior (i.e., improve their own utilities). Obviously, such strategic behaviors from mobile users will change the traffic volume in uplink transmissions for computation offloading. Considering that the uplink spectrum resources are limited and shared among multiple mobile users, any individual strategic behavior from a mobile user on varying its offloading ratio can directly affect the performance of the uplink transmission scheduling, which may further trigger frequent changes of offloading strategies from all mobile users, leading to system instability. Thus, in order to ensure the robustness and achieve a mutually satisfactory equilibrium, our designed mechanism has to satisfy the following essential requirements.

- *Incentive compatibility:* For each mobile user $i, \forall i \in \mathcal{N}$, its expected utility should be maximized when $\tilde{\beta}_i = \beta_i$, i.e.,

$$\mathbb{E}[U_i(\tilde{\beta}_i)] \leq \mathbb{E}[U_i(\beta_i)],\ \forall \tilde{\beta}_i \neq \beta_i, \tag{12}$$

where $\mathbb{E}[U_i(\tilde{\beta}_i)]$ indicates the expected utility of mobile user $i$ with computation offloading strategy $\tilde{\beta}_i$ and can be computed by substituting $\tilde{\beta}_i$ into (9).

- *Individual rationality:* For each mobile user $i, \forall i \in \mathcal{N}$, its expected utility should always be non-negative if it follows the network-determined computation offloading scheme $\boldsymbol{\beta}$, i.e.,

$$\mathbb{E}[U_i(\beta_i)] \geq 0. \tag{13}$$

By satisfying both incentive compatibility (12) and individual rationality (13), all mobile users will be willing to participate in the system and manage their computation offloading strictly following the designed mechanism.

### 3.3 Problem Formulation

The network social welfare ($\mathcal{SW}$) of the considered management framework for delay-sensitive mobile computation offloading and transmission scheduling is defined as

$$\mathcal{SW} = \mathbb{E}[U_{\text{reg}}] + \sum_{i=1}^{N} \mathbb{E}[U_i], \tag{14}$$

where $\mathbb{E}[U_{\text{reg}}]$ and $\sum_{i=1}^{N} \mathbb{E}[U_i]$ are the utilities of the network regulator and all mobile users, respectively.

Here, the utility of the network regulator $\mathbb{E}[U_{\text{reg}}]$ can be mathematically expressed as

$$\begin{aligned}\mathbb{E}[U_{\text{reg}}] &= \sum_{i=1}^{N} C_i^{\text{cont}} + \sum_{i=1}^{N} \beta_i \lambda_i \pi_i^{\text{tx}} - Y(\boldsymbol{\lambda}^{\text{offload}}) \\ &= \left( \sum_{i=1}^{N} \beta_i \lambda_i (\pi_i^{\text{cont}} + \pi_i^{\text{tx}}) \right) - Y(\boldsymbol{\lambda}^{\text{offload}}),\end{aligned} \tag{15}$$

where $\sum_{i=1}^{N} C_i^{\text{cont}}$ and $\sum_{i=1}^{N} \beta_i \lambda_i \pi_i^{\text{tx}}$ are the charges collected from mobile users for their computation offloading and uplink transmissions, respectively; $Y(\boldsymbol{\lambda}^{\text{offload}})$ denotes the network operating cost in providing cloud computation and wireless spectrum access. Following the convention in [43], [44], we assume that $Y(\boldsymbol{\lambda}^{\text{offload}})$ is a known increasing and convex function with respect to the data traffic $\boldsymbol{\lambda}^{\text{offload}}$.

Substituting (15) and (11) into (14) yields

$$\mathcal{SW} = \sum_{i=1}^{N} V_i(\beta_i) - \sum_{i=1}^{N} \beta_i \lambda_i \mathbb{E}[W_i(T_i)] - Y(\boldsymbol{\lambda}^{\text{offload}}). \tag{16}$$

In order to maximize the social welfare $\mathcal{SW}$, the network regulator has to carefully determine the computation offloading scheme $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)^6$, the uplink transmission scheduling discipline $\boldsymbol{\zeta}$ and the pricing rule $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, where $\pi_i = \pi_i^{\text{cont}} + \pi_i^{\text{tx}}, \forall i \in \mathcal{N}$. Note that it is not necessary to differentiate $\pi_i^{\text{cont}}$ and $\pi_i^{\text{tx}}$ for each mobile user $i, \forall i \in \mathcal{N}$, because both of them are service charges paid to the network regulator.

In summary, the problem of designing an efficient mechanism for joint delay-sensitive computation offloading and transmission scheduling can be formulated as

$$[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\pi}] = \arg\max\ \mathcal{SW} \tag{17}$$

$$s.t.,\ 0 \leq \beta_i \leq 1,\ \forall i \in \mathcal{N}, \tag{18}$$

$$\pi_i \geq 0,\ \forall i \in \mathcal{N}, \tag{19}$$

$$\beta_i = \arg\max_{\tilde{\beta}_i} \big( V_i(\tilde{\beta}_i) - \tilde{\beta}_i \lambda_i \pi_i - \tilde{\beta}_i \lambda_i \mathbb{E}[W_i(T_i)] \big), \tag{20}$$

$$V_i(\beta_i) - \beta_i \lambda_i \pi_i - \beta_i \lambda_i \mathbb{E}[W_i(T_i)] \geq 0,\ \forall i \in \mathcal{N}, \tag{21}$$

$$\boldsymbol{\lambda}^{\text{offload}} = (\beta_1 \lambda_1, \beta_2 \lambda_2, \ldots, \beta_N \lambda_N), \tag{22}$$

$$(T_1, T_2, \ldots, T_N) \in \boldsymbol{Q}(\boldsymbol{\lambda}^{\text{offload}}, \boldsymbol{\zeta}), \tag{23}$$

where constraints (18) and (19) show the ranges of each mobile user's computation offloading ratio and service charge, respectively; constraints (20) and (21) indicate the requirements for incentive compatibility and individual rationality, respectively, and are derived by substituting (11) into (12) and (13); constraint (22) illustrates the traffic for computation offloading, which also equals the traffic for uplink transmissions; and constraint (23) states that the experienced delay of each mobile user for offloading a computation task, i.e., $T_i, \forall i \in \mathcal{N}$, is determined by the queueing management system (as depicted in Fig. 1), denoted by $\boldsymbol{Q}(\cdot)$. Obviously, solving this problem directly is very challenging because i) $T_i, \forall i \in \mathcal{N}$, is an endogenous factor of the underlaying queueing system and it depends on the input $\boldsymbol{\lambda}^{\text{offload}}$ and the scheduling discipline $\boldsymbol{\zeta}$ (which are both decision-dependent); ii) $\boldsymbol{\zeta}$ is not a simple decision variable (or vector) but a complicated management policy of the queue; iii) to guarantee both incentive compatibility and individual rationality, $\boldsymbol{\pi}$ may need to be devised as a function with respect to the queueing dynamics. Therefore, in the following, we will propose a novel approach to design the mechanism, i.e., $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\pi}]$, which can meet all these requirements. Note that the payment term $\boldsymbol{\pi}$ is cancelled out in the network social welfare $\mathcal{SW}$ in (16), and thus it is not included in the objective function (17) but only included in constraints (20) and (21). This motivates us to decouple the original problem to a management problem for determining $[\boldsymbol{\beta}, \boldsymbol{\zeta}]$ and a pricing design problem for determining $\boldsymbol{\pi}$, and such decoupling will not affect the network-wide optimality, if we can ensure that the determination of $\boldsymbol{\pi}$ is based on the optimal $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$.

## 4 Multi-User Computation Offloading and Transmission Scheduling Mechanism

In this section, an efficient multi-user mobile computation offloading and transmission scheduling mechanism for delay-sensitive applications in mobile edge computing (named as MOTM) is proposed. We first study the optimal transmission scheduling discipline $\boldsymbol{\zeta}$ for minimizing the

[6] This centrally derived computation offloading scheme can be considered as a suggestion to mobile users. Although the computation offloading ratio is actually determined strategically by each mobile user, we will show that, under the designed mechanism, the individual optimal decision will be exactly the same as the network-wide optimal management scheme.

total delay cost of all mobile users given a fixed computation offloading decision. Then, based on the derived relationship between the transmission scheduling and the computation offloading, we formulate a convex optimization problem to determine the optimal computation offloading scheme $\boldsymbol{\beta}$ for maximizing the network social welfare. After that, an appropriate pricing rule $\boldsymbol{\pi}$ is devised which can regulate the noncooperative game behaviors of mobile users. Finally, we summarize the proposed MOTM and prove its feasibility and optimality.

### 4.1 Delay-Sensitive Transmission Scheduling

From (16), we can observe that the first and the third terms of the network social welfare ($\mathcal{SW}$) only depend on mobile users' computation offloading ratios $\boldsymbol{\beta}$ (or the offloaded traffics) regardless of any other decisions. This implies that these two terms become constants under a given computation offloading scheme. In addition, as explained that the payment terms have been cancelled out in $\mathcal{SW}$, and the design of the pricing rule is mainly for guaranteeing incentive compatibility and individual rationality (i.e., constraints (20) and (21)). Thus, by assuming a given computation offloading scheme, termed as $\tilde{\boldsymbol{\beta}}$, and ignoring constraints (20) and (21), the original problem formulated in (17), (18), (19), (20), (21), (22), and (23) can be relaxed as

$$\begin{aligned}\boldsymbol{\zeta} &= \arg\max \mathcal{SW}\\ &= \arg\max \sum_{i=1}^{N} V_i(\tilde{\beta}_i) - \sum_{i=1}^{N} \tilde{\beta}_i\lambda_i\mathbb{E}[W_i(T_i)] - Y(\tilde{\boldsymbol{\beta}})\\ &= \arg\min \sum_{i=1}^{N} \tilde{\beta}_i\lambda_i\mathbb{E}[W_i(T_i)]\end{aligned} \tag{24}$$

$$s.t.,\ \boldsymbol{\lambda}^{\text{offload}} = (\tilde{\beta}_1\lambda_1, \tilde{\beta}_2\lambda_2, \ldots, \tilde{\beta}_N\lambda_N), \tag{25}$$

$$(T_1, T_2, \ldots, T_N) \in \boldsymbol{Q}(\boldsymbol{\lambda}^{\text{offload}}, \boldsymbol{\zeta}), \tag{26}$$

which turns out to be a pure queueing scheduling problem with the objective of minimizing the total delay cost of mobile users. Obviously, after solving this problem, we can obtain the optimal transmission scheduling discipline $\boldsymbol{\zeta}$ and as well as the minimum delay cost under the given $\tilde{\boldsymbol{\beta}}$.

It is worth noting that, different from traditional delay-sensitive queueing scheduling problems [45], [46] which commonly limited their focuses on linear delay costs, in this paper, we consider a more practical scenario where mobile users are allowed to have heterogeneous and nonlinear delay cost functions, i.e., $W_i(\cdot), \forall i \in \mathcal{N}$. This greatly increases the difficulty in finding the optimal $\boldsymbol{\zeta}$ because the nonlinear delay sensitivities of mobile users cannot be reflected by the simple fixed and static priorities. Moreover, since it is required to derive the delay distributions of the queueing system, i.e., the distribution of $T_i, \forall i \in \mathcal{N}$, to quantify $\mathbb{E}[W_i(T_i)]$, the mean analysis [47], a widely used method under linear delay cost settings, is no longer applicable. In the following, a novel transmission scheduling discipline is thus proposed and analyzed.

Before the analysis, we first introduce some new notations. Denote $1/\mu_i = \mathbb{E}[S_i/R_i]$ as the mean uplink transmission time for one computation task of mobile user $i, \forall i \in \mathcal{N}$, and $\rho_i = \lambda_i^{\text{offload}}/\mu_i = \frac{\tilde{\beta}_i\lambda_i}{\mu_i}$ as its individual offloading traffic intensity (i.e., a ratio). Besides, define $L_i = \rho_i T_i$ and $L = \sum_{i\in\mathcal{N}} L_i$ as the workload of each mobile user $i$ and the total workload waiting for edge cloud computing (or uplink transmissions), respectively. In addition, describe the marginal delay cost of each mobile user $i$ as $W_i'(t) = \frac{\partial W_i(t)}{\partial t}$, where $W_i(\cdot)$ stands for the nonlinear delay cost function, and $t$ is the realization of its random delay $T_i$. Furthermore, let $1/\mu = \frac{\sum_{i\in\mathcal{N}}\mathbb{E}[S_i]}{\sum_{i\in\mathcal{N}}\mathbb{E}[R_i]}$ and $\rho = \sum_{i\in\mathcal{N}} \rho_i$.

We then build the transmission scheduling discipline $\boldsymbol{\zeta}$ according to a delay-dependent dynamic priority ranking rule as follows.

*Transmission Scheduling Discipline $\boldsymbol{\zeta}$:* Denote $a_i(\tau)$ as the amount of tasks that mobile user $i, \forall i \in \mathcal{N}$, has buffered up to the time instant $\tau$. Then, whenever a channel becomes free at time $\tau$, it will be scheduled for serving the uplink transmissions of mobile user $i$ with the highest priority index $\mathcal{I}_i(\tau)$, which is defined as

$$i = \arg\max_{i\in\mathcal{N}} \mathcal{I}_i(\tau) = \arg\max_{i\in\mathcal{N}} W_i'\left(\frac{a_i(\tau)}{\tilde{\beta}_i\lambda_i}\right)\mu_i. \tag{27}$$

*An Illustration Example of $\boldsymbol{\zeta}$:* Consider that there are two mobile users ($MU1$ and $MU2$) and one uplink channel available at time $\tau$. Assume that the offloading task rates of $MU1$ and $MU2$ are 2 tasks/sec. and 3 tasks/sec., respectively. The mean transmission time for one task of $MU1$ and $MU2$ are set as $1/\mu_1 = 0.1$ sec. and $1/\mu_2 = 0.2$ sec., respectively. Define the delay cost functions of $MU1$ and $MU2$ as $W_1(T_1) = T_1^2$ and $W_2(T_2) = 2T_2^2$, so that $W_1'(t) = 2t$ and $W_2'(t) = 4t$. If at time $\tau$, the amount of tasks temporarily stored in the buffers of $MU1$ and $MU2$ are $a_1(\tau) = 5$ and $a_2(\tau) = 3$, then according to (27), the priority indices of $MU1$ and $MU2$ can be calculated as

$$\mathcal{I}_1(\tau) = W_1'(5/2) \times 0.1 = 2 \times (5/2) \times 0.1 = 0.5, \tag{28}$$

$$\mathcal{I}_2(\tau) = W_2'(3/3) \times 0.2 = 4 \times (3/3) \times 0.2 = 0.8. \tag{29}$$

Since $\mathcal{I}_2(\tau) > \mathcal{I}_1(\tau)$, $MU2$ will be scheduled for transmission on the available channel starting from time $\tau$.

In the following theorem, we prove that $\boldsymbol{\zeta}$ is asymptotically optimal in minimizing the total delay cost of mobile users and derive the corresponding delay distributions.

**Theorem 1.** *Given the computation offloading decision $\tilde{\boldsymbol{\beta}}$ and by applying the proposed transmission scheduling discipline $\boldsymbol{\zeta}$,*

(a) *the total delay cost of all mobile users in the system, i.e., $\sum_{i=1}^{N}\tilde{\beta}_i\lambda_i\mathbb{E}[W_i(T_i)]$, can be asymptotically minimized, as the traffic intensity $\rho$ tends to 1;*

(b) *the delay distribution of each individual, i.e., the cumulative distribution function (CDF) of $T_i, \forall i \in \mathcal{N}$, can be approximated as*

$$F_{T_i}(t|\tilde{\boldsymbol{\beta}}) \approx F_L(g_i^{-1}(\rho_i t)|\tilde{\boldsymbol{\beta}}). \tag{30}$$

*Here, $F_L(\cdot)$ is the CDF of the total workload $L$, which can be expressed by Brownian approximation [48] as*

$$F_L(\ell|\tilde{\boldsymbol{\beta}}) \approx 1 - \rho e^{-\gamma\ell} \quad \text{with} \quad \gamma = \frac{\rho}{\mu(1-\rho)} \frac{\kappa_a^2 + \kappa_s^2}{2}, \tag{31}$$

*where $\kappa_a$ and $\kappa_s$ are the coefficients of variation of the offloading task interarrival time and transmission time, respectively. $g_i^{-1}(\cdot)$ denotes the inverse function of $g_i(\cdot)$ which maps the scheduling discipline $\boldsymbol{\zeta}$ to the allocation of the workload over each mobile user $i, \forall i \in \mathcal{N}$, and $\boldsymbol{g}(\cdot) = (g_1(\cdot), g_2(\cdot), \ldots, g_N(\cdot))$ can be obtained by solving the linear-constrained convex minimization problem:*

$$\boldsymbol{g}(L) = \underset{g_i(L)=L_i, \forall i \in \mathcal{N}}{\arg\min} \sum_{i=1}^{N} \tilde{\beta}_i \lambda_i \mathbb{E}[W_i(T_i)] \tag{32}$$

$$s.t., \ \sum\nolimits_{i=1}^{N} L_i = L, \tag{33}$$

$$L_i = \rho_i T_i, \ L_i \geq 0, \ \forall i \in \mathcal{N}. \tag{34}$$

**Proof.** Intuitively, the total workload $L$ of the queueing system is independent of the scheduling discipline, when the traffic intensity $\rho$ tends to 1 (i.e., the queueing system approaches the heavy traffic limit). However, the scheduling discipline does impact how the total workload $L$ is distributed among different mobile users. From [49], we know that as $\rho$ tends to 1, the ratio between the workload of each mobile user $i$ and the total workload, denoted by $\frac{L_i}{L}, \forall i \in \mathcal{N}$, becomes a constant.

Thus, the queueing scheduling problem for determining $\boldsymbol{\zeta}$ shown in (24), (25), and (26) can be transformed to the problem of allocating $L$ over mobile users, i.e., the determination of $\boldsymbol{g}(L) = (g_1(L), g_2(L), \ldots, g_N(L))$, with the aim of minimizing their total delay cost. This transformed problem can be formulated as illustrated in (32), (33), and (34). Since the objective function is convex (because $W_i(\cdot), \forall i \in \mathcal{N}$, is defined as a convex function) and all constraints are linear, this problem can be easily solved by checking the first-order conditions:

$$\frac{\tilde{\beta}_i \lambda_i}{\rho_i} W_i'\left(\frac{L_i}{\rho_i}\right) - \xi_i = \frac{\tilde{\beta}_j \lambda_j}{\rho_j} W_j'\left(\frac{L_j}{\rho_j}\right) - \xi_j, \ \forall i, j \in \mathcal{N}, \tag{35}$$

with $L_i \xi_i = 0$ and $\sum_{i=1}^{N} L_i = L$, where $\xi_i \geq 0$ is the Lagrange multiplier with respect to the non-negativity constraint on $L_i, \forall i \in \mathcal{N}$.

Furthermore, the instantaneous workload of each mobile user at any time $\tau$ can be computed as

$$L_i(\tau) = \frac{a_i(\tau)}{\mu_i}, \ \forall i \in \mathcal{N}, \forall \tau, \tag{36}$$

where $a_i(\tau)$ is the amount of tasks that have been stored in the buffer of mobile user $i$ up to time $\tau$. Then, we have

$$\frac{L_i(\tau)}{\rho_i} = \frac{a_i(\tau)}{\mu_i} \frac{\mu_i}{\tilde{\beta}_i \lambda_i} = \frac{a_i(\tau)}{\tilde{\beta}_i \lambda_i}, \ \forall i \in \mathcal{N}, \forall \tau. \tag{37}$$

Substituting (37) into (35), we can observe that the proposed scheduling discipline $\boldsymbol{\zeta}$ based on (27) actually tries to distribute the workloads of mobile users to satisfy the first-order conditions (35) at any time $\tau$. Therefore, $\boldsymbol{\zeta}$ is asymptotically optimal in minimizing the total delay cost of mobile users, and Theorem 1 (a) is proved.

With the asymptotically optimal solution of $\boldsymbol{g}(L) = (g_1(L), g_2(L), \ldots, g_N(L))$, we can derive the CDF of each mobile user's delay $T_i, \forall i \in \mathcal{N}$, as

$$\begin{aligned} F_{T_i}(t|\tilde{\boldsymbol{\beta}}) &= \mathit{Prob.}(T_i \leq t|\tilde{\boldsymbol{\beta}}) \\ &= \mathit{Prob.}(L_i \leq \rho_i t|\tilde{\boldsymbol{\beta}}) \\ &\approx \mathit{Prob.}(g_i(L) \leq \rho_i t|\tilde{\boldsymbol{\beta}}) \\ &= F_L(g_i^{-1}(\rho_i t)|\tilde{\boldsymbol{\beta}}), \end{aligned} \tag{38}$$

where $F_L(\ell|\tilde{\boldsymbol{\beta}})$ is the CDF of the total workload $L$ and can be directly expressed by Brownian approximation [48], as shown in (31). This completes the proof of Theorem 1 (b). □

Theorem 1 indicates that the proposed transmission scheduling discipline $\boldsymbol{\zeta}$ can asymptotically minimize the total delay cost of all mobile users, for any given computation offloading decision $\tilde{\boldsymbol{\beta}}$. Although the optimality of $\boldsymbol{\zeta}$ relies on the heavy traffic approximation (i.e., $\rho$ tends to 1), since it is widely known that the ever-increasing mobile data traffic is rapidly straining the capacity of existing cellular networks [6], it can be expected that $\boldsymbol{\zeta}$ will actually achieve a good performance under practical wireless settings.

Moreover, with the CDF of $T_i, \forall i \in \mathcal{N}$, as derived in Theorem 1 (b), the minimum total delay cost of all mobile users can be calculated as

$$\min \sum_{i=1}^{N} \tilde{\beta}_i \lambda_i \mathbb{E}[W_i(T_i)] = \sum_{i=1}^{N} \tilde{\beta}_i \lambda_i \int_{-\infty}^{\infty} W_i(t) dF_{T_i}(t|\tilde{\boldsymbol{\beta}}). \tag{39}$$

Note that (39) specifies a closed-form relationship between the computation offloading decision $\tilde{\boldsymbol{\beta}}$ and the minimized total delay cost of all mobile users under the proposed transmission scheduling discipline $\boldsymbol{\zeta}$. This relationship will be used in Section 4.2 to determine the optimal computation offloading scheme.

### 4.2 Optimal Computation Offloading Scheme

With the (asymptotically) optimal transmission scheduling discipline $\boldsymbol{\zeta}$ devised in Section 4.1, we are now able to determine the optimal computation offloading scheme $\boldsymbol{\beta}$ by solving

$$\begin{aligned} \boldsymbol{\beta} &= \arg\max \mathcal{SW} \\ &= \arg\max \sum_{i=1}^{N} V_i(\beta_i) - \sum_{i=1}^{N} \beta_i \lambda_i \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)] - Y(\boldsymbol{\beta}) \end{aligned} \tag{40}$$

$$s.t., \ 0 \leq \beta_i \leq 1, \ \forall i \in \mathcal{N}, \tag{41}$$

where $\sum_{i=1}^{N} \beta_i \lambda_i \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]$ denotes the minimized total delay cost of mobile users resulted by the optimal transmission scheduling discipline $\boldsymbol{\zeta}$ with respect to $\boldsymbol{\beta}$. Obviously, this problem (i.e., (40) and (41)) is a subproblem of the original one formulated in (17), (18), (19), (20), (21), (22), and (23) without taking into account the pricing rule and individual utilities of mobile users. In the following, we present some important characteristics of this problem and the corresponding solution.

**Theorem 2.** *There always exists a unique optimal solution of $\boldsymbol{\beta}$ which maximizes the network social welfare $\mathcal{SW}$.*

**Proof.** From (41), we can obviously see that the feasible set of the optimal computation offloading ratios is nonempty and compact. Moreover, i) $V_i(\beta_i)$ is linear and increasing with respect to $\beta_i, \forall i \in \mathcal{N}$, as shown in (10); ii) $\sum_{i=1}^{N} \beta_i \lambda_i \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]$ is convex and increasing with respect to $\beta_i, \forall i \in \mathcal{N}$, because queueing delays are naturally convex and increasing with respect to the traffic load, and the waiting cost $W_i(\cdot)$ is defined as an increasing convex function with respect to $\beta_i, \forall i \in \mathcal{N}$; and iii) $Y(\boldsymbol{\beta})$ is ordinarily convex and increasing with respect to $\beta_i, \forall i \in \mathcal{N}$, as explained in Section 3.3. All these together imply that the objective function (40) is concave with respect to the offloading ratios $\beta_i, \forall i \in \mathcal{N}$. Therefore, the problem formulated in (40) and (41), is a concave maximization problem with a nonempty and compact feasible set, so that there always exists a unique optimal solution of $\boldsymbol{\beta}$. □

**Theorem 3.** *Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_N^*)$ be the network-wide optimal computation offloading scheme, i.e., $\boldsymbol{\beta}^*$ maximizes the network social welfare $\mathcal{SW}$, then we must have*

$$\frac{\partial V_i(\beta_i)}{\partial \beta_i} = \lambda_i \mathbb{E}_{\boldsymbol{\beta}^*}^{\boldsymbol{\zeta}}[W_i(T_i)] + \sum_{j=1}^{N} \beta_j^* \lambda_j \frac{\partial \mathbb{E}_{\boldsymbol{\beta}^*}^{\boldsymbol{\zeta}}[W_j(T_j)]}{\partial \beta_i} + \frac{\partial Y(\boldsymbol{\beta}^*)}{\partial \beta_i} + \omega_i - \nu_i, \ \forall i \in \mathcal{N}, \tag{42}$$

$$\text{and} \quad 0 \leq \beta_i^* \leq 1, \ \forall i \in \mathcal{N}, \tag{43}$$

$$\omega_i \beta_i^* = 0, \ \forall i \in \mathcal{N}, \tag{44}$$

$$\nu_i (1 - \beta_i^*) = 0, \ \forall i \in \mathcal{N}, \tag{45}$$

*where $\omega_i \geq 0$ and $\nu_i \geq 0$ are Lagrange multipliers with respect to constraints $\beta_i \geq 0$ and $\beta_i \leq 1$, respectively.*

**Proof.** It has been proved in Thoerem 2 that the social welfare maximization problem formulated in (40) and (41), is a standard concave maximization problem. Thus, by applying Karush-Kuhn-Tucker (KKT) optimality conditions [50] along with some simple mathematical manipulations, we can obtain (42), (43), (44), and (45). □

Theorem 2 indicates the existence and uniqueness of the optimal solution for (40) and (41), and hence existing software-based optimization tools [51] can be directly employed to numerically determine the optimal computation offloading scheme. Besides, Theorem 3 illustrates the necessary and sufficient optimality conditions which will be used in Section 4.3 to theoretically verify the network-wide optimality of the designed pricing rule in regulating the strategic and noncooperative behaviors of mobile users.

### 4.3 Design of the Pricing Rule

After deriving the optimal joint computation offloading and transmission scheduling scheme, denoted by $[\boldsymbol{\beta}, \boldsymbol{\zeta}]$, as shown in Sections 4.1 and 4.2, our remaining problem is to determine the pricing rule $\boldsymbol{\pi}$ based on the optimal $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ to meet incentive compatibility and individual rationality (i.e., constraints (20) and (21)). Since the network optimum (i.e., the maximization of the network social welfare $\mathcal{SW}$) is not affected by the designed pricing rule, the main purpose of designing $\boldsymbol{\pi}$ is to induce all smart mobile users to follow the determined $[\boldsymbol{\beta}, \boldsymbol{\zeta}]$.

Naturally, as rational and selfish entities, mobile users may deviate from the management decisions made by the network regulator, if and only if they can benefit from such behaviors. Particularly, each mobile user $i, \forall i \in \mathcal{N}$, can strategically and independently determine its self-preferred computation offloading ratio $\beta_i \in [0, 1]$ with the objective of maximizing its own expected utility, which has been defined in Section 3 and is rewritten here as

$$U_i(\beta_i) = V_i(\beta_i) - \beta_i \lambda_i \pi_i - \beta_i \lambda_i \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)], \tag{46}$$

where $V_i(\beta_i)$ is the value gained by mobile user $i$ from computation offloading, calculated by (10); $\pi_i$ is the undetermined service charge on mobile user $i$ for each offloaded computation task; and $\mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]$ stands for the average delay cost of one computation task of mobile user $i$ when the optimal uplink transmission scheduling discipline $\boldsymbol{\zeta}$ is adopted. Note that, since the transmission scheduling is executed centrally by the network regulator, mobile users are not able to manipulate $\boldsymbol{\zeta}$, even though they are strategic[7].

To maximize the individual utility $U_i(\beta_i), \forall i \in \mathcal{N}$, smart mobile users will compete with each other by adjusting their own computation offloading ratios, i.e., $\beta_i, \forall i \in \mathcal{N}$, so as to possibly increase their valuation gains, while at the same time lower their service charges and waiting delay costs from computation offloading and uplink transmissions. This will obviously result in a distributed decision making problem, or more specifically, a noncooperative game on a queueing system with discipline $\boldsymbol{\zeta}$. Formally, this game can be represented as

$$\mathcal{G} \triangleq \{\mathcal{N}, \mathcal{B}, \{U_i(\beta_i, \boldsymbol{\beta}_{-i})\}_{i \in \mathcal{N}}\}, \tag{47}$$

where mobile users in the set $\mathcal{N}$ act as players; $\mathcal{B}$ denotes the strategy set of mobile users' computation offloading ratios $\beta_i, \forall i \in \mathcal{N}$; and $U_i(\beta_i, \boldsymbol{\beta}_{-i})$ is the utility function of mobile user $i, \forall i \in \mathcal{N}$, in terms of its strategy $\beta_i$ given that the strategy of all other mobile users is $\boldsymbol{\beta}_{-i} = \boldsymbol{\beta} \backslash \beta_i$. The Nash equilibrium (NE) of game $\mathcal{G}$ can be defined as follows.

**Definition 1 (NE).** *A strategy profile $\boldsymbol{\beta}^e = (\beta_1^e, \beta_2^e \ldots, \beta_N^e)$ is a Nash equilibrium of game $\mathcal{G}$ if for every mobile user $i, \forall i \in \mathcal{N}$, we have*

$$U_i(\beta_i^e, \boldsymbol{\beta}_{-i}^e) \geq U_i(\beta_i, \boldsymbol{\beta}_{-i}^e), \ \forall \beta_i \in [0, 1]. \tag{48}$$

It is obvious that the NE of game $\mathcal{G}$ largely depends on the pricing rule $\boldsymbol{\pi}$. Therefore, in order to guarantee that all mobile users will be willing to self-manage their computation offloading according to the network-wide optimal scheme so that both their own utilities and the social welfare $\mathcal{SW}$ can be maximized, it is clear that $\boldsymbol{\pi}$ should be designed

[7] In this paper, we do not consider malicious attacks and assume that all mobile users are trustworthy.

with the goal of making the network-wide optimal solution be an NE of game $\mathcal{G}$, i.e.,

$$\beta_i^* = \beta_i^e, \ \forall i \in \mathcal{N}, \tag{49}$$

where $\beta_i^e$ and $\beta_i^*$ are the outputs (i.e., computation offloading ratio on mobile user $i, \forall i \in \mathcal{N}$) of the NE of game $\mathcal{G}$ and the network-wide optimal computation offloading problem (40) and (41), respectively.

*Designed Pricing Rule* $\boldsymbol{\pi}^*$*: Let* $\boldsymbol{\beta}^*$ *denote the network-wide optimal computation offloading scheme (as devised in Section 4.2), then the optimal pricing rule* $\boldsymbol{\pi}^* = (\pi_1^*, \pi_2^*, \ldots, \pi_N^*)$ *for mobile users in computation offloading and transmission scheduling is defined as*

$$\pi_i^*(\beta_i) = \frac{1}{\lambda_i}\alpha_i + \frac{1}{\beta_i\lambda_i}\theta_i, \ \forall i \in \mathcal{N}, \tag{50}$$

*where*

$$\alpha_i = \frac{\partial Y(\boldsymbol{\beta}^*)}{\partial \beta_i} + \sum_{j=1, j\neq i}^{N} \beta_j^* \lambda_j \frac{\partial \mathbb{E}_{\boldsymbol{\beta}^*}^{\boldsymbol{\zeta}}[W_j(T_j)]}{\partial \beta_i}, \tag{51}$$

$$\theta_i \leq \bar{\theta}_i = V(\beta_i^*) - \beta_i^* \lambda_i \mathbb{E}_{\boldsymbol{\beta}^*}^{\boldsymbol{\zeta}}[W_i(T_i)] - \alpha_i \beta_i^*. \tag{52}$$

*Note that* $\alpha_i$ *and* $\theta_i$ *are coefficients that are independent of* $\beta_i$ *and determined by* $\boldsymbol{\beta}^*$ *only.*

Next, we prove that, by employing the designed pricing rule $\boldsymbol{\pi}^*$, the NE of game $\mathcal{G}$ always exists and its equilibrium conditions are satisfied by the network-wide optimal computation offloading scheme $\boldsymbol{\beta}^*$.

**Theorem 4.** *The game* $\mathcal{G} \triangleq \{\mathcal{N}, \mathcal{B}, \{U_i(\beta_i, \boldsymbol{\beta}_{-i})\}_{i\in\mathcal{N}}\}$ *with the designed pricing rule, i.e.,* $\boldsymbol{\pi} \triangleq \boldsymbol{\pi}^*$*, has at least one NE.*

**Proof.** Since the strategies of mobile users are their self-determined computation offloading ratios which must be within the range of $[0, 1]$, the strategy set $\mathcal{B}$ of game $\mathcal{G}$ can be expressed in the form of Cartesian product as

$$\mathcal{B} = \prod_{i=1}^{N} [0, 1] \subset \mathbb{R}^N, \tag{53}$$

which is obviously nonempty, convex and compact.

Substituting the pricing rule (50), (51), and (52) into the utility function (46) of mobile user $i, \forall i \in \mathcal{N}$, and taking the first-order derivative with respect to $\beta_i$, we have

$$\begin{aligned} \frac{\partial U_i}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i}\Big(V_i(\beta_i) - \beta_i\lambda_i \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)] - \beta_i\alpha_i - \theta_i\Big) \\ &= \frac{\partial V_i(\beta_i)}{\partial \beta_i} - \lambda_i \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)] - \beta_i\lambda_i \frac{\partial \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]}{\partial \beta_i} - \alpha_i. \end{aligned} \tag{54}$$

As shown in (10), $V_i(\beta_i)$ is linear with $\beta_i$. Thus, the second-order derivative of $U_i, \forall i \in \mathcal{N}$, can be calculated based on (54) as

$$\frac{\partial^2 U_i}{\partial \beta_i^2} = -2\lambda_i \frac{\partial \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]}{\partial \beta_i} - \beta_i\lambda_i \frac{\partial^2 \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]}{\partial \beta_i^2}. \tag{55}$$

Recall that the waiting cost $W_i(\cdot)$ is defined as an increasing convex function, and the queueing delay $T_i$ is naturally increasing and convex with respect to $\beta_i$, so that

$$\frac{\partial \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]}{\partial \beta_i} \geq 0, \ \frac{\partial^2 \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)]}{\partial \beta_i^2} \geq 0. \tag{56}$$

Substituting (56) into (55) yields $\frac{\partial^2 U_i}{\partial \beta_i^2} \leq 0$, which means that $U_i$ is concave with respect to $\beta_i$.

In conclusion, since the strategy set is nonempty, convex and compact, and the utility functions are continuous and concave, there exists at least one NE in game $\mathcal{G}$. □

**Theorem 5.** *With the adoption of the designed pricing rule, i.e.,* $\boldsymbol{\pi} \triangleq \boldsymbol{\pi}^*$*, the network-wide optimal computation offloading scheme* $\boldsymbol{\beta}^*$ *is an NE of game* $\mathcal{G}$*, namely*

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}^e. \tag{57}$$

**Proof.** Since the NE of game $\mathcal{G}$ exists (as proved in Theorem 4), the strategy of each mobile user $i, \forall i \in \mathcal{N}$ will converge to the equilibrium $\beta_i^e$, such that its individual utility is maximized, i.e.,

$$\begin{aligned} \beta_i^e &= \arg\max U_i \\ &= \arg\max V_i(\beta_i) - \beta_i\lambda_i\pi_i - \beta_i\lambda_i \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\zeta}}[W_i(T_i)] \end{aligned} \tag{58}$$

$$s.t., \ 0 \leq \beta_i \leq 1, \tag{59}$$

Substituting the pricing rule (50), (51), and (52) into (58) and applying the KKT conditions, we can show that $\beta_i^e, \forall i \in \mathcal{N}$, must satisfy

$$\begin{aligned} \frac{\partial V_i(\beta_i)}{\partial \beta_i} &= \lambda_i \mathbb{E}_{\boldsymbol{\beta}^e}^{\boldsymbol{\zeta}}[W_i(T_i)] + \beta_i^e\lambda_i \frac{\partial \mathbb{E}_{\boldsymbol{\beta}^e}^{\boldsymbol{\zeta}}[W_i(T_i)]}{\partial \beta_i} \\ &\quad + \frac{\partial}{\partial \beta_i}\left(\beta_i^e\lambda_i\left(\frac{1}{\lambda_i}\alpha_i + \frac{1}{\beta_i^e\lambda_i}\theta_i\right)\right) + \omega_i - \nu_i \\ &= \lambda_i \mathbb{E}_{\boldsymbol{\beta}^e}^{\boldsymbol{\zeta}}[W_i(T_i)] + \sum_{j=1}^{N} \beta_j^e\lambda_j \frac{\partial \mathbb{E}_{\boldsymbol{\beta}^e}^{\boldsymbol{\zeta}}[W_j(T_j)]}{\partial \beta_i} \\ &\quad + \frac{\partial Y(\boldsymbol{\beta}^e)}{\partial \beta_i} + \omega_i - \nu_i, \end{aligned} \tag{60}$$

$$\text{and} \quad 0 \leq \beta_i^e \leq 1, \tag{61}$$

$$\omega_i \beta_i^e = 0, \tag{62}$$

$$\nu_i(1 - \beta_i^e) = 0, \tag{63}$$

where $\omega_i \geq 0$ and $\nu_i \geq 0$ are Lagrange multipliers.

Obviously, the equilibrium conditions (60), (61), (62), and (63) for each mobile user $i, \forall i \in \mathcal{N}$, are exactly the same as the optimality conditions (42), (43), (44), and (45) for deriving $\boldsymbol{\beta}^*$. This means that $\boldsymbol{\beta}^*$, which satisfies (42), (43), (44), and (45), must also satisfy (60), (61), (62), and (63), $\forall i \in \mathcal{N}$, and thus we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}^e$. □

Theorems 4 and 5 indicate that the NE of game $\mathcal{G}$ can be obtained by directly solving the optimization problem defined in (40) and (41), rather than running the distributed decision making process among mobile users. In addition, it is not necessary to further prove the uniqueness of NE or guarantee a fast convergence speed of game $\mathcal{G}$, because in practice, the network regulator can always centrally determine $\boldsymbol{\beta}^*$ in advance and then induce all mobile users to follow it by adopting the designed pricing rule $\boldsymbol{\pi}$, or in other words, no iterative strategy adaption is needed.

### 4.4 Summary of MOTM

In summary, the proposed multi-user mobile computation offloading and transmission scheduling mechanism (MOTM) consists of the transmission scheduling discipline $\boldsymbol{\zeta}$ built based on the delay-dependent dynamic priority order (27), the computation offloading scheme $\boldsymbol{\beta}$ derived from the optimization problem (40) and (41) and the pricing rule $\boldsymbol{\pi} \triangleq \boldsymbol{\pi}^*$ designed in (50), (51), and (52) to make $\boldsymbol{\beta} \triangleq \boldsymbol{\beta}^* = \boldsymbol{\beta}^e$. In the following, we show that MOTM can indeed meet all requirements or constraints imposed by the original problem formulated in Section 3.3, namely, *incentive compatibility*, *individual rationality* and *optimality (i.e., network social welfare maximization)*.

**Theorem 6 (Incentive compatibility).** *All mobile users will be willing to manage their computation offloading according to the proposed MOTM, denoted by $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\pi}]$, because doing so can maximize their own utilities, i.e.,*

$$\beta_i = \arg\max_{\tilde{\beta}_i \in [0,1]} \Big( V_i(\tilde{\beta}_i) - \tilde{\beta}_i \lambda_i \pi_i - \tilde{\beta}_i \lambda_i \mathbb{E}^{\boldsymbol{\zeta}}_{\boldsymbol{\beta}}[W_i(T_i)] \Big). \tag{64}$$

**Proof.** It has already been proved in Theorem 5 that by determining the computation offloading scheme as $\boldsymbol{\beta} \triangleq \boldsymbol{\beta}^*$ along with the optimal transmission scheduling discipline $\boldsymbol{\zeta}$ and the correspondingly designed pricing rule $\boldsymbol{\pi}$, the NE conditions for the distributed strategy making process among mobile users can be satisfied, leading to $\boldsymbol{\beta} \triangleq \boldsymbol{\beta}^* = \boldsymbol{\beta}^e$. Since NE is ordinarily defined as the equilibrium such that all mobile users can maximize their own utilities, MOTM, which reaches the NE, obviously guarantees (64), and thus is incentive-compatible. □

**Theorem 7 (Individual rationality).** *With the implementation of the proposed MOTM, denoted by $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\pi}]$, the expected utility of each mobile user $i, \forall i \in \mathcal{N}$, is always non-negative, i.e.,*

$$\mathbb{E}[U_i] = V_i(\beta_i) - \beta_i \lambda_i \pi_i - \beta_i \lambda_i \mathbb{E}^{\boldsymbol{\zeta}}_{\boldsymbol{\beta}}[W_i(T_i)] \geq 0. \tag{65}$$

**Proof.** Substituting the designed pricing rule $\boldsymbol{\pi}$, i.e., (50), (51), and (52), into the expected utility of each mobile user $i, \forall i \in \mathcal{N}$, we have

$$\begin{aligned} \mathbb{E}[U_i] &= V_i(\beta_i) - \beta_i \lambda_i \mathbb{E}^{\boldsymbol{\zeta}}_{\boldsymbol{\beta}}[W_i(T_i)] - \beta_i \lambda_i \pi_i \\ &= V_i(\beta_i) - \beta_i \lambda_i \mathbb{E}^{\boldsymbol{\zeta}}_{\boldsymbol{\beta}}[W_i(T_i)] - \beta_i \alpha_i - \theta_i \\ &\geq V_i(\beta_i) - \beta_i \lambda_i \mathbb{E}^{\boldsymbol{\zeta}}_{\boldsymbol{\beta}}[W_i(T_i)] - \beta_i \alpha_i - \bar{\theta}_i \\ &= V_i(\beta_i) - \beta_i \lambda_i \mathbb{E}^{\boldsymbol{\zeta}}_{\boldsymbol{\beta}}[W_i(T_i)] - V_i(\beta_i) + \beta_i \lambda_i \mathbb{E}^{\boldsymbol{\zeta}}_{\boldsymbol{\beta}}[W_i(T_i)] \\ &= 0, \end{aligned} \tag{66}$$

which completes the proof. □

**Theorem 8 (Optimality).** *The proposed MOTM, denoted by $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\pi}]$, can maximize the network social welfare $\mathcal{SW}$ in the joint management of delay-sensitive computation offloading and transmission scheduling.*

**Proof.** The reasons why the social welfare $\mathcal{SW}$ can be maximized by the proposed mechanism MOTM, denoted by $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\pi}]$, are twofold: i) the computation offloading and transmission scheduling decisions, i.e., $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$, have been jointly optimized with the objective of maximizing the social welfare $\mathcal{SW}$, as shown in Sections 4.1 and 4.2; and ii) it has been proved in Theorems 6 and 7 that all mobile users are willing to follow the network-wide optimal management decisions $[\boldsymbol{\beta}, \boldsymbol{\zeta}]$ under the specifically designed pricing rule $\boldsymbol{\pi}$. Therefore, the proposed mechanism is robust and the network optimality (i.e., the maximization of network social welfare $\mathcal{SW}$) determined by $[\boldsymbol{\beta}, \boldsymbol{\zeta}]$ can be maintained. □

## 5 Simulation Results

In this section, simulations are conducted to evaluate the performance of the proposed mechanism in mobile computation offloading and transmission scheduling for delay-sensitive applications. All results are obtained by using Monte Carlo simulations over 100 runs with various system parameters.

### 5.1 Simulation Settings

Consider a MATLAB-based simulation environment for a mobile edge computing system with $N$ mobile users running delay-sensitive computational applications and $M$ channels dedicated for computation offloading, where $N$ and $M$ are in ranges from 50 to 100 and 5 to 15, respectively. According to the 4G cellular network characteristics [52], the transmission power of each mobile user is set as 100 mW and the uplink transmission rate is determined randomly from 2 to 5 Mbps. Based on the configurations of mobile edge computing assisted video gaming applications [12], [53], the average values of the packet size and required CPU cycles of a computation task are approximated as 500 Kb and 1000 Megacycles, respectively. For different mobile users, their computation tasks may be heterogeneous, and thus we further allow the actual values of the packet size and required CPU cycles of each computation task from each mobile user to be taken randomly within $[400, 600]$ Kb and $[800, 1200]$ Megacycles, respectively. In addition, assume that the arrival rate of computation tasks at each mobile user $i, \forall i \in \mathcal{N}$, is selected within $[2, 5]$ per minutes and define the delay cost function as $W_i(T_i) = \epsilon_i T_i^2$, where $\epsilon_i$ reflects the marginal delay sensitivity of each mobile user $i$ and is chosen randomly over $[1, 6]$. Besides, for simplicity, let the mean valuation of one computation task be 1, the network operation cost be 0, and the energy cost coefficients $\varphi_i = \psi_i = 1$. Similar settings have also been employed in [7], [8], [54]. Note that some parameters may be varied for different evaluation purposes.

### 5.2 Performance Evaluations

To show the performance of our proposed transmission scheduling discipline $\boldsymbol{\zeta}$ in minimizing the total delay cost of all mobile users, Fig. 2 compares it with two existing scheduling disciplines, i.e., the first-come first-serve (FCFS)

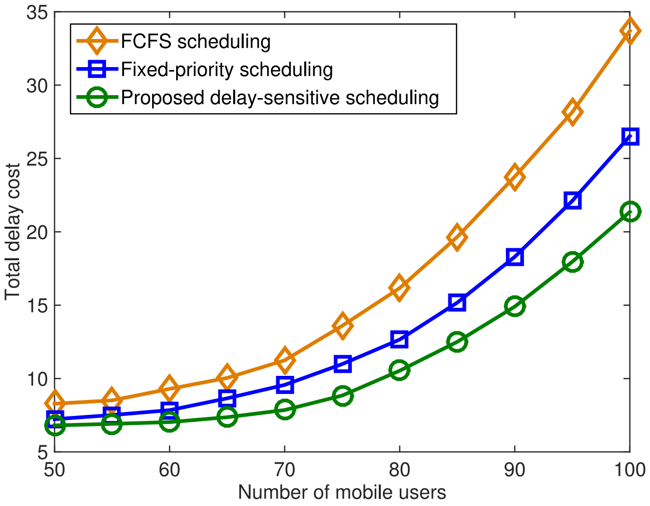


Fig. 2. Performance comparison of the different scheduling disciplines.

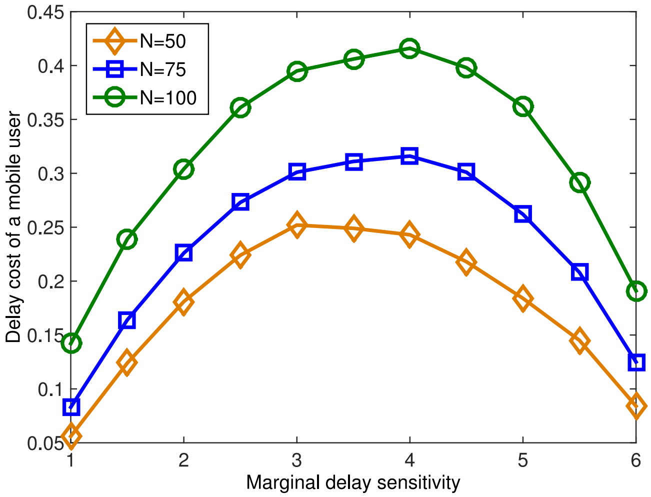


Fig. 3. Delay cost of a mobile user in MOTM.

scheduling [55] which treats all uplink transmissions equally without considering their heterogeneous delay sensitivities, and the fixed-priority scheduling [33] which maintains a fixed transmission order based on mobile users' marginal delay sensitivities but regardless of their experienced delays in computation offloading. It is intuitive that the total waiting costs increase with the number of mobile users for all three scheduling disciplines. This is because when the number of mobile users is larger, the queueing system is more congested so that waiting delays become longer. Besides, the FCFS scheduling leads to the highest delay costs due to ignoring potential priorities among mobile users indicated by their different delay sensitivities. The fixed-priority scheduling outperforms FCFS because of the consideration of mobile users' marginal delay sensitivities. More importantly, it is shown that the proposed delay-sensitive transmission scheduling discipline $\zeta$ achieves the best performance in minimizing the total delay cost. This is because the proposed scheduling discipline $\zeta$ well balance the waiting costs of all mobile users in computation offloading by taking into account both their marginal delay sensitivities and experienced delays.

Fig. 3 investigates the delay cost of a mobile user with different marginal delay sensitivities, when the proposed mechanism, MOTM, is applied. It can be observed that the mobile user's delay cost first increases and then decreases with the increase of its marginal delay sensitivity. According to the definition, the delay cost of a mobile user is a function of both its marginal delay sensitivity and experienced delay. Thus, the waiting cost is considerably small for two extreme cases: i) when the marginal delay sensitivity tends to zero, i.e., the mobile user is running an application which does not care about any delay so that no delay cost is introduced; ii) when the marginal delay sensitivity is extremely large, i.e., the mobile user is highly sensitive to the delay so that its computation tasks will be mostly processed locally (to avoid potential transmission and scheduling delays) or offloaded with the highest priority via the uplink transmission. Furthermore, since the queueing system is more congested with a larger number of mobile users $N$, each mobile user will experience a longer delay and hence suffer a higher delay cost when $N$ increases, as shown in Fig. 3.

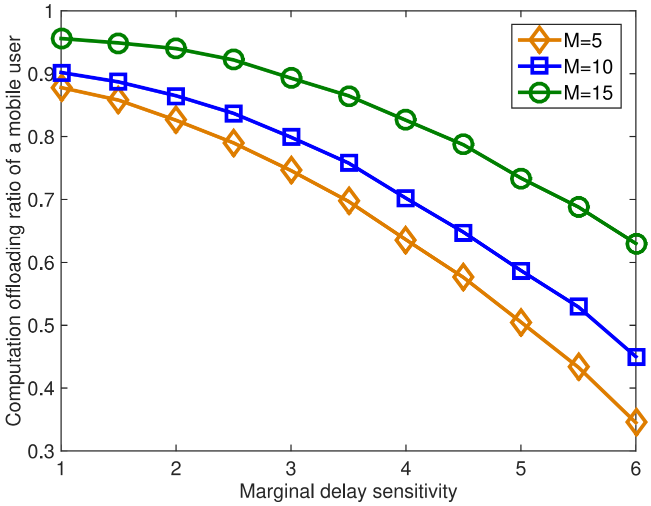


Fig. 4. Computation offloading ratio of a mobile user in MOTM.

Fig. 4 illustrates the computation offloading ratio of a mobile user with different marginal delay sensitivities in the proposed MOTM. It can be seen from this figure that a mobile user's computation offloading ratio decreases with the increase of its marginal delay sensitivity. Naturally, mobile edge computing introduces a longer delay than local processing due to the wireless transmissions between the mobile user and the edge cloud. Thus, the mobile user with a higher delay sensitivity would prefer to process more computation tasks locally and correspondingly decrease its computation offloading ratio. Besides, we can see that such decreasing trend is more obvious when the marginal delay sensitivity gets larger. This is because to maximize the individual utility, the proposed MOTM can well balance the local energy consumption and the offloading delay cost of each mobile user. As a result, when the delay cost of a mobile user becomes dominant (as its marginal delay sensitivity increases), its computation offloading ratio decreases significantly. In addition, since it is intuitive that a larger number of channels (i.e., a larger $M$) always leads to a better scheduling performance in computation offloading (i.e., a shorter delay), Fig. 5 also shows that given the

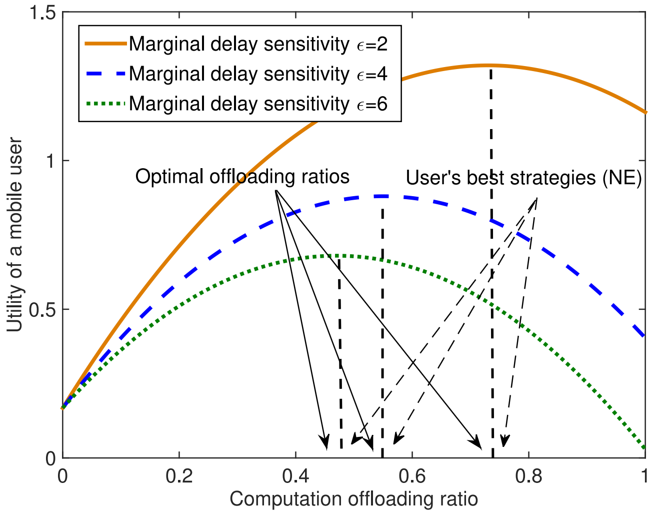


Fig. 5. Optimal offloading strategies of mobile users in MOTM.

same marginal delay sensitivity, the offloading ratio of a mobile user increases with $M$.

Fig. 5 examines the incentive compatibility of the proposed MOTM by showing the utility of a mobile user with different computation offloading ratios. In the considered system, each mobile user may strategically deviate the network-wide optimal management scheme by choosing its own computation offloading ratio for potentially increasing its individual utility. The trend of curves in Fig. 5 indicates that the utility of a mobile user first increases with the computation offloading ratio. This is because increasing the offloading ratio can lower the local processing cost and exploit more benefits from edge computing so that a higher utility can be obtained. However, after a certain point (i.e., the offloading ratio determined by the optimal management scheme as presented in Fig. 4), since the cost of computation offloading (including the induced delay cost and service charge) becomes dominant, the utility decreases. Obviously, the computation offloading ratio which can maximize the individual utility is the best strategy (i.e., NE) that will be adopted by the mobile user. Therefore, Fig. 5 reveals that the network-wide optimal solution also reaches the NE of each mobile user, which proves Theorem 6.

In Figs. 6 and 7, the superiority of applying the proposed MOTM in multi-user delay-sensitive mobile edge computing system is demonstrated. For comparison purpose, the local computing scheme without enabling edge computing capability (local computing), the decentralized game based computation offloading scheme (DCO) [7] and the energy-efficient centralized computation offloading scheme (ECO) [10] are simulated as benchmarks. DCO manages the computation offloading and uplink wireless transmissions by running a distributed decision making process among mobile users, while ECO centrally optimizes the joint resource allocations in mobile edge computing by taking into account the heterogeneities of users' priorities, local processing costs and channel conditions. However, both of them rely on the assumption of quasi-static network scenarios, so that their management decisions are made myopically by ignoring the impacts of random task arrivals and buffering delays.

Fig. 6 compares the network social welfare (i.e., $\mathcal{SW}$) produced by different computation offloading and transmission scheduling management schemes with respect to the number of mobile users. Intuitively, $\mathcal{SW}$ increases with the number of mobile users because more users implies more computation tasks and higher total valuations of task processing (including both local and edge computing). Besides, it is shown that local computing results in the lowest $\mathcal{SW}$. This is because without the help of edge computing, mobile users have to suffer tremendously high energy consumptions in utilizing local CPU resources. In contrast, due to the employment of edge computing, $\mathcal{SW}$ is obviously higher for all other schemes (i.e., DCO, ECO and the proposed MOTM), in which computation offloading is enabled. Furthermore, ECO outperforms DCO because of the consideration of global optimality in the centralized management and the presumption that mobile users will never behave selfishly and strategically. Moreover, we can see that the proposed MOTM achieves the best performance in Fig. 6. This is because MOTM can maximize $\mathcal{SW}$ by considering the long-term tradeoffs of the system in a stochastic network scenario, and can also guarantee that no individual mobile user has the incentive to deviate the optimal solution. We further compare these four management schemes with respect to the number of channels in Fig. 7, and similar observations as in Fig. 6 can

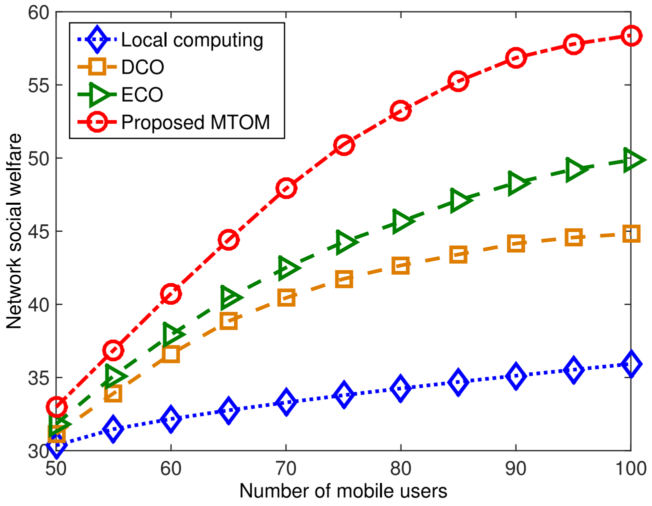


Fig. 6. Comparison on $\mathcal{SW}$ with different number of mobile users.

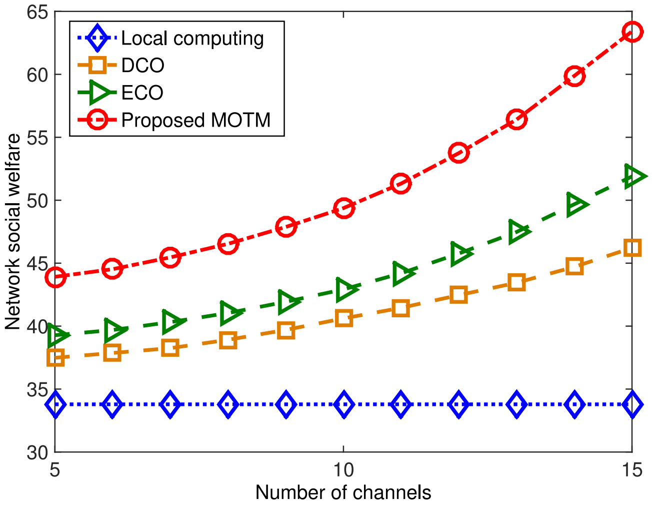


Fig. 7. Comparison on $\mathcal{SW}$ with different number of channels.

be obtained. It is worth noting that since more channels indicates a better scheduling performance in computation offloading, $\mathcal{SW}$ increases with the number of channels for all management schemes, except for the local computing scheme. This is because local computing does not require the support of wireless transmissions, so that its performance is independent of the number of channels.

## 6 Conclusion and Future Work

In this paper, the joint computation offloading and transmission scheduling for delay-sensitive applications in mobile edge computing has been studied. To characterize the dynamic management of the system with potential network uncertainties, a queueing model is formulated. By considering tradeoffs between local and edge computing, wireless features and noncooperative game behaviors of smart mobile users, we propose a novel mechanism, namely MOTM, to jointly determine the computation offloading scheme, the transmission scheduling discipline and the pricing rule. Both theoretical analyses and simulation results show that our proposed mechanism can guarantee that no individual mobile user has the incentive to strategically deviate the network-wide optimal management and can largely improve the social welfare compared to the counterparts.

In the future work, we will further integrate the cloud computing management in the designed mechanism. Specifically, if the cloud computing resource is limited and the offloaded tasks are overwhelmed, a congestion delay at the cloud may happen and will also contribute to the total computation offloading delay. In this case, the overall system may need to be formulated as a tandem queueing model with a multi-server multi-class priority queue for uplink transmission scheduling followed by another multi-server multi-class priority queue for cloud computing service. This motivates us to analyze the joint distribution of two queueing delays and develop an algorithm for jointly determining the optimal priority-aware disciplines for both queues.

## Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Discovery and CRD Grants.

## References

[1] S. Wang and S. Dey, “Adaptive mobile cloud computing to enable rich mobile multimedia applications,” *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 870–883, Jun. 2013.

[2] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, “A survey of mobile cloud computing: architecture, applications, and approaches,” *Wireless Commun. Mobile Comput.*, vol. 13, no. 18, pp. 1587–1611, 2013.

[3] P. Mach and Z. Becvar, “Mobile edge computing: A survey on architecture and computation offloading,” *IEEE Commun. Surveys Tut.*, vol. 19, no. 3, pp. 1628–1656, Jul.-Sep. 2017.

[4] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. S. Shen, “Cooperative edge caching in user-centric clustered mobile networks,” *IEEE Trans. Mobile Comput.*, vol. 17, no. 8, pp. 1791–1805, Aug. 2018.

[5] M. V. Barbera, S. Kosta, A. Mei, and J. Stefa, “To offload or not to offload? the bandwidth and energy costs of mobile cloud computing,” in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1285–1293.

[6] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper,” 2017.

[7] 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.

[8] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, “Computation offloading and resource allocation in Wireless cellular networks with mobile edge computing,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.

[9] 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.

[10] 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.

[11] D. Huang, P. Wang, and D. Niyato, “A dynamic offloading algorithm for mobile computing,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 1991–1995, Jun. 2012.

[12] 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.

[13] J. Liu, J. Wan, B. Zeng, Q. Wang, H. Song, and M. Qiu, “A scalable and quick-response software defined vehicular network assisted by mobile edge computing,” *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 94–100, Jul. 2017.

[14] C. Yi and J. Cai, “Two-stage spectrum sharing with combinatorial auction and stackelberg game in recall-based cognitive radio networks,” *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3740–3752, Nov. 2014.

[15] C. Yi, Z. Zhao, J. Cai, R. L. de Faria, and G. M. Zhang, “Priority-aware pricing-based capacity sharing scheme for beyond-wireless body area networks,” *Comput. Netw.*, vol. 98, pp. 29–43, 2016.

[16] C. Yi and J. Cai, “Ascending-price progressive spectrum auction for cognitive radio networks with power-constrained multiradio secondary users,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 781–794, Jan. 2018.

[17] H. Zhang, Y. Zhang, Y. Gu, D. Niyato, and Z. Han, “A hierarchical game framework for resource management in fog computing,” *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 52–57, Aug. 2017.

[18] L. Tang and S. He, “Multi-user computation offloading in mobile edge computing: A behavioral perspective,” *IEEE Netw.*, vol. 32, no. 1, pp. 48–53, Jan. 2018.

[19] J. Moura and D. Hutchison, “Game theory for multi-access edge computing: Survey, use cases, and future trends,” *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 260–288, 1Q 2019.

[20] K. Li, “A game theoretic approach to computation offloading strategy optimization for non-cooperative users in mobile edge computing,” *IEEE Trans. Sustainable Comput.*, to be published, doi: 10.1109/TSUSC.2018.2868655.

[21] M. Liu and Y. Liu, “Price-based distributed offloading for mobile-edge computing with computation capacity constraints,” *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 420–423, Jun. 2018.

[22] J. Zheng, Y. Cai, Y. Wu, and X. S. Shen, “Dynamic computation offloading for mobile cloud computing: A stochastic game-theoretic approach,” *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 771–786, Apr. 2019.

[23] X. Chen, “Decentralized computation offloading game for mobile cloud computing,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 4, pp. 974–983, Apr. 2015.

[24] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. S. Quek, “Offloading in mobile edge computing: Task allocation and computational frequency scaling,” *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3571–3584, Aug. 2017.

[25] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, “Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994–6009, Sep. 2017.

[26] S. Chen, Y. Wang, and M. Pedram, “A semi-markovian decision process based control method for offloading tasks from mobile devices to the cloud,” in *Proc. IEEE Global Commun. Conf.*, Dec. 2013, pp. 2885–2890.

[27] L. Liu, Z. Chang, and X. Guo, “Socially aware dynamic computation offloading scheme for fog computing system with energy harvesting devices,” *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1869–1879, Jun. 2018.

[28] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi, “Multiobjective optimization for computation offloading in fog computing,” *IEEE Internet Things J.*, vol. 5, no. 1, pp. 283–294, Feb. 2018.

[29] M. Mitra, "Mechanism design in queueing problems," *Econ. Theory*, vol. 17, no. 2, pp. 277–305, Mar. 2001.

[30] C. Yi, S. Huang, and J. Cai, "An incentive mechanism integrating joint power, channel and link management for social-aware D2D content sharing and proactive caching," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 789–802, Apr. 2018.

[31] D. Li, W. Saad, I. Guvenc, A. Mehbodniya, and F. Adachi, "Decentralized energy allocation for wireless networks with renewable energy powered base stations," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2126–2142, Jun. 2015.

[32] C. Yi and J. Cai, "Transmission management of delay-sensitive medical packets in beyond wireless body area networks: A queueing game approach," *IEEE Trans. Mobile Comput.*, vol. 17, no. 9, pp. 2209–2222, Sep. 2018.

[33] C. Yi, A. S. Alfa, and J. Cai, "An incentive-compatible mechanism for transmission scheduling of delay-sensitive medical packets in e-health networks," *IEEE Trans. Mobile Comput.*, vol. 15, no. 10, pp. 2424–2436, Oct. 2016.

[34] C. Yi and J. Cai, "A priority-aware truthful mechanism for supporting multi-class delay-sensitive medical packet transmissions in e-health networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 9, pp. 2422–2435, Sep. 2017.

[35] S. Wang and S. Dey, "Cloud mobile gaming: Modeling and measuring user experience in mobile wireless networks," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 16, no. 1, pp. 10–21, 2012.

[36] C. Wang, F. R. Yu, C. Liang, Q. Chen, and L. Tang, "Joint computation offloading and interference management in wireless cellular networks with mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7432–7445, Aug. 2017.

[37] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 2, no. 3, pp. 257–269, Jul. 2003.

[38] Z. Han, D. Niyato, W. Saad, T. Baar, and A. Hjrungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[39] J. Hu, L. L. Yang, and L. Hanzo, "Delay analysis of social group multicast-aided content dissemination in cellular system," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1660–1673, Apr. 2016.

[40] 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.

[41] J. S. Harsini and F. Lahouti, "Adaptive transmission policy design for delay-sensitive and bursty packet traffic over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 776–786, Feb. 2009.

[42] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 1078–1086.

[43] S. M. Betz and H. V. Poor, "Energy efficient communications in CDMA networks: A game theoretic analysis considering operating costs," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5181–5190, Oct. 2008.

[44] H. Xu and B. Li, "Dynamic cloud pricing for revenue maximization," *IEEE Trans. Cloud Comput.*, vol. 1, no. 2, pp. 158–171, Jul.-Dec. 2013.

[45] C. Buyukkoc, P. Variaya, and J. Walrand, "$c\mu$ rule revisited," *Adv. Appl. Prob.*, vol. 17, no. 1, pp. 237–238, 1985.

[46] Y. Masuda and S. Whang, "Dynamic pricing for network service: Equilibrium and stability," *Manag. Sci.*, vol. 45, no. 6, pp. 857–869, 1999.

[47] M. Reiser and S. S. Lavenberg, "Mean-value analysis of closed multichain queuing networks," *J. ACM*, vol. 27, no. 2, pp. 313–322, 1980.

[48] J. Harrison, *Brownian Motion and Stochastic Flow Systems*, Hoboken, NJ, USA: Wiley, 1985.

[49] H. C. Gromoll, "Diffusion approximation for a processor sharing queue in heavy traffic," *Ann. Appl. Probability*, vol. 14, no. 2, pp. 555–611, 2004.

[50] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, Hoboken, NJ, USA: Wiley, 2013.

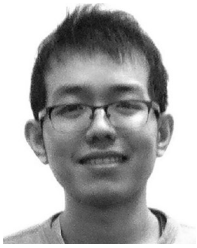
[51] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2008, http://cvxr.com/cvx/

[52] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, Cambridge, MA, USA: Academic Press, 2013.

[53] M. Viitanen, J. Vanne, T. D. Hamalainen, and A. Kulmala, "Low latency edge rendering scheme for interactive 360 degree virtual reality gaming," in *Proc. IEEE 38th Int. Conf. Distrib. Comput. Syst.*, Jul. 2018, pp. 1557–1560.

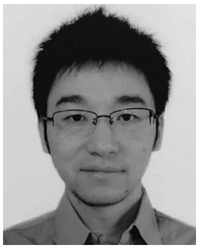
[54] S. Huang, C. Yi, and J. Cai, "A sequential posted price mechanism for D2D content sharing communications," in *Proc. IEEE Global Commun. Conf.*, Dec. 2016, pp. 1–6.

[55] A. S. Alfa, *Queueing theory for Telecommunications: Discrete Time Modelling of a Single Node System*, Berlin, Germany: Springer, 2010.



**Changyan Yi** (S'16-M'18) received the BSc degree from the Guilin University of Electronic Technology, China, in 2012, and the MSc and PhD degrees from the University of Manitoba, MB, Canada, in 2014 and 2018, respectively. He is currently working as a research associate in the Department of Electrical and Computer Engineering, University of Manitoba, Canada. He was awarded the Chinese Government Award for Outstanding Students Abroad in 2017, A. Keith Dixon Graduate Scholarship in Engineering for 2017-2018, Edward R. Toporeck Graduate Fellowship in Engineering for 2014-2017 (four times), University of Manitoba Graduate Fellowship (UMGF) for 2015-2018, and IEEE ComSoc Student Travel Grant for IEEE Globecom 2016. His research interests include algorithmic game theory, queueing theory and their applications in radio resource management, wireless transmission scheduling, and network economics. He is a member of the IEEE.

**Jun Cai** (M'04-SM'14) received the BSc and the MSc degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1996 and 1999, respectively, and the PhD degree in electrical engineering from the University of Waterloo, ON, Canada, in 2004. From June 2004 to April 2006, he was with McMaster University, Hamilton, ON, as a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellow. Since July 2006, he has been with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada, where he is currently a professor. His current research interests include energy-efficient and green communications, dynamic spectrum management and cognitive radio, radio resource management in wireless communications networks, and performance analysis. He served as the TPC co-chair for the IEEE VTC-Fall 2012 Wireless Applications and Services Track, the IEEE Globecom 2010 Wireless Communications Symposium, and the IWCMC 2008 General Symposium; the publicity co-chair for IWCMC in 2010, 2011, 2013, and 2014, respectively; and the registration chair for QShine in 2005. He also served on the editorial board of the *Journal of Computer Systems, Networks, and Communications* and as a guest editor of the special issue of the *Association for Computing Machinery Mobile Networks and Applications*. He received the Best Paper Award from Chinacom in 2013, the Rh Award for outstanding contributions to research in applied sciences in 2012 from the University of Manitoba, and the Outstanding Service Award from IEEE Globecom in 2010. He is a senior member of the IEEE.



**Zhou Su** received the PhD degree from Waseda University, Tokyo, Japan, in 2003. He is an associate editor of *IET Communications*, and associate editor of the *IEICE Transactions on Communications*. He is the chair of the Multimedia Services and Applications over Emerging Networks Interest Group (MENIG) of the IEEE Comsoc Society, the Multimedia Communications Technical Committee. He also served as the co-chair of several international conferences including IEEE VTC Spring 2016, IEEE CCNC2011, etc. He is a TPC member of some flagship conferences including IEEE INFOCOM, IEEE ICC, IEEE Globecom, etc. His research interests include multimedia communication, wireless communication, and network traffic. He received the best paper award of IEEE CyberSciTech2017, WiCon2016, CHINACOM2008, and the Funai Information Technology Award for Young Researchers in 2009.