IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 December 2022, accepted 21 December 2022, date of publication 27 December 2022, date of current version 2 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232731

RESEARCH ARTICLE

Massive MIMO-NOMA Based MEC in Task Offloading for Delay Minimization

SAADET SİMAY YILMAZ, (Graduate Student Member, IEEE), AND BERNA ÖZBEK, (Senior Member, IEEE)

Electrical and Electronics Engineering Department, Izmir Institute of Technology, 35430 Izmir, Türkiye

Corresponding author: Saadet Simay Yılmaz (simayyilmaz@iyte.edu.tr)

This work was supported by the European Union Horizon 2020, RISE 2018 Scheme (H2020-MSCA-RISE-2018) through the Marie Skłodowska-Curie Grant 823903 (RECENT).

ABSTRACT Mobile edge computing (MEC) has been considered a promising technology to reduce task offloading and computing delay by enabling mobile devices to offload their computation-intensive tasks. Non-orthogonal multiple access (NOMA) is regarded as a promising method of increasing spectrum efficiency, while Massive multiple-input multiple-output (MIMO) can support a larger number of users for simultaneous offloading. These two technologies can effectively facilitate offloading and further improve the performance of MEC systems. In this work, we propose a NOMA and Massive MIMO assisted MEC system for delay-sensitive applications. Our objective is to minimize the overall computing and transmission delay under users' transmit power and MEC computing capability. Through the pairing scheme for Massive MIMO-NOMA, the users with the higher channel gain can offload all their data, while the users with the lower channel gain can offload a portion of their data to the MEC. Performance results are provided regarding to the sum data rate and overall system delay compared with the orthogonal multiple access (OMA)-MIMO based and Massive MIMO (M-MIMO) based MEC systems.

INDEX TERMS MEC, Massive MIMO, NOMA, offloading.

I. INTRODUCTION

Real-time and streaming video traffic, as well as internet of things (IoT) applications, have developed rapidly with the advent of the sixth-generation (6G) era. Augmented reality (AR), mobile online gaming and face recognition are examples of applications that generally have heavy computation needs and strict latency requirements [1]. However, the limited computing capability of mobile devices reduces the quality of the user experience, resulting in excessive delay and power consumption. One of the possible solutions is to enable mobile devices for offloading their computation-intensive tasks to a remote cloud center. However, the existing mobile cloud computing faces some issues, such as high latency due to the long propagation distance, low scalability and an increased burden on fronthaul links due to the centralized deployment of the cloud center [2]. As a result, conventional cloud computing will not be sufficient to meet the communication and computation requirements of 6G networks [3].

Mobile Edge Computing (MEC) has been a key solution for 6G communication systems to mitigate the limitations and concerns of conventional cloud computing. The MEC provides cloud-computing capabilities within the Radio Access Network (RAN) and eliminates the requirement for traffic routed through the core network. By moving the computing and storage features to the edge, MEC also provides a distributed and decentralized service environment characterized by low latency and high rate access [4]. Furthermore, the energy consumption of mobile devices is reduced by offloading computation-intensive tasks to a MEC server for execution [3].

Massive multiple-input multiple-output (MIMO) has the potential to be part of 5G and beyond communication systems. By deploying a large number of antennas at the base station (BS), Massive MIMO serves multiple users using the same time-frequency resource and significantly improves high data rates with simple linear processing as well as the

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Masini.

energy efficiencies of the system. In addition to that, Non-orthogonal multiple access (NOMA) is one of the promising radio access technology for 6G systems, since orthogonal multiple access (OMA) systems have low spectrum efficiency and can only support a limited number of users. In this article, we focus on power-domain NOMA, where multiple users are multiplexed at the same time and frequency with different power levels. The successive interference cancellation (SIC) can be utilized to mitigate this co-channel interference (CCI) [5].

In MEC systems, several users may access the same server for task offloading, which requires massive connectivity and a multiple-access strategy. Therefore, efficient and stable wireless communication is needed to satisfy the seamless task offloading to provide a transmission framework for improving system throughput and reducing overall delay. Due to the real-time processing requirements of MEC systems, it is promising to integrate Massive MIMO and NOMA with MEC.

The studies in [6], [7], [8], [9], [10], [11], [12], and [13] have analyzed single antenna-based MEC systems with the NOMA technology. In [6], sub-channel scheduling, task assignment and power allocation have been investigated for OMA and NOMA based MEC systems to minimize the total energy consumption. In [7], decentralized computation offloading in a NOMA-based MEC system has been examined, where long-term average network computation cost is minimized in terms of power consumption and buffering delay. Similarly, the authors in [8] have provided the energy consumption minimization problem for NOMA-assisted MEC systems, in which multiple tasks with different latency requirements are offloaded to several BSs through NOMA. The authors in [9] have formulated the latency minimization problem in a NOMA-based MEC system under maximum transmission power and maximum energy available for local computing and offloading. The offloading efficiency maximization problem for a clustered NOMA-enabled MEC system has been introduced in [10]. The main aim is to maximize network offloading efficiency with power allocation and computing resource allocation under the delay and transmit power constraints for partial offloading mode. The study in [11] has investigated the task delay minimization problem for NOMA-enabled multi-user MEC networks with the partial offloading policy. However, the delay minimization problem has been solved for a two-user case and a user pairing scheme has not been considered. Similarly, the task offloading time minimization problem has been formulated under transmit power and offloading data rate constraints in multi-user NOMA-enabled MEC systems in [12]. We have studied a multi-helper cooperative MEC system based on NOMA to maximize the total offloading data under the latency and power constraints in our previous work presented in [13]. Nonetheless, all these studies consider single antenna-based offloading with the NOMA.

On the other hand, the work in [14] has considered computation offloading via multi-antenna NOMA to improve multi-user MEC systems' performance, in which the BS has four antennas. The weighted sum-energy minimization problem has been formulated for partial offloading and binary offloading.

The works in [15], [16], [17], [18], [19], [20], and [21] consider Massive MIMO assisted MEC system. In particular, the optimization of energy consumption and the minimization of maximum delay for a Massive MIMO assisted MEC system have been presented in [15] under a maximum transmit power constraint. In [16], energy-efficient beamforming and resource allocation have been studied for multi-access edge computing systems consisting of multi-antenna access points (APs) and single-antenna users. The authors have considered maximizing the energy efficiency (EE) of the MEC system. Similarly, in [17], the minimization of the total energy consumption of the multi-user MEC system has been formulated under the condition of satisfying the minimum delay constraint considering uplink and downlink transmissions. The authors of [18] have considered a single-cell MIMO system with perfect and imperfect channel state information (CSI). They have formulated the minimization of the maximum weighted energy consumption subject to the available computing resources and allowable latency. The following studies have considered the delay minimization for Massive MIMO based MEC systems. Specifically, the authors in [19] have studied joint communication and computation resource allocation problems considering the computing resource at the MEC server, the pilot and data transmission power for a single-cell Massive MIMO based MEC network. The problem is formulated to minimize the maximum offloading delay over multiple users. Moreover, to minimize the maximum offload computing delay for all users, the authors in [20] have presented a single-cell multi-user Massive MIMO-MEC network. In [21], the authors have considered a binary offloading scheme in which the users and BS are equipped with a multi-antenna. The authors have formulated to minimize the overall cost of the weighted sum of energy consumption and time delay by jointly considering offloading decision-making, multi-user MIMO (MU-MIMO) transmit precoding design and computation resource allocation. Although these studies improve MEC system performance through Massive MIMO, they do not consider the NOMA system, which improves spectral efficiency in Massive MIMO-based MEC system.

The previous studies focused on only the NOMA or Massive MIMO-based MEC systems, while we propose a NOMA and Massive MIMO assisted MEC system for delay-sensitive applications to minimize the overall computing and transmission delay for remote computing in which all users in each cluster offload the computing tasks to the MEC server, considering both offloading and computing phases. The main motivation for this paper is to construct an efficient MEC mechanism where not only the users with higher channel gain and computational-intensive tasks but also cell-edge users with lower channel gain can offload their tasks through user pairing, offloading and computation scheme. The study of [22] has provided the Massive MIMO and NOMA-based MEC system to reduce the overall delay among all users without considering the cluster concept, while we apply the same transmission delay for the users in the same cluster in NOMA. In addition to that, a partial offloading scheme has been adopted in [22], while we present an offloading for remote computing scheme.

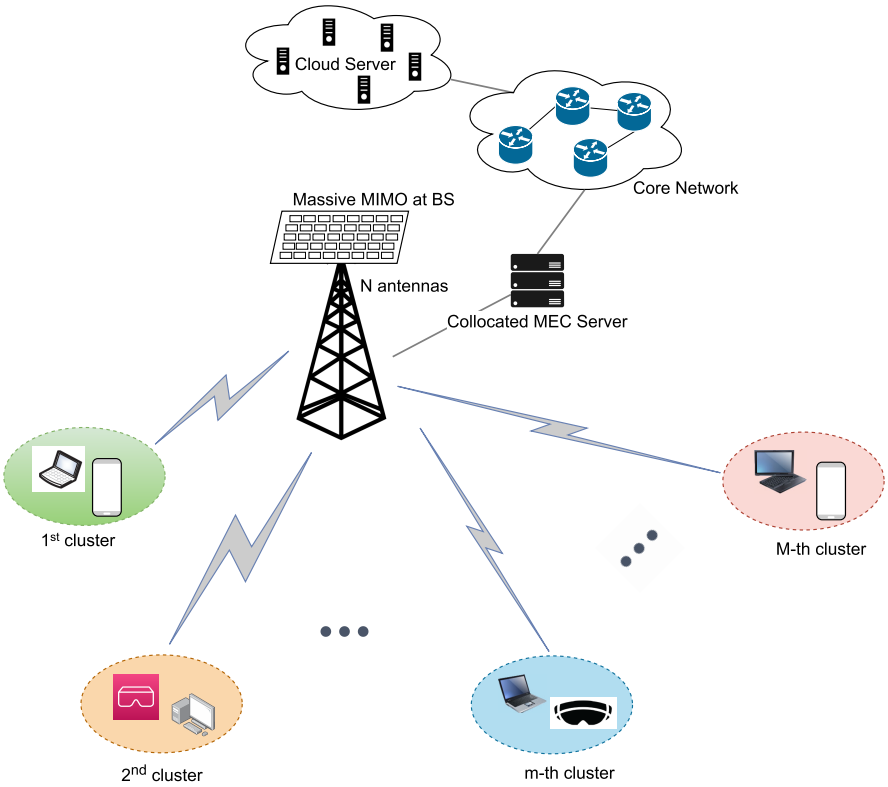

FIGURE 1. Massive MIMO-NOMA assisted MEC system.

To be specific, in this article, we present a framework for a MEC system with Massive MIMO and NOMA technology to demonstrate the advantages of massive connectivity, higher spectral efficiency and lower delay. To this end, multiple users in the system, even the user at the cell-edge, can offload their tasks to the MEC server under an overall delay constraint.

The main contributions of this work can be summarized as follows:

1) We present a Massive MIMO-NOMA assisted MEC system for remote computing. To the best of the authors' knowledge, Massive MIMO-NOMA assisted MEC system by employing an efficient computing scheme to minimize overall delay minimization has not been studied yet.
2) Specifically, the users with relatively higher channel gains are called as strong and the users at the cell-edge are called as weak. While pairing the weak users with the strong ones in each cluster, the weak users can offload a portion of their data to the MEC. In each cluster, the strong user determines the transmission delay to offload its data, while a portion of the weak user's data is also offloaded during this transmission.
3) We formulate the problem of minimizing the overall computing and transmission delay over the Massive MIMO based NOMA system under the computing capability and transmit power constraints. Then, the optimization problem is transformed into a linear problem and solved with a convex optimization tool, i.e., the interior-point method.
4) The simulation results demonstrate the performance gain achieved by the combination of Massive MIMO and NOMA on task offloading in terms of the overall delay and sum data rate compared with the Massive MIMO (M-MIMO) based MEC and OMA based MEC systems.

The rest of the article is organized as follows. Section II presents the uplink Massive MIMO-NOMA assisted MEC framework. We propose the problem formulation followed by the solution in Section III. Simulation results are illustrated and discussed in Section IV. Finally, Section V concludes the article.

II. SYSTEM MODEL

As shown in Fig. 1, we consider an uplink Massive MIMO-NOMA assisted MEC system for remote computing in which a BS is equipped with an antenna array of N elements and serves K single-antenna users, under the assumption that $K << N$. The sets of users and antennas are denoted as $\mathcal{K} = \{1, \ldots, K\}$ and $\mathcal{N} = \{1, \ldots, N\}$, respectively. The available bandwidth for the system is B. In this article, we consider the case of two users in each cluster for NOMA having totally M clusters with $M \leq N$ with $m \in \{1, \ldots, M\}$. Then, the number of clusters, M, is determined as $K/2$. Accordingly, in the proposed system, an efficient user clustering is applied.

We apply the High-High channel gain user pairing strategy for the proposed framework to pair these K users [23]. Specifically, the channel gains of K users are sorted in descending order. Then, these K users are divided into two sets, each including M users. The set of the first M users

with the higher channel gains is defined by the strong user set, while the weak user set includes the set of remaining M users with lower channel gains from $M+1$ to K. Then, the two-user cluster is formed by pairing the users from each set. Accordingly, the first cluster is formed by pairing the first user of the strong user set with the first user of the weak user set, and so on through the High-High channel gain users pairing strategy. Specifically, the first cluster includes the 1st user and $M+1$th user, and the Mth cluster contains the Mth user and Kth user.

Accordingly, the uplink channel vectors belonging to each set are denoted as $\mathbf{h}_{m,j}$ between the BS and the user in the corresponding set of $j=1,2$ for each m^{th} cluster. The channel matrices belonging to the strong user set, $\mathbf{H}_1$, and weak user set, $\mathbf{H}_2$, in all clusters are given by

$$\mathbf{H}_1 = \left[\mathbf{h}_{1,1} \dots \mathbf{h}_{m,1} \dots \mathbf{h}_{M,1}\right] \tag{1}$$

$$\mathbf{H}_2 = \left[\mathbf{h}_{1,2} \dots \mathbf{h}_{m,2} \dots \mathbf{h}_{M,2}\right] \tag{2}$$

The overall $N \times M$ channel matrices $\mathbf{H}_j$, whose element $\mathbf{h}_{m,j}$ with size $N \times 1$ uplink channel vector, can be given as $\mathbf{H}_j = \mathbf{G}_j \ \mathbf{D}_j^{1/2}$ [1], [24]. Here, $\mathbf{D}_j = \text{diag}\left\{L_{1,j}, L_{2,j}, \dots, L_{M,j}\right\} \in \mathbb{R}^{M\times M}$ is a diagonal matrix and represents the large-scale fading component including path loss where $L_{m,j}$ is the path-loss coefficient for the user in the j^{th} set and m^{th} cluster. $\mathbf{G}_j$ is a $N \times M$ matrix including $N \times 1$ vector of $\mathbf{g}_{m,j}$ belonging to each user in the j^{th} set and m^{th} cluster. Each element of $\mathbf{g}_{m,j}$ represents the small-scale Rayleigh fading component of the channel and is modeled independent and identically distributed random variables of $\mathcal{CN}(0,1)$. Thus, we have $\mathbf{h}_{m,j} = \mathbf{g}_{m,j}\sqrt{L_{m,j}}$.

In the proposed framework, the delay mainly includes transmission time and computing time in a remote computing scheme. We focus on the uplink transmission and do not consider the required time to transmit the computing data from the MEC server to the user in the downlink phase.

A. TRANSMISSION SCHEME

The proposed framework includes both offloading phase and computing phase. Firstly, we provide the offloading phase where the users transmit their tasks to the MEC server.

The received signal at the BS can be expressed as:

$$\mathbf{y} = \sum_{m=1}^{M}\sum_{j=1}^{2} \mathbf{h}_{m,j}\sqrt{\alpha_{m,j}}s_{m,j} + \mathbf{n}, \tag{3}$$

where $\alpha_{m,j}$ is the power allocation factor of the user in the j^{th} set and m^{th} cluster within the range $0 < \alpha_{m,j} \leq 1$. The uplink symbol of the user in the j^{th} set and m^{th} cluster is $s_{m,j}$ having $\mathbb{E}[|s_{m,j}|^2] \leq P$ where P represents the maximum transmit power per user. $\mathbf{n}$ represents the $N\times 1$ additive white Gaussian noise (AWGN) with $\mathcal{CN}\left(0,\sigma_n^2\right)$.

We can re-write (3) as:

$$\mathbf{y} = (\mathbf{H}_1\mathbf{s}_1 + \mathbf{H}_2\mathbf{s}_2) + \mathbf{n}, \tag{4}$$

where $\mathbf{s}_1 = \left[\sqrt{\alpha_{1,1}}s_{1,1} \dots \sqrt{\alpha_{m,1}}s_{m,1} \dots \sqrt{\alpha_{M,1}}s_{M,1}\right]^T$ and $\mathbf{s}_2 = \left[\sqrt{\alpha_{1,2}}s_{1,2} \dots \sqrt{\alpha_{m,2}}s_{m,2} \dots \sqrt{\alpha_{M,2}}s_{M,2}\right]^T$ are the $M \times 1$ transmitted signal vectors for the strong and weak user set, respectively.

We use SIC decoding at the BS to extract the users' data in both the strong and weak sets. Firstly, the signals of the strong users are decoded by treating the weak users as interference. After that, SIC is performed by subtracting the signals of strong users from the received signal to decode the weak users' data.

The zero-forcing (ZF) postcoding technique is employed at the BS to mitigate inter-user interference. It is assumed that the BS can have the perfect CSI belonging to all users. In this way, ZF postcoding matrix $\mathbf{W}_j'$ is defined as

$$\mathbf{W}_j' = \mathbf{H}_j^H\left(\mathbf{H}_j\mathbf{H}_j^H\right)^{-1}, \tag{5}$$

where the normalized ZF postcoding matrix is expressed as

$$\mathbf{W}_j = \left[\mathbf{w}_{1,j}{}^T \dots \mathbf{w}_{m,j}{}^T \dots \mathbf{w}_{M,j}{}^T\right]^T, \tag{6}$$

with $\mathbf{w}_{m,j}$ is the ZF postcoder vector with the length $1\times N$ for the user in the j^{th} set and m^{th} cluster. It is given by $\mathbf{w}_{m,j} = \frac{\mathbf{w}'_{m,j}}{\left\|\mathbf{w}'_{m,j}\right\|}$ where $\mathbf{w}'_{m,j}$ is the m^{th} row of $\mathbf{W}'_j$.

The interference between the strong users is eliminated with $\mathbf{W}_1$. Accordingly, the received signal vector of the strong set, $\mathbf{r}_1 = \left[r_{1,1} \ \dots r_{m,1} \dots r_{M,1}\right]^T$, can be expressed as

$$\mathbf{r}_1 = \mathbf{W}_1\mathbf{y} = \mathbf{W}_1\mathbf{H}_1\mathbf{s}_1 + \mathbf{W}_1\mathbf{H}_2\mathbf{s}_2 + \mathbf{W}_1\mathbf{n}. \tag{7}$$

Thus, the received signal of the strong user in the m^{th} cluster is given by

$$r_{m,1} = \mathbf{w}_{m,1}\mathbf{h}_{m,1}\sqrt{\alpha_{m,1}}s_{m,1} + \sum_{i=1}^{M}\mathbf{w}_{m,1}\mathbf{h}_{i,2}\sqrt{\alpha_{m,2}}s_{i,2} + \mathbf{w}_{m,1}\mathbf{n} \tag{8}$$

where the desired signals of the strong user set and the inter-set interference caused by the weak user set are represented in the first term and the second term, respectively.

For the strong user in the m^{th} cluster, the received instantaneous signal to interference plus noise ratio (SINR) is given by

$$\text{SINR}_{m,1} = \frac{\left|\mathbf{w}_{m,1}\mathbf{h}_{m,1}\right|^2\alpha_{m,1}P}{\alpha_{m,2}P\sum_{i=1}^{M}\left|\mathbf{w}_{m,1}\mathbf{h}_{i,2}\right|^2 + \sigma_n^2}. \tag{9}$$

After performing SIC to remove interference from strong users to weak users, the interference between the weak users is eliminated with $\mathbf{W}_2$. Then, the received signal vector of the weak set, $\mathbf{r}_2 = \left[r_{1,2} \ \dots r_{m,2} \dots r_{M,2}\right]^T$, is expressed as

$$\mathbf{r}_2 = \mathbf{W}_2\mathbf{H}_2\mathbf{s}_2 + \mathbf{W}_2\mathbf{n}. \tag{10}$$

Accordingly, the received signal of the weak user in the m^{th} cluster is given by

$$r_{m,2} = \mathbf{w}_{m,2}\mathbf{h}_{m,2}\sqrt{\alpha_{m,2}}s_{m,2} + \mathbf{w}_{m,2}\mathbf{n}. \tag{11}$$

For the weak user in the m^{th} cluster, the received instantaneous SINR is defined by:

$$\text{SINR}_{m,2} = \frac{\left|\mathbf{w}_{m,2}\mathbf{h}_{m,2}\right|^2\alpha_{m,2}P}{\sigma_n^2}. \tag{12}$$

The data rate of the user in the j^{th} set and the m^{th} cluster is hence given by,

$$R_{m,j} = \mathbb{E}\left\{B \log_2\left(1 + \text{SINR}_{m,j}\right)\right\}. \tag{13}$$

Hence, the sum data rate in the system is expressed as

$$R_{\text{sum}} = \sum_{m=1}^{M} \sum_{j=1}^{2} R_{m,j}. \tag{14}$$

III. PROBLEM FORMULATION AND PROPOSED SOLUTION

In this section, we propose the computation scheme for the Massive MIMO-NOMA assisted MEC system. Then, we present the optimization problem and give the solution based on the interior-point algorithm to tackle the problem of minimizing overall delay.

A. COMPUTATION SCHEME

The main objective of the proposed system is to execute the data belonging to strong users under delay constraints while offloading a portion of weak users' data to MEC at the same transmission time. Accordingly, the strong user determines the transmission delay in each cluster, m, to offload its data to the MEC. The total task size for the strong user is initially defined as $D_{m,1}$. The transmission time to offload $D_{m,1}$ for the strong user in the m^{th} cluster is given by,

$$T^t_{m,1} = \frac{D_{m,1}}{R_{m,1}} \tag{15}$$

Since the transmission delay in each cluster is determined by the strong user, we calculate the offloaded data by the weak user as follows:

$$T^t_{m,1}\, R_{m,2} = D_{m,2} \tag{16}$$

When the users' data is received, the MEC server allocates its computing resources to the tasks and the computing is performed for each cluster. The computing time at the MEC server belonging to m^{th} cluster is expressed as:

$$T^c_m = \frac{(D_{m,1} + D_{m,2})\, C_{mec}}{f^{mec}_m} \tag{17}$$

where C_{mec} denotes the number of CPU cycles required to calculate one bit in the MEC server, which is also named the computation intensity. f^{mec}_m denotes the CPU frequency allocated to the m^{th} cluster by MEC.

Finally, the total time to perform the task in the m^{th} cluster is expressed as:

$$T_m = T^t_{m,1} + T^c_m \tag{18}$$

Our aim is to minimize the overall computing and transmission delay in all clusters by jointly optimizing the users' transmit power and MEC computing capacity. Then, we define the optimization problem as follows:

$$\min_{\boldsymbol{\alpha},\mathbf{F}} \max_{m} (T_m) \tag{19}$$

$$\text{s.t.} \quad 0 < \alpha_{m,j} \leq 1, \quad j = \{1,2\}, \quad m \in \{1,\ldots,M\} \tag{19a}$$

$$\sum_{m=1}^{M} f^{mec}_m \leq F_{max} \tag{19b}$$

$$R_{m,2} < R_{m,1}, \quad m \in \{1,\ldots,M\}. \tag{19c}$$

where $\boldsymbol{\alpha} = \left[\alpha_{1,1}, \alpha_{2,1}, \ldots, \alpha_{M,1}, \ \alpha_{1,2}, \alpha_{2,2}, \ldots, \alpha_{M,2}\right]$, $\mathbf{F} = \left[f^{mec}_1, f^{mec}_2, \ldots, f^{mec}_M\right]$ and F_{max} is the total CPU computing capacity at the MEC server. The constraint (19a) shows the range of power allocation factors for each user in the j^{th} set and m^{th} cluster. The constraint (19b) gives the total computing resources. The constraint (19c) ensures that the data rate of strong users in the m^{th} cluster is higher than that of weak users in the same cluster.

In order to reduce the computing time T^c_m, we can use all available CPU computing resources and then the constraint (19b) is re-defined as:

$$\sum_{m=1}^{M} f^{mec}_m = F_{max} \tag{20}$$

Then, we share the computing resources among the clusters equally as follows:

$$f^{mec}_m = \frac{F_{max}}{M} \tag{21}$$

To solve the optimization problem (19), we transform the problem into a minimization problem by introducing an auxiliary variable ξ as follows:

$$\min_{\boldsymbol{\alpha}} \ \xi \tag{22}$$

$$\text{s.t.} \quad \text{(19a) (19c) and (21)},$$

$$T_m \leq \xi, \quad m \in \{1,\ldots,M\}. \tag{22a}$$

Thus, we have transformed the non-convex problem (19) into a convex problem using the auxiliary variable. On this basis, the solution to the problem is found by applying the interior-point method subject to the nonlinear inequalities constraints.

Algorithm 1 outlines the detailed steps of the interior-point method. In the minimization problem, $\mathbf{x}$ is determined as a vector of the following components; $\mathbf{x} = \left[\alpha_{m,j}, \xi\right], \forall m, j$. A feasible solution to Problem (22) is the vector $\mathbf{x}$ satisfying all the constraints. The initial values $\mathbf{x}^0$ are decided by determining the lower and upper bounds range for each component $\alpha_{m,j}$, ξ in $\mathbf{x}$. For this method, the initial values of power allocation factors, $\alpha_{m,j}$, and the overall delay, ξ, are defined to satisfy the constraints.

In this way, an approximate problem $f_\mu(\mathbf{x}, \mathbf{s})$ is defined with a barrier parameter μ as follows:

$$f_\mu(\mathbf{x}, \mathbf{s}) = \xi - \mu \sum_{r} \ln(s_r) \tag{23}$$

where $\mathbf{s} = [s_1, s_2, \ldots, s_K] > 0$ describes the slack variables which convert the inequality constraints to the equality constraints. The slack variables are achieved through the

Algorithm 1 Minimization of the Overall Delay Through Interior-Point Algorithm

Input: $\mathbf{x} = [\alpha_{m,j}, \xi]$, f_m^{mec}, $D_{m,1}$, $\mathbf{h}_{m,j}$ for $j = \{1, 2\}$ and $m \in \{1, \ldots, M\}$.
Output: $\boldsymbol{\alpha}^*$and ξ^*
Initialization Step
1: Slack variables, $\mathbf{s}^0 = [s_1^0, s_2^0, \ldots, s_K^0] > 0$
2: Rearrange (19c) and (22a) as $\mathbf{G} = [G_1, G_2, \ldots, G_K]$
3: Select initial feasible points $\mathbf{x}^0$ as $\mathbf{G}(\mathbf{x}^0) < 0$
4: Choose a barrier parameter, $\mu^0 > 0$ and a convergence tolerance, $\varepsilon > 0$
5: Set u = 0
Main Step

$$\min_{\mathbf{x},\mathbf{s}} \quad f_\mu(\mathbf{x}, \mathbf{s}),$$
$$s.t. \quad G_r(\mathbf{x}) + s_r = 0, \quad r \in \{1, \ldots, K\}$$

6: **while** $\left|f_\mu(\mathbf{x}^u, \mathbf{s}^u) - f_\mu(\mathbf{x}^{u+1}, \mathbf{s}^{u+1})\right| < \varepsilon$ **do**
7: Define the Lagrange function $\mathcal{L}(\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda})$ of $f_\mu(\mathbf{x}, \mathbf{s})$ using Lagrange multipliers $\boldsymbol{\lambda}$, then use the Karush-Kuhn-Tucker (KKT) conditions to solve

$$\mathcal{L}(\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda}) = \xi^u - \mu^u \sum_r \ln(s_r^u) - \lambda_r^u (G_r(\mathbf{x}^u) + s_r^u)$$

8: Solve $f_\mu(\mathbf{x}, \mathbf{s})$ by decreasing μ:
· Starting from $\mathbf{x}^0$, find the point that minimizes $f_\mu(\mathbf{x}, \mathbf{s})$ with an iterative descent method and call them the new variables, $\mathbf{x}^{u+1}$, $\mathbf{s}^{u+1}$ and $\boldsymbol{\lambda}^{u+1}$
9: $\mu^{u+1} = \sigma \mu^u$, where $\sigma \in (0, 1)$
10: $u = u + 1$
11: **end while**

algorithm to guarantee the positiveness of these variables. For any nonlinear inequality constraint, there is a slack variable. The nonlinear inequality constraints in (19c) and (22a) are rearranged as $R_{m,2} - R_{m,1} < 0$ and $T_m - \xi \leq 0$ for $m \in \{1, \ldots, M\}$, respectively. Then, each of these nonlinear inequalities is called $\mathbf{G} = [G_1, G_2, \ldots, G_K]$.

After finding the optimized solutions through Algorithm 1, $\boldsymbol{\alpha}^*$and ξ^*, we calculate the total data of strong users, D_1, and the total data of weak users, D_2, over all clusters as given below, respectively.

$$D_1 = \sum_{m=1}^{M} D_{m,1} \tag{24}$$

and

$$D_2 = \sum_{m=1}^{M} D_{m,2} \tag{25}$$

Thus, the total executed data in the MEC system is given by,

$$D = D_1 + D_2. \tag{26}$$

The complexity of the interior-point method in Algorithm 1 can be determined as $\mathcal{O}\left(\sqrt{n}\,\frac{1}{\varepsilon}\right)$ iterations, where n is the number of variables in the problem, depending on mainly the number of users, K, in the system [25]. Thus, the number of users and the choice of the convergence tolerance, ε, affect the complexity. In the algorithm, convergence tolerance, ε, is selected as 10^{-6}.

OMA-MIMO based MEC scheme: The system performance is compared with the OMA-MIMO based MEC scheme [26] as in Fig. 2. We serve the same number of users within two-time slots and equally share computing resources among the users.

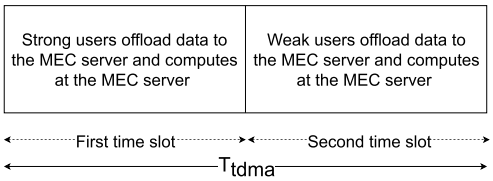

FIGURE 2. Time slot allocation for OMA-MIMO based MEC scheme.

The transmission rates of the m^{th} strong user, $\hat{R}_{m,1}$, and m^{th} weak user, $\hat{R}_{m,2}$, are given below, respectively.

$$\hat{R}_{m,1} = \frac{1}{2}\mathbb{E}\left\{B \log_2\left(1 + \frac{\left|\mathbf{w}_{m,1}\mathbf{h}_{m,1}\right|^2 P}{{\sigma_n}^2}\right)\right\}. \tag{27}$$

$$\hat{R}_{m,2} = \frac{1}{2}\mathbb{E}\left\{B \log_2\left(1 + \frac{\left|\mathbf{w}_{m,2}\mathbf{h}_{m,2}\right|^2 P}{{\sigma_n}^2}\right)\right\}. \tag{28}$$

Thus, the sum data rate for OMA-MIMO based MEC scheme is given by,

$$\hat{R}_{\text{sum}} = \sum_{m=1}^{M}\sum_{j=1}^{2} \hat{R}_{m,j}. \tag{29}$$

Under these circumstances, the transmission time and the computing time are calculated and rearranged. The transmission time to offload $D_{m,1}$ for the m^{th} strong user is given by,

$$\hat{T}_{m,1}^t = \frac{D_{m,1}}{\hat{R}_{m,1}} \tag{30}$$

The computing time at the MEC server to compute data belonging to m^{th} strong user is expressed as:

$$\hat{T}_{m,1}^c = \frac{D_{m,1}\, C_{mec}}{f^{mec}} \tag{31}$$

where $f^{mec} = \frac{F_{max}}{K}$.

Then, the total time for m^{th} strong user in the OMA-MIMO based MEC scheme is given as $\hat{T}_{m,1} = \hat{T}_{m,1}^t + \hat{T}_{m,1}^c$.

Similarly, the transmission time to offload $D_{m,2}$ for the m^{th} weak user is given by,

$$\hat{T}_{m,2}^t = \frac{D_{m,2}}{\hat{R}_{m,2}} \tag{32}$$

The computing time at the MEC server to compute data belonging to m^{th} weak user is expressed as:

$$\hat{T}_{m,2}^c = \frac{D_{m,2}\, C_{mec}}{f^{mec}} \tag{33}$$

The total time for the m^{th} weak user in the OMA-MIMO based MEC scheme is given as $\hat{T}_{m,2} = \hat{T}^t_{m,2} + \hat{T}^c_{m,2}$.

Thus, the overall delay for the OMA-MIMO based MEC scheme is expressed as

$$T_{tdma} = \max_{\forall m} \left\{ \hat{T}_{m,1} \right\} + \max_{\forall m} \left\{ \hat{T}_{m,2} \right\}. \tag{34}$$

Massive MIMO (M-MIMO) based MEC system: The system performance is also compared with the M-MIMO based MEC system. The received instantaneous SINR is defined by:

$$\widetilde{SINR}_k = \frac{|\mathbf{v}_k \, \mathbf{h}_k|^2 \, \alpha_k P}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} \alpha_i P \, |\mathbf{v}_k \mathbf{h}_i|^2 + {\sigma_n}^2}, \quad \forall k \in \mathcal{K}. \tag{35}$$

where $\mathbf{v}_k$ is the normalized ZF postcoder vector with the size $1 \times N$ and the ZF postcoding matrix is $\mathbf{V} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$. The normalized ZF postcoding vector is given as $\mathbf{v}_k = \frac{\mathbf{V}_k}{\|\mathbf{V}_k\|}$, where $\mathbf{V}_k$ is the k^{th} row of $\mathbf{V}$. Here, the channel matrix is $\mathbf{H} = [\mathbf{H}_1 \, \mathbf{H}_2]$ including $\mathbf{h}_k$ that is the k^{th} column of $\mathbf{H}$ with size $N \times 1$ represents either strong or weak user channel vector. α_k is the power allocation factor, which represents either strong or weak user power allocation factor.

Then, the data rate of the k^{th} user is calculated as follows:

$$\widetilde{R}_k = \mathbb{E}\left\{ B \log_2 \left(1 + \widetilde{SINR}_k \right) \right\}, \quad \forall k \in \mathcal{K}. \tag{36}$$

The transmission time to offload D_k for the k^{th} user is given by,

$$\widetilde{T}^t_k = \frac{D_k}{\widetilde{R}_k} \tag{37}$$

where D_k is the total task size which represents either strong or weak user task size.

The computing time at the MEC server to compute data belonging to k^{th} user is expressed as:

$$\widetilde{T}^c_k = \frac{D_k \, C_{mec}}{f^{mec}} \tag{38}$$

Then, the total time for k^{th} user in the M-MIMO based MEC scheme is given as $\widetilde{T}_k = \widetilde{T}^t_k + \widetilde{T}^c_k$.

Thus, the overall delay the M-MIMO based MEC scheme is expressed as

$$T_{m-mimo} = \max_{\forall k \in \mathcal{K}} \widetilde{T}_k. \tag{39}$$

IV. PERFORMANCE RESULTS

This section provides the simulation results to illustrate the performance of the proposed Massive MIMO-NOMA assisted based MEC framework compared with the M-MIMO based MEC and the OMA-MIMO based MEC systems. For the M-MIMO based MEC system, the same transmit powers as in the proposed scheme are used, while for the OMA-MIMO based MEC the power of all users is chosen as P, as in [26]. Consequently, the OMA-MIMO based MEC uses approximately 50% higher transmit power than the proposed M-MIMO NOMA-based and M-MIMO-based MEC schemes.

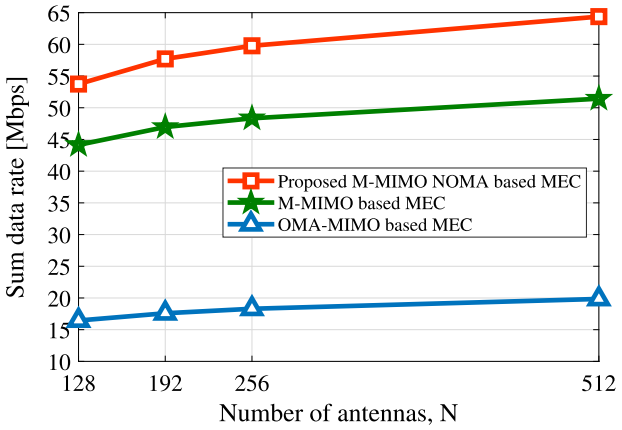

FIGURE 3. The sum data rate versus the number of antennas, N, for K = 16, M = 8 and P = 0 dBm.

The system parameters are given in Table 1. The users are uniformly located in the considered area within a radius of 300 m. The noise power spectral density is −174 dBm/Hz. The path loss is determined by $L_{m,j} = 30.6 + 36.7 \log_{10}(d_{m,j}\,(m))$ where $d_{m,j}$ corresponds to the distance between the user and the BS [15].

TABLE 1. Simulation parameters.

Parameter	Value
K	16
B	1 MHz
P	0 dBm
F_{max}	20 GHz
C_{mec}	100 cycles/bit
$D_{m,1}$	1 Mbits
ξ_{max}	5 sec

Fig. 3 provides the sum data rate of MEC systems for the different numbers of antennas at the BS for K = 16. As shown in the figure, the proposed M-MIMO-NOMA based MEC achieves a higher sum data rate than the M-MIMO based MEC and OMA-MIMO based MEC for all N values. Specifically, for N = 128, the proposed M-MIMO-NOMA based MEC achieves 54 Mbps, while the M-MIMO based MEC provides 44 Mbps and the OMA-MIMO based scheme enables 16 Mbps. Similarly, for N = 512, the proposed M-MIMO-NOMA based MEC achieves around 13 Mbps and 45 Mbps higher sum data rates than its counterparts with the M-MIMO based MEC and the OMA-MIMO based MEC, respectively. The simulation results show the advantages of NOMA in the MEC system based on achievable data rates.

Fig. 4 presents the overall delay versus the number of antennas at the BS for K = 16. The overall delay decreases as the number of antennas increases since the data rates of strong users are increased in the Massive MIMO system, which reduces the transmission delay. Specifically, the proposed MEC with N = 512 reduces the overall delay by 18%

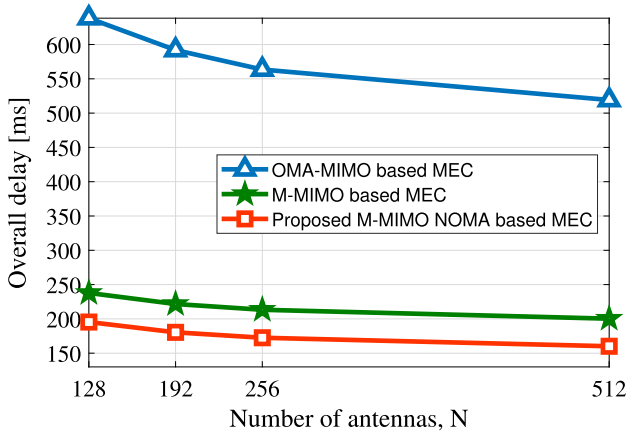

FIGURE 4. The overall delay versus the number of antennas, N, for K = 16, M = 8 and P = 0 dBm.

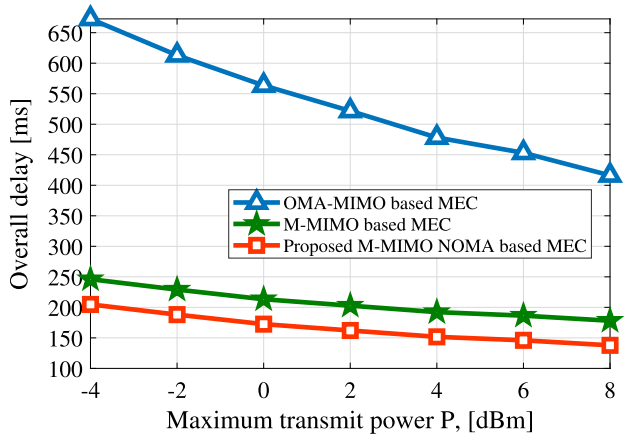

FIGURE 5. The overall delay versus the maximum transmit power, P for N = 256, K = 16, M = 8.

compared with the case of N = 128 by serving multiple users simultaneously. Moreover, the proposed M-MIMO-NOMA based MEC outperforms the M-MIMO based MEC and the OMA-MIMO based MEC in terms of the overall delay for all N values, e.g., the proposed framework reduces the overall delay by 443 ms and 359 ms compared with the OMA-MIMO based MEC for N = 128 and N = 512, respectively. Accordingly, the proposed M-MIMO-NOMA based MEC reduces the overall delay by 42.4 ms and 40.2 ms compared with the M-MIMO based MEC for N = 128 and N = 512, respectively. These performance results confirm the benefits of the proposed joint Massive MIMO and NOMA based MEC system in terms of the overall delay.

In Fig. 5, the overall delay versus the maximum transmit power, P, is shown for N = 256 and K = 16. When the maximum transmit power is increased, the data rate of strong users significantly increases, resulting in a reduction in transmission delay and, thus overall delay. The proposed MEC framework reduces the overall delay by 17% at P = −4 dBm and 23% at P = 8 dBm compared with the M-MIMO based MEC scheme. Similarly, the overall delay of the proposed system is reduced by 70% at P = −4 dBm and 67% at P = 8 dBm compared with the OMA-MIMO based MEC.

TABLE 2. The average transmit power per strong user and per weak user for N and K = 16, M = 8, P = 0 dBm.

N	P_s [dBm]	P_w [dBm]
128	-1.05	-6.56
192	-0.91	-7.24
256	-0.84	-7.63
512	-0.75	-8.08

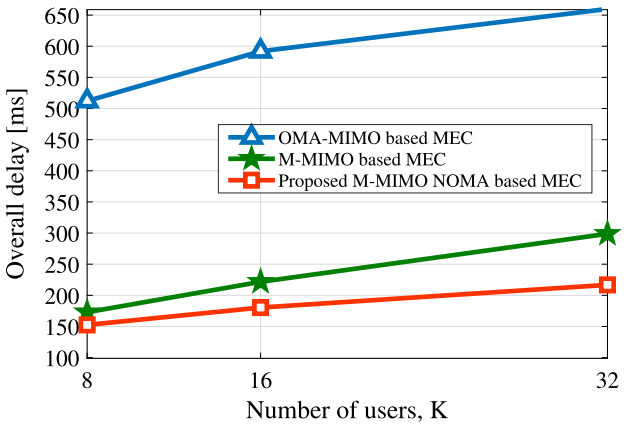

FIGURE 6. The overall delay versus the number of users, K for N = 256 and P = 0 dBm.

Table 2 gives the average transmit power per strong user, P_s, and per weak user, P_w, for the different number of antennas at P = 0 dBm. The average transmit power of strong user is higher than those of weak user due to the given constraint (19c). Furthermore, the average transmit power belonging to strong users increases when the number of antennas is increased, resulting in an increased data rate of the strong users and, thus reducing transmission delay.

Fig. 6 investigates the effect of the number of users, K, on the overall delay for N = 256. With the increasing number of users, the CPU frequency allocated by MEC to the m^{th} cluster for the proposed M-MIMO NOMA based MEC and k^{th} user for the OMA-MIMO based MEC and M-MIMO based MEC systems decreases. This causes an increasing computing delay at the MEC server. Thus, it results in a higher overall delay.

V. CONCLUSION

In this paper, a joint NOMA and Massive MIMO assisted MEC system with a remote computing scheme has been proposed for delay-sensitive 6G applications. We have shown that the combination of NOMA and MEC improves the system performance by simultaneously serving K users with N antennas. By combining Massive MIMO and MEC technologies, more users can offload computational-intensive tasks simultaneously to the MEC while reducing the overall delay. We have formulated the overall computing and transmission delay minimization problem for Massive MIMO-NOMA assisted MEC systems. In this way, the proposed framework enables both cell-center users and cell-edge users to offload their tasks to the MEC server by applying an efficient user pairing, offloading and computation scheme.

The simulation results verify the benefits of the proposed joint Massive MIMO and NOMA with the MEC system. As a future work, the proposed algorithm can be extended to densely deployed scenarios through user pairing algorithms to allocate more than two users for each cluster. Furthermore, the channel impairments and different channel models can be examined in the proposed MEC system.

REFERENCES

[1] M. Zeng, W. Hao, O. A. Dobre, Z. Ding, and H. V. Poor, "Massive MIMO-assisted mobile edge computing: Exciting possibilities for computation offloading," *IEEE Veh. Technol. Mag.*, vol. 15, no. 2, pp. 31–38, Jun. 2020.

[2] 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.

[3] J. Ren, G. Yu, Y. Cai, and Y. He, "Latency optimization for resource allocation in mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5506–5519, Aug. 2018.

[4] R. Malik and M. Vu, "Energy-efficient computation offloading in delay-constrained massive MIMO enabled edge network using data partitioning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6977–6991, Oct. 2020.

[5] N. Madani and S. Sodagari, "Performance analysis of non-orthogonal multiple access with underlaid device-to-device communications," *IEEE Access*, vol. 6, pp. 39820–39826, 2018.

[6] K. Wang, F. Fang, D. B. D. Costa, and Z. Ding, "Sub-channel scheduling, task assignment, and power allocation for OMA-based and NOMA-based MEC systems," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2692–2708, Apr. 2021.

[7] Z. Chen, L. Zhang, Y. Pei, C. Jiang, and L. Yin, "NOMA-based multi-user mobile edge computation offloading via cooperative multi-agent deep reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 1, pp. 350–364, Mar. 2022.

[8] H. Qiu, S. Gao, Y. Chen, and G. Tu, "Energy-efficient rate allocation for NOMA-MEC offloading under outage constraints," *IEEE Commun. Lett.*, vol. 26, no. 11, pp. 2710–2714, Nov. 2022.

[9] A. Tiwari, T. Goyal, and S. Gurugopinath, "Latency minimization in uplink non-orthogonal multiple access-based mobile edge computing," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2020, pp. 1–5.

[10] M. W. Baidas, "Offloading-efficiency maximization for mobile edge computing in clustered NOMA networks," in *Proc. 11th IEEE Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Nov. 2020, pp. 0101–0107.

[11] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal resource allocation for delay minimization in NOMA-MEC networks," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7867–7881, Dec. 2020.

[12] T. Irum, M. U. Ejaz, and M. Elkashlan, "Minimizing task offloading delay in NOMA-MEC wireless systems," in *Proc. 4th Global Power, Energy Commun. Conf. (GPECOM)*, Jun. 2022, pp. 632–637.

[13] S. S. Yilmaz and B. Özbek, "Multi-helper NOMA for cooperative mobile edge computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9819–9828, Jul. 2022.

[14] 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. 2018.

[15] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Delay minimization for massive MIMO assisted mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6788–6792, Jun. 2020.

[16] H. Lim and T. Hwang, "Energy-efficient beamforming and resource allocation for multi-antenna MEC systems," *IEEE Access*, vol. 10, pp. 18008–18022, 2022.

[17] D. Gao, H. Cheng, Z. Han, and S. Yang, "Resource optimization for the multi-user MIMO systems assisted edge cloud computing," in *Proc. IEEE 6th Int. Conf. Signal Image Process. (ICSIP)*, Oct. 2021, pp. 948–953.

[18] N. T. Ti, L. B. Le, and Q. Le-Trung, "Computation offloading in MIMO based mobile edge computing systems under perfect and imperfect CSI estimation," *IEEE Trans. Services Comput.*, vol. 14, no. 6, pp. 2011–2025, Nov. 2021.

[19] T. Huang, Y. Zhang, H. Wu, W. Jiang, C. Yao, M. Xu, and J. Feng, "Joint pilot and data transmission power control and computing resource allocation for the massive MIMO based MEC network," in *Proc. IEEE 19th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2019, pp. 860–865.

[20] W. Feng, J. Zheng, and W. Jiang, "Joint pilot and data transmission power control and computing resource allocation algorithm for massive MIMO-MEC networks," *IEEE Access*, vol. 8, pp. 80801–80811, 2020.

[21] C. Ding, J.-B. Wang, H. Zhang, M. Lin, and J. Wang, "Joint MU-MIMO precoding and resource allocation for mobile-edge computing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1639–1654, Mar. 2021.

[22] D. Li, N. Qin, B. Li, X. Jing, C. Du, and C. Wan, "Resource allocation method based on massive MIMO NOMA MEC on distribution communication network," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 634, no. 1, Feb. 2021, Art. no. 012069.

[23] A. Rauniyar, P. Engelstad, and O. N. Østerbø, "An adaptive user pairing strategy for uplink non-orthogonal multiple access," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, Aug. 2020, pp. 1–7.

[24] A. Khansefid and H. Minn, "Performance bounds for massive MIMO uplink," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 632–636.

[25] 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.

[26] B. Kim, W. Chung, S. Lim, S. Suh, J. Kwun, S. Choi, and D. Hong, "Uplink NOMA with multi-antenna," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–5.

SAADET SİMAY YILMAZ (Graduate Student Member, IEEE) received the B.S. degree from the Department of Electrical and Electronics, the degree 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 a 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 engineering with the Department of Electrical and Electronics Engineering, Izmir Institute of Technology, Türkiye, and working in the field of wireless communication systems for more than 15 years. Under her supervision, 15 master's thesis and two Ph.D. dissertations have been completed. She has been awarded as the Marie-Curie Intra-European (EIF) Fellow by the European Commission for two years, in 2010. She has coordinated one international and four national projects, worked as a Consultant for three Eureka-Celtic projects and two national industry driven projects. She is also supervising three Ph.D. and one master's students, conducting one international project under H2020-MSCA-RISE Programme, from 2018 to 2023, and coordinating one international project under Horizon Europe MSCA-DN Programme, from 2022 to 2027. She has published more than 100 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.