Multi-Helper NOMA for Cooperative Mobile Edge Computing

Saadet Simay Yılmaz, *Member, IEEE*, and Berna Özbek, *Senior Member, IEEE*

***Abstract*— The next-generation wireless networks are expected to support a number of computation-intensive and delay-sensitive applications such as virtual reality (VR), autonomous driving, telesurgery and unmanned aerial vehicles (UAVs). Since many devices are computation and power limited, mobile edge computing (MEC) has been deemed as a promising way to enhance computation service. In this paper, we propose a novel cooperative MEC that exploits the combination of non-orthogonal multiple access (NOMA) and multiple helpers. In the proposed system featuring a user, multiple helpers and a base station (BS), the user can simultaneously offload its computation-intensive tasks to the helpers using NOMA when there is no strong direct transmission link between the user and the BS. Then, the helpers can compute and offload these tasks through NOMA. Thus, in the proposed scheme, the computation and offloading modes at the helpers are determined with respect to the optimized task offloading decision factor. The simulation results show that the proposed NOMA-based cooperative MEC significantly increases the total offloading data under the latency constraints compared to the benchmark schemes featuring one helper with strong direct transmission link.**

***Index Terms*— Cooperative mobile edge computing (MEC), low latency, multi-helper, non-orthogonal multiple access (NOMA), offloading.**

I. Introduction

RECENT advances in the next-generation wireless technologies have motivated several computationally intensive and latency-critical applications such as virtual reality (VR), augmented reality (AR), autonomous driving, telesurgery, unmanned aerial vehicles (UAVs) and Internet of Things (IoT) [1]. These applications require ultra-low-latency communication, computation and control among many wireless devices. Since the devices have small physical sizes, limited computation capacities and limited power sources, it is a challenging task to handle intensive computation loads at the user side.

To overcome these limitations, cloud computing offers one possible solution by offloading the computation-intensive tasks from users to the cloud. However, because of the data propagation through wide area networks, the long propagation distances between the devices and the centralized cloud, cloud computing can cause excessive latency computation and heavy traffic loads at the backhaul networks. Therefore, cloud computing may not support latency-critical applications.

Manuscript received 17 April 2021; revised 5 August 2021; accepted 15 September 2021. Date of publication 11 October 2021; date of current version 8 July 2022. This work was supported by the European Union Horizon 2020 through the RISE 2018 Scheme (H2020-MSCA-RISE-2018) under Marie Skłodowska-Curie Grant 823903 (RECENT). The Associate Editor for this article was S. Mumtaz. *(Corresponding author: Saadet Simay Yılmaz.)*

The authors are with the Department of Electrical and Electronics Engineering, Izmir Institute of Technology, 35430 Izmir, Turkey (e-mail: simayyilmaz@iyte.edu.tr; bernaozbek@iyte.edu.tr).

Digital Object Identifier 10.1109/TITS.2021.3116421

To resolve this issue, mobile edge computing (MEC) has been a promising solution to enable computation-intensive and latency-critical applications. MEC utilizes the powerful computing capabilities such as a MEC server integrated into the base station (BS) within the radio access network. Compared with cloud computing, in MEC systems, users offload computation-intensive tasks to the powerful MEC servers in proximity to base stations (BSs) and access points (APs) for execution, which avoids data delivery over the backhaul networks and reduces latency [2]. MEC can considerably decrease computation latency and significantly reduce the traffic loads at the backhaul networks [3]. One of the other benefits of the MEC system is that by offloading computation-intensive tasks, the energy consumption of the devices can be significantly saved. Besides, since the tasks can be computed at the adjacent BS instead of the remote cloud center, the congestion in the core network can be effectively relieved, thereby leading to the reduction of the latency. However, there are some main cases that need to be overcome. Firstly, the computation resource at the BS cannot be always sufficient to support all devices. Secondly, there could be no strong direct transmission link to the BS.

On the other hand, non-orthogonal multiple access (NOMA) has been recognized as a critical technology for next-generation wireless communication systems to meet extremely high data rate requirements. Compared to conventional orthogonal multiple access (OMA), such as time division multiple access (TDMA) and orthogonal frequency division multiple access (OFDMA), NOMA can achieve higher spectral efficiency by allowing multiple users to operate simultaneously in the same time-frequency resource with the different power levels [4]. This can be achieved by applying superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver. The motivation of employing the NOMA technology is to reduce the offloading latency and improve the performance of the MEC based systems. Therefore, these two communication techniques, MEC and NOMA, can be combined to provide gains in terms of the total offloading data and the latency.

A. State-of-the-Art on Computation Offloading in MEC

Recently, several works have been published addressing the issue of latency minimization, task offloading, resource management and energy consumption in MEC-enabled networks. A computation efficiency metric has been presented in [5] to maximize the energy efficiency used for both computing and communication. Motivated by the performance gains of applying NOMA over OMA, a multi-user uplink NOMA-based MEC network has been analyzed in [6]. Both NOMA uplink and downlink transmissions have been applied in [7] to reduce the latency and energy consumption considering NOMA-based MEC offloading. Then, completion time and energy minimization have been respectively optimized in [8] and in [9] for the different users with different computation requirements in NOMA-based MEC networks. However, in [6]–[9], only one cluster of users has been considered and the resource allocation among the different clusters of users forming NOMA has been ignored.

Since each resource is suggested to be multiplexed by a small number of users due to decoding complexity and error propagation, the importance of resource allocation among the different clusters of users forming NOMA has been stated in [10], [11]. Thus, the total energy minimization problem has been studied in [12] for an uplink NOMA-based MEC network by considering resource allocation for the different clusters with two users to perform NOMA. In [13], the total energy consumption has been minimized considering the NOMA-based transmission in both task uploading and downloading. The authors in [14] have presented a total energy consumption minimization problem while achieving the computation latency constraint. Additionally, in [15], a problem has been formulated for the purpose of maximizing the probability of successful computation by jointly optimizing the offloading time consumption, the power allocation and the offloading ratios. Then, a joint radio and computation resource allocation has been stated in [16] for a multi-user MEC system with computation interference issue regarding the sum offloading rate maximization and the sum energy minimization problems.

Although the above studies have demonstrated the effectiveness of MEC through NOMA in enhancing the computation performance of wireless networks, the computation resource at the BS cannot be always sufficient to support all devices. In addition, these studies have investigated the task offloading scheme between devices and BS, regardless of cooperative computing. The studies in [17]–[20] have presented the cooperative computing for MEC-enabled networks. The authors in [17] have studied the cooperation of multiple MEC-enabled BSs to enhance the computation offloading service with extra tasks. In [18], cooperative relay-aided transmission with a device-to-device (D2D) edge computing offloading architecture has been introduced. The study in [19] has demonstrated the D2D-assisted and NOMA-based MEC system to reduce the computing load of the edge server. In [20], MEC networks with wireless backhaul have been considered where users can offload their computational tasks to the MEC server through a small cell base station (SBS) so that computation resource at the MEC server is shared among offloading users.

To deal with this issue, another potential solution is to offload part of the computation-intensive tasks to helpers via NOMA. In the light of this solution, in [21], a cooperative edge computing in both computation and communication has been given for a basic three-node MEC system consisting of one user node, one helper node and one AP node with a MEC server integrated. The main objective is to minimize the total energy consumption while satisfying the user's computation latency constraint based on TDMA transmission protocol. In [22], a basic three-node MEC with NOMA-based cooperative edge computing has been presented to maximize sum offloading data subject to the latency constraints. Similarly, the authors of [23] have represented a NOMA-aided user cooperation scheme in a three-node MEC wireless power transfer (WPT) system based on the energy consumption minimization problem. The only difference between the system models in [22] and [23] is the available computation tasks at the helper. The helper does not have computation tasks in [23]. However, these studies [21]–[23] are based on only one helper scenario. In addition to that, the user can directly offload its own tasks to the BS in these three-node MEC systems. On the other hand, different from the basic three-node MEC system, a cooperative task computation framework has been considered in [24] with the purpose of maximizing the number of accomplished tasks and minimizing the power consumption of users. In this cooperative task computation system between the user and the BS, one user can help the other user for task computation via D2D transmissions.

Furthermore, the MEC-based architecture has been recently recognized as a promising solution for many intelligent transportation systems (ITSs). The authors in [25] and [26] have introduced the MEC technology into vehicular networks in order to enable computing offloading in computation-intensive and latency-critical applications on resource-limited vehicles.

B. Motivations and Contributions

Motivated by the above discussions, as shown in Fig. 1, the user can have computation-intensive mobile applications, such as autonomous driving, gesture recognition and three-dimension (3D) modeling, i.e., real time navigation and real time traffic monitoring. Also, these tasks can also be executed under latency constraints. Therefore, it is a challenging task for the devices to handle these intensive computation loads with the latency requirements. If there were a strong direct transmission link between the user and the MEC server namely the three-node MEC system, the user would offload its computation-intensive and delay-sensitive tasks to a nearby MEC server for remote task execution. However, for the user which is at the cell edge, there is no strong direct transmission link between the user and the MEC server. Therefore, the cell-edge user chooses to offload its computationally intensive and latency-critical tasks to the server through the aid of helpers.

To be specific, in this paper, by considering the state-of-the-art work listed above, we propose a new paradigm of cooperative MEC with multiple helpers based on NOMA. In the proposed framework, a user simultaneously offloads its tasks using NOMA to many helpers at the first slot. Then, the helpers can both compute and offload the user's tasks at the second slot. In this way, NOMA is adopted for offloading in both time slots to increase the total offloading data. Furthermore, since the distances between the user and the helpers affect the maximization of the total offloading data, adjusting

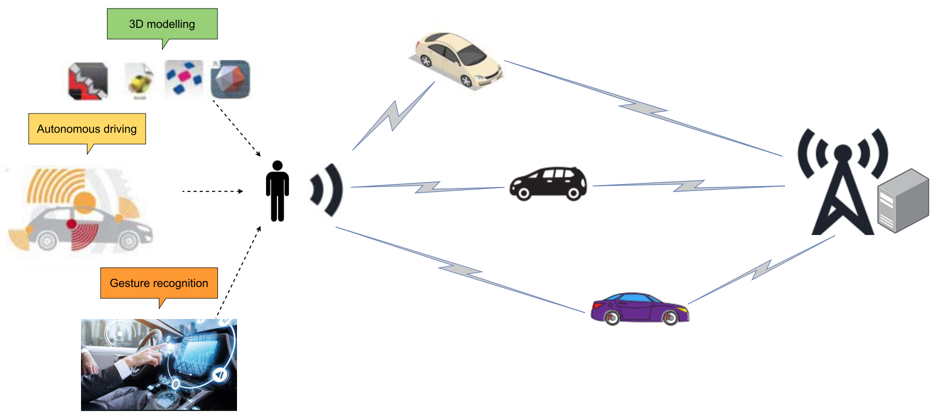

Fig. 1. The scenario of the cooperative MEC based on NOMA with multiple helpers.

the distances is important while achieving latency constraints of the user's applications. As a result, the communication resource optimization is performed to improve the user computation experience through emerging MEC systems. The main contributions in this work are summarized as follows:

1) We propose a multi-helper cooperative MEC system based on NOMA to maximize the total offloading data under the latency and power constraints while in the literature [21], [22], and [23], only one helper case for the MEC system has been considered and the NOMA scheme is only applied for the link between the user and the helper while it is not used for the link between the helper and the BS.
2) Afterward, we analyze the proposed framework under the optimum distances of the multiple helpers to maximize the total offloading data.
3) We provide simulation results to show the superiority of the proposed framework in terms of the total offloading data compared with the benchmarking solutions in [21], [22] and [23].

The rest of this paper is organized as follows. In Section II, we give the cooperative system model with multiple helpers. The proposed problem formulation and the optimum solution are presented in Section III. The extensive performance results of the proposed framework are provided in Section IV. Finally, the conclusions are drawn in Section V.

II. System Model

In this paper, we consider a NOMA-based cooperative MEC system which consists of a BS integrated with a MEC server, one user and M helpers, as illustrated in Fig. 2.

Let $\mathcal{M} = \{1, 2, \ldots, M\}$ denote the set of helpers. In the system, all helpers are identical and have the same hardware specifications. These helper nodes can be a vehicle or a device in a vehicle which have certain computation and communication resources. Besides, the user can be considered as a requesting vehicle and the BS integrated with a MEC server can be called a serving vehicle in congested traffic. For convenience, we introduce them as the user, helpers and the BS throughout the paper.

We assume that the user has individual computation tasks with data size, L_u to complete successfully under a common

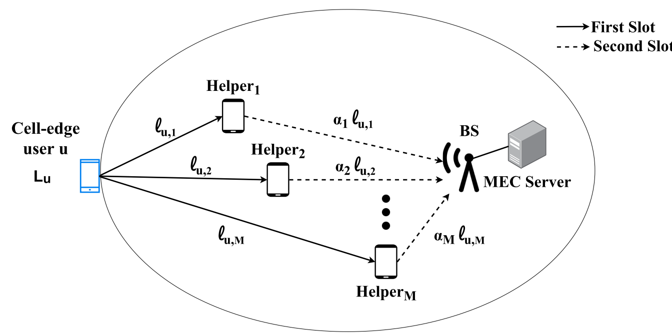

Fig. 2. The proposed cooperative MEC based on NOMA with M helpers.

latency constraint, but the helpers, Helper$_m$ $\forall m \in \mathcal{M}$, do not have computation tasks. The BS is integrated with the MEC server to execute the computation-intensive tasks offloaded by the helpers. It is assumed that there is no strong direct transmission link between the user and the MEC server since the user is at the cell edge. The user sends a certain part of its tasks to the MEC server through the helpers. Besides, these helpers are at the cell-center.

We consider the partial computation offloading operation mode that assumes the computational tasks can be divided into two independent parts. One part is executed locally, while the other part is offloaded to the helpers. In Fig. 2, the user simultaneously offloads $\ell_{u,m}$ part of its own data L_u, to mth helper and locally computes the rest of the data. Each helper computes and offloads the tasks received from the user. The Helper$_m$ uses task offloading decision factor α_m to decide the fraction of the offloading and the computing modes. That is, for a particular time instant, the Helper$_m$ can have $0 < \alpha_m < 1$ to act the double modes. In the system, α_m is obtained through the optimization algorithm and cannot be 0 or 1 which indicates that the helpers both perform some portions of the offloaded tasks from the cell-edge user and establish transmission to offload some portion of the tasks.

For a cooperative edge computing system, the cooperation between the user and the helpers is essential. The main motivation of the proposed cooperative MEC system is that, for a common latency constraint T, the user offloads its own tasks to the MEC server with the aid of M helpers. Latency

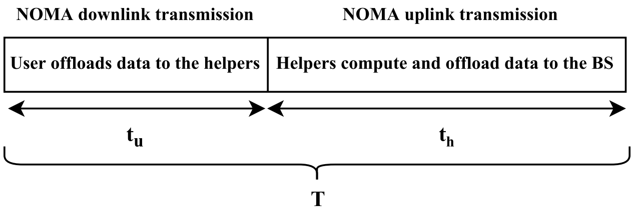

Fig. 3. Latency constraint scheme.

constraint, T, is divided into two time slots such as t_u and t_h as shown in Fig. 3. In the first time slot, t_u, the user offloads the tasks to the helpers. Then, the helpers offload and compute a certain part of the user's tasks in the second time slot, t_h.

In the first time slot which performs NOMA downlink transmission, without loss of generality, M helpers are sorted as $g_{u,1} \geq g_{u,2} \geq \cdots \geq g_{u,M}$ where $g_{u,m}$ is the channel gain between the user and the Helper$_m$, $\forall m \in \mathcal{M}$. The channel gain is modeled by using Rayleigh distribution with a variance of distance-dependent path loss coefficient. The corresponding channel state information (CSI) is assumed to be available at each receiving node.

In the first time slot t_u, the offloaded data from the user to the Helper$_m$ is calculated as $\ell_{u,m} = t_u R_m^D$, and the transmission rate R_m^D is given by [22], [27]–[32] and defined as follows:

$$R_m^D = \log_2\left(1 + \frac{P_m^D\, g_{u,m}}{g_{u,m} \sum_{j=1}^{m-1} P_j^D + \sigma^2}\right) \tag{1}$$

where P_m^D denotes the downlink transmit power of the Helper$_m$, and σ^2 is the variance of zero-mean complex additive white Gaussian noise (AWGN).

In the second time slot which performs NOMA uplink transmission, depending on α_m values, all helpers simultaneously offload their data, $\ell_{m,o} = \alpha_m(\ell_{u,m})$, to the BS relying on the uplink NOMA scheme. Under this scheme, the helpers are sorted based on their channel gains, namely $g_{M,o} \geq g_{M-1,o} \geq \cdots \geq g_{1,o}$ where $g_{m,o}$ is the channel gain between the Helper$_m$ and the BS. Then, the BS utilizes the SIC technique to decode the data coming from the helpers. According to the principle of the SIC, the BS first decodes the information from the helper with the larger channel gain and then removes it from the other helpers' signals. Therefore, the transmission rate of the Helper$_m$ is given by [4], [14], [33], [34]:

$$R_m^U = \log_2\left(1 + \frac{P_m^U\, g_{m,o}}{\sum_{j=1}^{m-1} P_j^U\, g_{j,o} + \sigma^2}\right) \tag{2}$$

where P_m^U is the uplink transmit power of the Helper$_m$.

III. Proposed Framework and Problem Formulation

In this section, a total offloading data (TOD) maximization problem with given constraints is formulated for the proposed cooperative MEC based on NOMA with the M helpers. In the proposed framework, since there is no strong direct transmission link between the cell-edge user and the MEC server, the user sends a certain part of its tasks to the MEC server through the cell-center helpers. The helpers do not act as a pure relay since they can compute some portions of the offloaded tasks from the cell-edge user and also offload some portions of these computation tasks to a MEC server. In the proposed approach, we only focus on the offloading part at the user side rather than computing since our aim is to maximize TOD.

The objective function is TOD that is sum of the user's and helpers' offloading data and defined as function $w(.)$ given below

$$w(\mathbf{t}, \mathbf{P}) = \sum_{m=1}^{M} \left(t_u R_m^D + t_h R_m^U\right) \tag{3}$$

Since the data $\alpha_m \left(t_u R_m^D\right)$ will be offloaded from the Helper$_m$ to the BS in the uplink transmission, the function $w(.)$ can be rewritten as the following function f(.):

$$f(\mathbf{t}, \mathbf{P}, \boldsymbol{\alpha}) = \sum_{m=1}^{M} \left(t_u R_m^D + \alpha_m (t_u R_m^D)\right) \tag{4}$$

Accordingly, TOD maximization problem subject to the latency constraints and power allocation factors is given by

$$\max_{\mathbf{t},\mathbf{P},\boldsymbol{\alpha}} \quad f(\mathbf{t}, \mathbf{P}, \boldsymbol{\alpha}) \tag{5}$$
$$\text{s.t. } R_m^D \geq R_{th,m}^D, \quad \forall m \in \mathcal{M} \tag{5a}$$
$$\sum_{m=1}^{M} P_m^D \leq P_u \tag{5b}$$
$$P_m^U \leq P_h, \quad \forall m \in \mathcal{M} \tag{5c}$$
$$t_u + t_h \leq T \tag{5d}$$
$$\frac{(1-\alpha_m)\, t_u R_m^D\, \phi}{f_m} \leq T - t_u, \quad \forall m \in \mathcal{M} \tag{5e}$$
$$0 < \alpha_m < 1, \quad \forall m \in \mathcal{M}. \tag{5f}$$

where $\mathbf{t} = [t_u, t_h]$, $\mathbf{P} = \left[P_1^D, \ldots, P_M^D, P_1^U, \ldots, P_M^U\right]$ and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_M]$, P_u is the maximum total transmit power of the user, and P_h is the maximum transmit power of each helper. ϕ denotes the number of central processing unit (CPU) cycles required to calculate 1 bit of data. f_m represents the computing power of the helpers, also known as the CPU frequency.

Constraint (5a) gives the minimum data rate constraints in downlink that should not be less than a given threshold rate $R_{th,m}^D$. Constraints (5b) and (5c) represent the power constraints in NOMA downlink and NOMA uplink transmissions, respectively. Constraint (5d) shows the total offloading time constraint for both the user and the helpers. Constraint (5e) represents the local computing time constraint of each helper. Constraint (5f) indicates the range of the task offloading decision factor for each helper.

In the proposed approach, we firstly investigate the optimum distances of the helpers which maximize TOD assuming there is no energy restriction on the helpers. After determining the optimum distances, we select the helpers having these distances. Then, the optimization algorithm is applied to find the optimal time, power allocation factors and the task offloading decision factor.

A. Problem Solution

In this section, we derive the solution to Problem (5) for the proposed system with the M helpers NOMA-based cooperative MEC. Firstly, we give the solution for the case of $M = 2$ and then generalize to the M helpers. Since the Problem (5) is a constrained nonlinear multivariable problem, we apply the interior point method. In this method, violation of inequality constraints is prevented by adding a barrier term to the objective function that ensures the optimal unconstrained values to be in the feasible space.

1) NOMA Downlink Transmission: Since Problem (5) is an increasing function with respect to P_m^D, the constraint associated with the downlink transmit power can be written as $P_1^D + P_2^D = P_u$ to maximize the objective function. Thus, in the downlink transmission, $\beta \in (0, 1)$ is the power allocation factor and becomes one of the optimization parameters in the Problem (5). Then, the allocated downlink transmit power, P_1^D, for the Helper_1 is determined by βP_u while the allocated downlink transmit power, P_2^D, for the Helper_2 is calculated as $(1-\beta)\, P_u$.

The achievable rate of the Helper_m in an OMA system [27] is given by

$$R_m^{OMA} = \frac{1}{2}\log_2\left(1 + \frac{P_u\, g_{u,m}}{\sigma^2}\right) \tag{6}$$

where the factor $\frac{1}{2}$ is due to the fact that conventional OMA results in a multiplexing loss of $\frac{1}{2}$.

In the downlink transmission, the achievable rate in the NOMA system should be no less than in the OMA system. In this case, $R_{th,m}^D$ is set as R_m^{OMA}. Then, the range of β can be obtained from the constraint (5a) directly as follows:

$$\begin{aligned} R_1{}^D &\geq R_1{}^{OMA} \\ \log_2\left(1 + \frac{\beta P_u\, g_{u,1}}{\sigma^2}\right) &\geq \frac{1}{2}\log_2\left(1 + \frac{P_u\, g_{u,1}}{\sigma^2}\right) \\ \beta &\geq \frac{\left(\sqrt{1 + \frac{P_u g_{u,1}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,1}} \end{aligned} \tag{7}$$

$$\begin{aligned} R_2{}^D &\geq R_2{}^{OMA} \\ \log_2\left(1 + \frac{(1-\beta)\, P_u\, g_{u,2}}{\beta P_u g_{u,2} + \sigma^2}\right) &\geq \frac{1}{2}\log_2\left(1 + \frac{P_u\, g_{u,2}}{\sigma^2}\right) \\ \beta &\leq \frac{\left(\sqrt{1 + \frac{P_u g_{u,2}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,2}} \end{aligned} \tag{8}$$

Thus, the range of β is given by:

$$\frac{\left(\sqrt{1 + \frac{P_u g_{u,1}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,1}} \leq \beta \leq \frac{\left(\sqrt{1 + \frac{P_u g_{u,2}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,2}} \tag{9}$$

Then, we can extend the downlink power allocation factors to the M helpers. Assume that n, m represent the index of each helper. The superscript of $\beta_m^{n,m}$ denotes the indices of the pairing helpers and the upper bound of $\beta_m^{n,m}$ is given as z_m [27]. Thus, it is written as

$$\begin{aligned} \beta_2^{1,2} &= \frac{\left(\sqrt{1 + \frac{P_u g_{u,2}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,2}} \triangleq z_2 \\ \beta_3^{1,3} &= \beta_3^{2,3} = \frac{\left(\sqrt{1 + \frac{P_u g_{u,3}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,3}} \triangleq z_3 \\ &\ \ \vdots \\ \beta_M^{1,M} &= \beta_M^{2,M} = \cdots = \beta_M^{M-1,M} \\ &= \frac{\left(\sqrt{1 + \frac{P_u g_{u,M}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,M}} \triangleq z_M \end{aligned} \tag{10}$$

According to the channel gain assumption in the downlink transmission, the order of z_m in (10) for $\forall m \in \mathcal{M}$ is given as

$$z_1 \leq z_2 \leq \cdots \leq z_M \tag{11}$$

where $z_1 = \frac{\left(\sqrt{1+\frac{P_u g_{u,1}}{\sigma^2}} - 1\right)\sigma^2}{P_u\, g_{u,1}}$.

In this way, the range of the power allocation factors, β_m, of each helper is determined as:

$$z_{m-1} \leq \beta_m \leq z_{m+1} \tag{12}$$

where $1 \leq m \leq M-1,\ M \geq 3$ and z_0 denotes 0.

Therefore, the downlink transmit power of each helper is written as $\beta_m P_u$ while the downlink transmit power of the Helper_M is found from the expression as $P_u - \left(\sum_{m=1}^{M-1} \beta_m P_u\right)$.

It can be noticed that Problem (5) is non-decreasing with respect to t_u. In order to maximize the total offloading data, the total latency constraint, T, is the sum of t_u and t_h since the range of t_u is given in $0 < t_u < T$. Thus, the constraint (5d) becomes:

$$t_u + t_h = T \tag{13}$$

Thus, the constraint (5e) can be rewritten as

$$\frac{(1-\alpha_m)\, t_u R_m^D\, \phi}{f_m} \leq t_h, \quad \forall m \in \mathcal{M} \tag{14}$$

2) NOMA Uplink Transmission: In the second time slot t_h since the data $\alpha_m\left(t_u R_m^D\right)$ will be offloaded from the Helper_m to the BS with transmission rate R_m^U, we can obtain the following equations:

$$\alpha_1\left(t_u R_1^D\right) = (T - t_u)\log_2\left(1 + \frac{P_1^U g_{1,o}}{\sigma^2}\right) \tag{15}$$

$$\alpha_2\left(t_u R_2^D\right) = (T - t_u)\log_2\left(1 + \frac{P_2^U g_{2,o}}{P_1^U g_{1,o} + \sigma^2}\right) \tag{16}$$

On the other hand, the uplink transmission power must be lower than or equal to P_h in the uplink as shown in constraint (5c). Thus, we can derive the nonlinear constraints related to the optimization Problem (5) for the uplink transmission powers, P_1^U and P_2^U, using (15) and (16) as follows:

$$\frac{\left(2^{\frac{\alpha_1\left(t_u R_1^D\right)}{t_h}}-1\right)\sigma^2}{g_{1,o}} \leq P_h \tag{17}$$

$$\frac{\left(2^{\frac{\alpha_2\left(t_u R_2^D\right)}{t_h}}-1\right)\left(2^{\frac{\alpha_1\left(t_u R_1^D\right)}{t_h}}\right)\sigma^2}{g_{2,o}} \leq P_h \tag{18}$$

Then, we can extend the uplink transmission powers given in constraint (5c) to the M helpers, $\forall m \in \mathcal{M}$ as

$$\frac{\left(2^{\frac{\alpha_m\left(t_u R_m^D\right)}{t_h}}-1\right)\prod_{i=1}^{m-1}\left(2^{\frac{\alpha_i\left(t_u R_i^D\right)}{t_h}}\right)\sigma^2}{g_{m,o}} \leq P_h \tag{19}$$

Furthermore, the range of the task offloading decision factor for each helper is determined to provide better cooperation through the helpers in the system. Thus, constraint (5f) is redefined as follows:

$$0.3 \leq \alpha_m \leq 0.7, \quad \forall m \in \mathcal{M}. \tag{20}$$

Accordingly, the optimization Problem in (5) can be rewritten as follows:

$$\max_{\mathbf{t},\mathbf{P},\boldsymbol{\alpha}} \quad f(\mathbf{t},\mathbf{P},\boldsymbol{\alpha}) \tag{21}$$

$$\text{s.t. } (12),\ (14),\ (19),\ (20). \tag{22}$$

It is also equivalent to minimizing -f(.), thus the corresponding optimization problem can be efficiently solved by using some standard nonlinear programming optimization tools [35]. The minimum of a constrained nonlinear multivariate function can be determined using the interior-point method.

In the interior-point method, a log-barrier term is used for the inequality constraints, and the problem with inequality constraints can be reduced to having only equality constraints [36]. Barrier functions are usually a logarithmic function and can be used to transform a constrained problem into a sequence of unconstrained problems. These functions avoid the iterates from leaving the feasible region by acting as a barrier.

The details of the interior-point algorithm are summarized in Algorithm 1 for the multi-helper scenario. In the minimization problem, $\mathbf{x}$ is defined as a vector of the components; $\mathbf{x} = [t_u, \beta_m, \alpha_m]$, $\forall m \in \mathcal{M}$. The vector $\mathbf{x}$ satisfying all the constraints is called a feasible solution to the Problem (21). The initial values $\mathbf{x}^0$ are determined by defining lower and upper bounds range for each component t_u, β_m and α_m in $\mathbf{x}$.

Then, we define an approximate problem $f_\mu(\mathbf{x},\mathbf{s})$ with a barrier parameter μ as follows:

$$f_\mu(\mathbf{x},\mathbf{s}) = -f(\mathbf{x}) - \mu \sum_m \ln(s_m) \tag{23}$$

where $\mathbf{s} = [s_1, s_2, \ldots, s_M] > 0$ represents the slack variables. The slack variables are used to transform the inequality constraints to the equality constraints. $\mathbf{s}$ is obtained through the algorithm to guarantee the positiveness of the slack variables. There exists a slack variable against each nonlinear inequality constraint. The nonlinear inequality constraints in (19) are called as $\mathbf{G} = [G_1, G_2, \ldots, G_M]$.

The general idea is that as the barrier parameter, μ decreases, the minimum to the approximate problem $f_\mu(\mathbf{x},\mathbf{s})$ will approach the minimum to the original Problem (21).

As a result of Algorithm 1, the optimum output values are obtained as $t_u^*, \beta_m^*, \alpha_m^*$, $\forall m \in \mathcal{M}$ in order to provide maximum total offloading data under the constraints.

Algorithm 1 Optimizing TOD Based on Interior-Point Algorithm for M Helpers

Input: $g_{u,m}, g_{m,o}$; $\forall m \in \mathcal{M}$, $\mathbf{x} = [t_u, \beta_m, \alpha_m]$, $\forall m \in \mathcal{M}$

• **Initialization Step**

Slack variables, $\mathbf{s}^0 = [s_1^0, s_2^0, \ldots, s_M^0] > 0$

Rearrange the inequality constraints in (19) as $\mathbf{G} = [G_1, G_2, \ldots, G_M]$

Choose initial feasible points $\mathbf{x}^0$ with $\mathbf{G}(\mathbf{x}^0) < 0$

Select a convergence tolerance, $\varepsilon > 0$

Select the barrier parameter $\mu^0 > 0$

Set j=0

• **Main Step**

Reformulate the objective function in (21) to the approximate problem using a barrier function with $\mathbf{s}$

$$\min_{\mathbf{x},\mathbf{s}} \quad f_\mu(\mathbf{x},\mathbf{s}),$$
$$s.t. \quad G_m(\mathbf{x}) + s_m = 0,\ \forall m \in \mathcal{M}$$

while $\left|f_\mu(\mathbf{x}^j,\mathbf{s}^j) - f_\mu(\mathbf{x}^{j+1},\mathbf{s}^{j+1})\right| < \varepsilon$ **do**

Define the Lagrange function $\mathcal{L}(\mathbf{x},\mathbf{s},\boldsymbol{\lambda})$ of $f_\mu(\mathbf{x},\mathbf{s})$ with Lagrange multipliers $\boldsymbol{\lambda}$ and solve the corresponding Karush-Kuhn-Tucker (KKT) equations

$$\mathcal{L}(\mathbf{x},\mathbf{s},\boldsymbol{\lambda}) = -f(\mathbf{x^j}) - \mu^j \sum_m \ln(s_m^j) + \lambda_m^j(G_m(\mathbf{x^j}) + s_m^j)$$

Solve the approximate problem by decreasing μ:

• Starting from $\mathbf{x}^0$, use an unconstrained search technique such as an iterative descent method applicable to unconstrained problems including steepest descent or Newton's method to find the point that minimizes $f_\mu(\mathbf{x},\mathbf{s})$ and call them as the new variables, $\mathbf{x}^{j+1}$, $\mathbf{s}^{j+1}$ and $\boldsymbol{\lambda}^{j+1}$

$\mu^{j+1} = \sigma\mu^j$, where $\sigma \in (0,1)$

$j = j + 1$

end while

Output: $t_u^*, \beta_m^*, \alpha_m^*$; $\forall m \in \mathcal{M}$

The complexity of the interior-point method can be given [37] as $\mathcal{O}(\sqrt{n}\log\frac{1}{\varepsilon})$ iterations, where n is the number of variables in the problem, depends on the number of helpers, M, in the system. Thus, the number of helpers and the choice of the convergence tolerance, ε, affect the complexity. In the algorithm, convergence tolerance, ε, is selected as 10^{-6}.

IV. Performance Evaluation

In this section, we present the simulation results to evaluate the performance of the proposed system. The simulation parameters are listed in Table I.

TABLE I
SIMULATION PARAMETERS

System Parameters	Values
Carrier frequency, f_c	3.5GHz
Transmission bandwidth, B	1MHz
Noise variance, σ^2	-159dBm
CPU frequency, f_m	1GHz
Required CPU cycles of tasks, ϕ	1000cycles/bit
Maximum total transmit power of the user, P_u	0.5W
Maximum transmit power of the helper, P_h	0.8W

For both the user and helpers, the channel is modeled by using Rayleigh fading with distance-dependent path loss whose parameters depend on whether the receiver is the BS or the helper.

When the receiver is the BS, the path loss model [38] is defined for the distance between the Helper$_m$ and the BS, $d_{m,o}$ $[m]$.

$$L(d_{m,o}) = 36.7\log_{10}\left(d_{m,o}\right) + 22.7 + 26\log_{10}\left(f_c\right) \quad [dB] \tag{24}$$

When the receiver is the helper, the path loss model [39] is given for the distance between the user and the corresponding Helper$_m$, $d_{u,m}$ $[m]$.

$$L(d_{u,m}) = \max(PL(d_{u,m}), PL\_B1(d_{u,m})) \quad [dB] \tag{25}$$

where PL is defined as:

$$PL(d_{u,m}) = 20\log_{10}\left(d_{u,m}\right) + 46.4 + 20\log_{10}\left(f_c/5\right) \tag{26}$$

and PL_B1 is expressed as follows:

$$\begin{aligned} PL\_B1(d_{u,m}) = &\left(44.9 - 6.55\log_{10}\left(h_{MS}\right)\right)\log_{10}\left(d_{u,m}\right) \\ &+ 5.83\log_{10}\left(h_{MS}\right) + 14.78 \\ &+ 34.97\log_{10}\left(f_c\right) \end{aligned} \tag{27}$$

where h_{MS} is the device antenna height and h_{BS} is the BS antenna height.

The performance results of the proposed scheme are compared with the one helper based MEC systems in [21] and [23]. In Figures 6 and 7, the proposed system is labeled as *two helpers with NOMA*, whereas the benchmark systems studied in [21] is labeled as *one helper with TDMA* and [23] is labeled as *one helper with direct link*. Moreover, TDMA based cooperative offloading scheme is also performed for the two helpers case, namely, *two helpers with TDMA*. In the TDMA scheme, four time slots are needed to offload the user's tasks to the BS through the two helpers.

The distance between the user and the BS, d, is selected at 750 meters. Accordingly, the cell-center helpers are deployed between $\frac{d}{3}$ and $\frac{2d}{3}$. It is assumed that the helpers cannot be close to the user and the MEC server from a distance of 50 meters. In the proposed system, the distances between the user and the helpers are determined according to TOD results. On the other hand, in the one helper systems in [21] and [23], the helper is located in the middle of the user and BS, which is 375 meters.

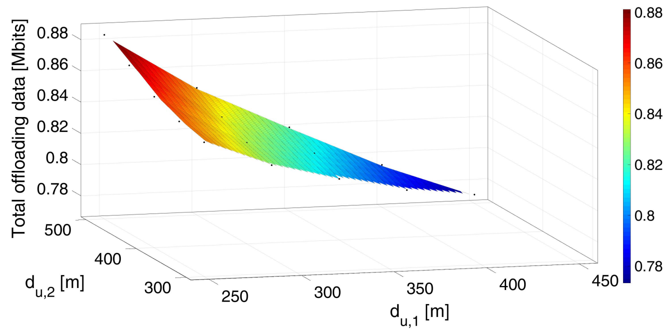

Fig. 4. Average total offloading data versus $d_{u,1}$ and $d_{u,2}$, T = 50ms.

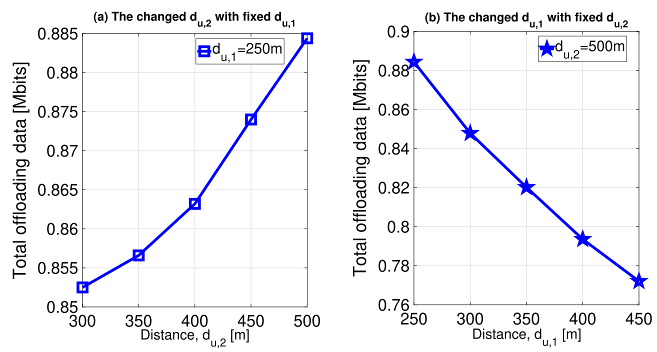

Fig. 5. The average total offloading data versus fixed $d_{u,1}$ and $d_{u,2}$ distance pairs, T = 50ms.

TOD is obtained for various distances using Algorithm 1 by dividing d into 50 meters intervals. In this way, we decide where the helpers should be located to maximize TOD. The total offloading data is shown in Fig. 4 as a function of different choices of $d_{u,1}$ and $d_{u,2}$ for T = 50ms. In the figure, $d_{u,1}$ is the distance between the user and the Helper$_1$, whereas $d_{u,2}$ denotes the distance between the user and the Helper$_2$. According to Fig. 4, the possible distance pairs giving the highest total offloading data are the case of fixed $d_{u,1}$ = 250m or $d_{u,2}$ = 500m which are plotted in detail in Fig. 5.

Fig. 5 illustrates the total offloading data versus the different $d_{u,1}$ and $d_{u,2}$ distance pairs for the proposed system for T = 50ms. The increasing curve in Fig. 5(a) represents $d_{u,1}$ = 250m is fixed and $d_{u,2}$ is changing through the x-axis. On the other hand, the descending curve in Fig. 5(b) represents $d_{u,2}$ = 500m is fixed and $d_{u,1}$ is changing through the x-axis. Fig. 5(a) indicates that the closer the Helper$_2$ is to the BS, the higher the total offloading data. Fig. 5(b) shows that the closer the Helper$_1$ is to the user, the higher the total offloading data. In other words, when the channel gain between the user and the Helper$_1$ becomes stronger in the downlink transmission and the channel gain between the Helper$_2$ and the BS becomes stronger in the uplink transmission, we obtain the maximum total offloading data.

By considering TOD results given in Fig. 5, the optimum distance pair, $\left[d_{u,1}, d_{u,2}\right]$, is chosen to have the maximum total offloading data as [250, 500]. For the remaining part of the simulation results, the optimum distance pair [250, 500] is utilized.

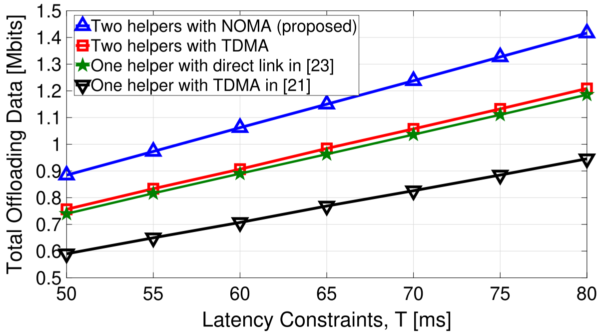

Fig. 6. The average total offloading data versus latency.

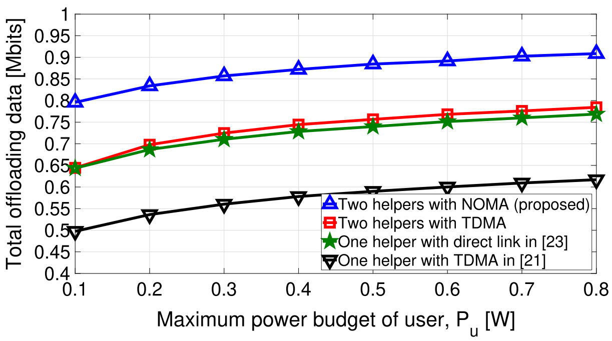

Fig. 7. The average total offloading data versus maximum power budget of user, $\mathrm{T} = 50$ms.

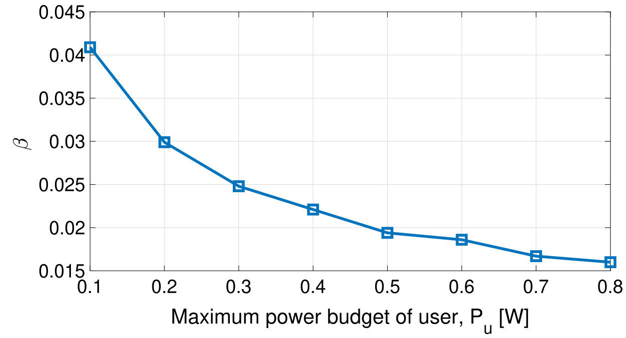

Fig. 8. β versus the maximum power budget of user, $\mathrm{d}_{u,1} = 250$m, $\mathrm{d}_{u,2} = 500$m, and $\mathrm{T} = 50$ms.

In Fig. 6, we provide the total offloading data results considering the latency constraints. The latency, T, is a scalable value depending on applications, in this paper, it is taken as between 50ms and 80ms. Increasing T stretches the latency constraint since the tasks of the user and the helpers can be executed more flexibly. As shown in Fig. 6, the system performance is improved in all systems when T increases since more user tasks are offloaded to the helpers. It is observed that two helpers case with NOMA or TDMA schemes outperform the one helper systems. Furthermore, TOD results demonstrate that the proposed two helpers with NOMA system provides the best performance. This is due to the fact that the user's computation-intensive tasks are distributed between two helpers instead of offloading to the BS directly and two helpers can offload more tasks to the BS using NOMA. The total offloading data in the proposed system increases dramatically comparing with the one helper with direct link in [23] system as 0.14Mbits at $\mathrm{T} = 50$ms and 0.23Mbits at $\mathrm{T} = 80$ms. Besides, the proposed two helpers with NOMA system achieves higher total offloading data performance on around 17% compared to the two helpers with TDMA scheme. The proposed two helpers with NOMA system provides a 0.29Mbits higher TOD at $\mathrm{T} = 50$ms and 0.47Mbits higher TOD at $\mathrm{T} = 80$ms compared to the one helper with TDMA in [21].

We also investigate the effect of the maximum transmit power of the user, P_u, on the total offloading data and the downlink power allocation factor, β, for $\mathrm{T} = 50$ms. Fig. 7 shows that when P_u increases, the total offloading data increases for all schemes due to increasing in the downlink transmission rates, R_m^D. Specifically, the proposed two helpers with NOMA system achieves 24% and 18% improvement in TOD over the one helper with direct link in [23] system by using the same transmit power of the user, $\mathrm{P}_u = 0.1$W and $\mathrm{P}_u = 0.8$W, respectively. Besides, the proposed algorithm with NOMA outperforms its TDMA counterpart by 24% and 16% in TOD for $\mathrm{P}_u = 0.1$W and $\mathrm{P}_u = 0.8$W, respectively. It is shown that the proposed two helpers with NOMA system can provide 60% and 47.3% higher TOD than the one helper with TDMA in [21] for $\mathrm{P}_u = 0.1$W and $\mathrm{P}_u = 0.8$W, respectively.

Fig. 8 illustrates that larger P_u results in a smaller β. The reason is that increasing P_u leads to allocate more power to the Helper_2. Then, the offloading data, $\ell_{u,1}$ and $\ell_{u,2}$ are allocated fairly in the proposed system.

The performance results based on uplink transmit powers, downlink transmit powers, the amount of the offloading data at the user and the helpers side and TOD are given in Table II for the different distance pairs for $\mathrm{T} = 50$ms and $\mathrm{P}_u = 0.5$W. Table II implies that there is an optimization between time constraints; t_h is lower than t_u so that the user can offload more data to the helpers. Furthermore, the uplink transmit powers, P_m^U, are allocated proportionally to the channel gain between the Helper_m and the BS.

In Table II, the distance pair of $[250, 300]$ provides more user data offloaded to the helpers, while less data is offloaded to the BS due to longer distance between the helpers and the BS. This inference is easily comprehended from the optimized task offloading decision factor, $\boldsymbol{\alpha}^*$. Thus, it results in less TOD compared to the optimum distance pair of $[250, 500]$. Furthermore, although the optimized task offloading decision factor, $\boldsymbol{\alpha}^*$ takes the maximum value, TOD is lower for the distance pair of $[450, 500]$. The reason is that since the user has to complete the computation-intensive tasks under the latency constraint, when the helpers are far away from the user, less data will be offloaded to the helpers to meet the latency constraint.

We also discuss the total computing data since it is important to execute the amount of the user's task. This variable indicates that how much of the user's data is cooperatively executed at both the BS and helpers side under the latency constraint. For the fixed optimum distance pair $[250, 500]$ and $\mathrm{T} = 50$ms, the proposed two helpers with cooperation scheme achieves higher total computing data performance on

TABLE II
PERFORMANCE RESULTS FOR DIFFERENT $d_{u,1}$ AND $d_{u,2}$ DISTANCE PAIRS, FOR T = 50ms, $P_u = 0.5$W

$\mathbf{d_{u,1}}$ [m]	$\mathbf{d_{u,2}}$ [m]	$\mathbf{t_u^*}$ [ms]	α_1^*	α_2^*	$\mathbf{P_1^D}$ [dBm]	$\mathbf{P_2^D}$ [dBm]	$\mathbf{P_1^U}$ [dBm]	$\mathbf{P_2^U}$ [dBm]	$\ell_{u,1}$ [Mbits]	$\ell_{u,2}$ [Mbits]	$\ell_{1,o}$ [Mbits]	$\ell_{2,o}$ [Mbits]	**TOD** [Mbits]
250	300	33.5	0.53	0.51	6.3	27	16.5	29	0.33	0.24	0.17	0.12	0.85
250	400	32.4	0.63	0.6	8.3	27	19.4	29	0.33	0.21	0.20	0.13	0.86
250	500	32.5	0.67	0.66	9.9	26.9	22.6	29	0.34	0.19	0.23	0.13	**0.88**
300	350	32.4	0.64	0.62	7.8	27	15.7	29	0.29	0.22	0.18	0.13	0.83
350	400	32.4	0.68	0.67	9	27	14.7	29	0.28	0.20	0.19	0.14	0.81
350	500	33.3	0.69	0.69	10.3	26.9	18.3	29	0.29	0.19	0.20	0.13	0.82
450	500	34.2	0.7	0.7	11.1	26.9	14.8	29	0.26	0.19	0.18	0.13	0.77

around 18% compared to the without cooperation case where the optimized $\boldsymbol{\alpha}^*$ values equal to 1 and the helpers act as a relay. This result indicates that the cooperation becomes more important to execute more user's data.

The performance of TOD is compared through multiple helpers for the cases $M = 2$ and $M = 3$. The optimum distances for $M = 3$ are determined as $d_{u,1} = 250$m, $d_{u,3} = 300$m and $d_{u,2} = 500$m to maximize TOD. Thus, we show that $M = 3$ provides a gain of about 15kbits on TOD compared to $M = 2$ case for T = 50ms. The contribution of adding one helper to the proposed framework is affected by power constraints.

V. CONCLUSION

In this paper, we have proposed the multi-helper NOMA-based cooperative MEC system, where the helpers compute and offload the user's computational tasks. Specifically, we have developed the efficient framework to maximize the total offloading data subject to the latency constraints. Furthermore, the optimum distances of the helpers have been determined to obtain the maximum total offloading data. The importance of cooperation has been discussed through the total computing data which indicates the amount of user's executed data in the system. The simulation results have demonstrated that the proposed system has better performance compared to the one helper systems in terms of the total offloading data. Through the proposed framework, we have shown that both the total offloading and computing data under latency constraints are improved by employing more than one helper in the NOMA-based cooperative MEC system.

REFERENCES

[1] H. Liao *et al.*, "Learning-based context-aware resource allocation for edge-computing-empowered industrial IoT," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4260–4277, May 2020.
[2] L. Zhang, J. Wu, S. Mumtaz, J. Li, H. Gacanin, and J. J. P. C. Rodrigues, "Edge-to-edge cooperative artificial intelligence in smart cities with on-demand learning offloading," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.
[3] F. Wang, J. Xu, and Z. Ding, "Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems," *IEEE Trans. Commun*, vol. 67, no. 3, pp. 2450–2463, Mar. 2019.
[4] Z. Yang, J. Hou, and M. Shikh-Bahaei, "Energy efficient resource allocation for mobile-edge computation networks with NOMA," in *Proc. IEEE Globecom Workshops*, Dec. 2018, pp. 1–7.
[5] H. Sun, F. Zhou, and R. Q. Hu, "Joint offloading and computation energy efficiency maximization in a mobile edge computing system," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3052–3056, Mar. 2019.
[6] F. Wang, J. Xu, and Z. Ding, "Optimized multiuser computation offloading with multi-antenna NOMA," in *Proc. IEEE Global Commun. Conf. (GLOBECOM) Workshop*, Singapore, Dec. 2017, pp. 1–7.
[7] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 375–390, Jan. 2019.
[8] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1875–1879, Dec. 2018.
[9] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for NOMA–MEC offloading," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6207–6211, Jun. 2019.
[10] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "On multiple users scheduling using superposition coding over Rayleigh fading channels," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 733–736, Apr. 2013.
[11] D. Wan, M. Wen, X. Cheng, S. Mumtaz, and M. Guizani, "A promising non-orthogonal multiple access based networking architecture: Motivation, conception, and evolution," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 152–159, Oct. 2019.
[12] Z. Yang, C. Pan, J. Hou, and M. Shikh-Bahaei, "Efficient resource allocation for mobile-edge computing networks with NOMA: Completion time and energy minimization," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7771–7784, Nov. 2019.
[13] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient NOMA-based mobile edge computing offloading," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 310–313, Feb. 2019.
[14] S. Mao, S. Leng, and Y. Zhang, "Joint communication and computation resource optimization for NOMA-assisted mobile edge computing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.
[15] Y. Ye, G. Lu, R. Q. Hu, and L. Shi, "On the performance and optimization for MEC networks using uplink NOMA," in *Proc. IEEE ICC Workshop*, May 2019, pp. 1–6.
[16] Z. Liang, Y. Liu, T.-M. Lok, and K. Huang, "Multiuser computation offloading and downloading for edge computing with virtualization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4298–4311, Sep. 2019.
[17] W. Fan, Y. Liu, B. Tang, F. Wu, and Z. Wang, "Computation offloading based on cooperations of mobile edge computing-enabled base stations," *IEEE Access*, vol. 6, pp. 22622–22633, 2018.
[18] J. Wen, C. Ren, and A. K. Sangaiah, "Energy-efficient device-to-device edge computing network: An approach offloading both traffic and computation," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 96–102, Sep. 2018.
[19] X. Diao, J. Zheng, Y. Wu, and Y. Cai, "Joint computing resource, power, and channel allocations for D2D-assisted and NOMA-based mobile edge computing," *IEEE Access*, vol. 7, pp. 9243–9257, 2019.
[20] Q.-V. Pham, L. B. Le, S.-H. Chung, and W.-J. Hwang, "Mobile edge computing with wireless backhaul: Joint task offloading and resource allocation," *IEEE Access*, vol. 7, pp. 16444–16459, 2019.
[21] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4188–4200, Jun. 2019.

[22] Y. Liu, "Exploiting NOMA for cooperative edge computing," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 99–103, Oct. 2019.

[23] B. Li, F. Si, W. Zhao, and H. Zhang, "Wireless powered mobile edge computing with NOMA and user cooperation," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1957–1961, Feb. 2021.

[24] Y. Pan, C. Pan, K. Wang, H. Zhu, and J. Wang, "Cost minimization for cooperative computation framework in MEC networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3670–3684, Jun. 2021.

[25] X. Han *et al.*, "Reliability-aware joint optimization for cooperative vehicular communication and computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 5437–5446, Aug. 2021.

[26] P. Dai, K. Hu, X. Wu, H. Xing, F. Teng, and Z. Yu, "A probabilistic approach for cooperative computation offloading in MEC-assisted vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, early access, Aug. 28, 2020, doi: 10.1109/TITS.2020.3017172.

[27] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, Apr. 2019.

[28] Y. Huang and Y. Liu, "User cooperation for NOMA-based mobile edge computing," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Dec. 2018, pp. 395–400.

[29] Y. Huang, Y. Liu, and F. Chen, "NOMA-aided mobile edge computing via user cooperation," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2221–2235, Apr. 2020.

[30] N. Glei and R. Belgacem Chibani, "Power allocation for energy-efficient downlink NOMA systems," in *Proc. 19th Int. Conf. Sci. Techn. Autom. Control Comput. Eng. (STA)*, Mar. 2019, pp. 611–613.

[31] T. Manglayev, R. C. Kizilirmak, and Y. H. Kho, "Optimum power allocation for non-orthogonal multiple access (NOMA)," in *Proc. IEEE 10th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2016, pp. 1–4.

[32] F. Zhu, Y. Huang, Y. Liu, and X. Zhang, "Cooperative computation offloading in NOMA-based edge computing," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2020, pp. 208–213.

[33] H. Zuo and X. Tao, "Power allocation optimization for uplink non-orthogonal multiple access systems," in *Proc. 9th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2017, pp. 1–5.

[34] I. Azam, M. B. Shahab, and S. Y. Shin, "User pairing and power allocation for capacity maximization in uplink NOMA," in *Proc. 42nd Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2019, pp. 690–694.

[35] A. Grace, *Optim. Toolbox: For Use With MATLAB: User's Guide*. Natick, MA, USA: Math Works Inc., Nov. 1990.

[36] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006.

[37] G. Lesaja, "Introducing interior-point methods for introductory operations research courses and/or linear programming courses," *Open Oper. Res. J.*, vol. 3, no. 1, pp. 1–12, Sep. 2009.

[38] *Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects*, Standard TR 36.814, 3GPP, 2010.

[39] *Study on LTE Device to Device Proximity Services; Radio Aspects*, Standard TR 36.843, 3rd Generation Partnership Project, 2014.

Saadet Simay Yılmaz (Member, IEEE) received the B.S. degree from the Department of Electrical and Electronics, Izmir Institute of Technology, Izmir, Turkey, the minor degree certificate from the Department of Chemistry, Izmir Institute of Technology, in 2014, and the M.S. degree from the Department of Electrical and Electronics, Izmir Institute of Technology, in 2017, where she is currently pursuing the Ph.D. degree. She is also working as a Research Assistant with the Department of Electrical and Electronics Engineering, Izmir Institute of Technology. During the M.S. program, she worked on the project supported by the Republic of Turkey Ministry of Science, Industry and Technology under SAN-TEZ 0686.STZ.2014 Programme. Her current research interests include massive MIMO systems and mobile edge computing.

Berna Özbek (Senior Member, IEEE) is currently an Associate Professor in telecommunication with the Department of Electrical and Electronics Engineering, Izmir Institute of Technology, Turkey, and also working in the field of wireless communication systems for more than 15 years. She has coordinated one international and four national projects, served as a Consultant for three Eureka-Celtic projects and two national industry driven projects. Under her supervision, 15 master's thesis and two Ph.D. dissertations have been completed. She is supervising three Ph.D. and one master's students and conducting one international project under Horizon 2020-MCRISE Programme from 2018 to 2022. She has published more than 90 peer-reviewed articles, one book, one book chapter, and two patents. Her research interests include interference management, resource allocation, limited feedback links, device-to-device communications, physical layer security, massive MIMO systems, and mmWave communications. She has been awarded as Marie-Curie Intra-European (EIF) Fellow by the European Commission for two years in the project Interference Management Techniques for Multicell Networks in 2010. She is serving as a referee for several international journals and on numerous TPCs for IEEE sponsored conferences, the European Commission, the Turkish Republic of the Ministry of Trade and Industry, and the Scientific and Technological Research Council of Turkey.