Energy-Efficient Resource Allocation for NOMA-MEC Networks With Imperfect CSI

Fang Fang, Member, IEEE, Kaidi Wang, Member, IEEE, Zhiguo Ding, Fellow, IEEE, and Victor C. M. Leung, Life Fellow, IEEE

Abstract—The combination of non-orthogonal multiple access (NOMA) and multi-access edge computing (MEC) can significantly improve the system performance including communication coverage, spectrum efficiency, etc. In this article, we focus on energy-efficient resource allocation for a multi-user multi-BS NOMA-MEC network with imperfect channel state information (CSI), where each user can upload its tasks to multiple base stations (BSs). We propose an optimization scheme, including task assignment, power allocation and user association, to minimize energy consumption. Specifically, we transform the probabilistic problem into a non-probabilistic one. To efficiently solve this nonconvex energy minimization problem, we first investigate the one-user two-BS case and derive the optimal closed-form expressions of task assignment and power allocation via the bilevel programming method. Subsequently, based on the derived optimal solution, we propose a low complexity algorithm for the user association in the multi-user multi-BS scenario. Simulations demonstrate that the proposed algorithm can yield much better performance than the conventional OMA scheme and the identical results with lower complexity from the exhaustive search with the small number of BSs.

Index Terms—Energy-efficient, imperfect channel state information (CSI), multi-access edge computing (MEC), non-orthogonal multiple access (NOMA), resource allocation.

I. INTRODUCTION

In the last decade, increasing applications and services such as virtual reality (VR), augmented reality (AR), autonomous vehicle and wireless healthcare in Internet of Things (IoT) have emerged in the evolution of wireless communication networks. However, due to the limited communication and storage resources and finite processing capabilities, most devices (e.g., sensors and wearable devices) cannot support ultra-low-latency and high-reliable communications. As a result, multi-access edge computing (MEC) has been proposed as a promising solution to enhance the computing capability of mobile devices with computation-intensive and latency-critical tasks [1], [2]. The performance gains on latency and energy consumption reduction motivated researchers to seamlessly apply MEC into wireless communications [3], [4]. To further improve spectrum efficiency, both non-orthogonal multiple access (NOMA) uplink transmission and NOMA downlink transmission have been proposed to apply in MEC [5]. The integration of NOMA into MEC networks can achieve superior performance on latency and energy consumption reduction over the traditional OMA-MEC system. This motivated researchers to investigate the performance gain of NOMA-MEC networks [6], [7].

A. Related Literature

In MEC networks, the devices with computation-intensive tasks need to offload (download) partial/entire tasks (task results) to (from) the MEC server located in close proximity [8], [9]. Depending on the partitionability and dependence of tasks, the offloading models can be classified into binary offloading and partial offloading. In the binary offloading model, the task cannot be partitioned and needs to be offloaded as an entire task to the MEC server [10]. While in the partial offloading model, the task can be partitioned into multiple tasks, and parts of them can be offloaded to the MEC server for remote executions, then the remaining tasks can be executed locally at mobile devices [11]. The offloaded computation tasks can be executed at the MEC server, usually the base station (BS). Then computation results can be downloaded from the BS to users [12], [13].

Appropriate communication and computation resource optimization can efficiently improve the system performance of NOMA-MEC networks, which attracts extensive researchers to carry out research works on NOMA-MEC. There are two categories according to different objectives: 1. Task delay minimization [14]–[17]; 2. Energy consumption minimization [18]–[26]; Regarding single input and single output (SISO) NOMA-MEC systems, a hybrid NOMA transmission scheme was proposed to minimize delay [14] and energy consumption [18] by considering fully offloading tasks to the MEC server. In particular, the optimal expressions of offloading...
power and time allocation were derived for the hybrid NOMA system, where a user can first offload parts of its tasks by pure NOMA, then offload the remaining by OMA. Both partial offloading [15], [16] and binary offloading [17]–[19] were considered to minimize the task delay in NOMA-MEC networks. In [15], the system overall delay was minimized by an efficient layered algorithm for the NOMA-enabled multi-access MEC network. By adopting partial offloading, the task completion time was minimized by the proposed bisection method based algorithm for multi-user NOMA-MEC networks [16], where the optimal task assignment and power allocation expressions were derived for the two-user case. Besides, binary offloading was considered to minimize the maximum task execution latency by optimizing successive interference cancellation (SIC) ordering and computation resource for NOMA-MEC networks [17]. Regarding the multiple antenna model, binary offloading and partial offloading were both considered in [19], where a Lagrangian-based algorithm and a greed method based algorithm were proposed to minimize the system energy consumption. Furthermore, the future wireless network is expected to achieve massive connectivity, which requires each BS to serve a large number of mobile devices by providing remote executions in NOMA-MEC networks [20], [21]. The energy efficiency optimization problem was investigated in NOMA-MEC networks [22], [23]. In addition, the energy consumption minimization problem was studied for heterogeneous NOMA-MEC networks [24]. From NOMA transmission perspective, there are two main applications of NOMA in MEC, i.e., NOMA uplink transmission and NOMA downlink transmission. The former one indicates that multiple users transmit signals to one single MEC server by using NOMA, which can ensure that multiple users complete their offloading simultaneously. While the latter one refers to the scenario where one user offloads its tasks to multiple MEC servers by using the NOMA protocol. Most existing works focus on NOMA uplink transmission enabled MEC networks while only a handful of research works investigated NOMA downlink transmission in MEC [25]–[27]. Besides, the perfect channel state information (CSI) is difficult to obtain in practice. To fully exploit the benefit of NOMA downlink transmission in MEC, such as supporting multiple servers, in this work, we mainly focus on energy consumption in NOMA downlink transmission enabled MEC.1

B. Motivations and Contributions

Since perfect CSI is challenging to obtain in practice, in this article, we consider imperfect CSI and investigate the resource allocation including task assignment, offloading power allocation and user association for a multi-user multi-BS NOMA-MEC network. It is worth to mention that the proposed resource allocation scheme is centralized. Similar to the emerging cloud radio access network (C-RAN), all BSs are connected by one central unit. The global system information including the estimated CSI, the computation tasks and communication resource at devices and computing resource at BSs can be obtained at the central unit via high-capacity fronthaul links such as fiber links [28]. The resource optimization is implemented in a centralized manner. The centralized unit performs the proposed algorithm to make decisions for users and sends the decisions to BSs, which will broadcast the decisions to all the associated users by one pilot sequence. Considering imperfect CSI, we propose an energy-efficient resource allocation scheme for a multi-user multi-BS network. After transforming the original probabilistic problem into a non-probabilistic problem, we first focus on one user two-BS case and propose the closed-form solution to the energy minimization problem. Subsequently, we focus on the multi-BS scenario, and a low-complexity user association is proposed to group each user with two BSs. Thus the obtained closed-form solution can be applied in each group to minimize the system energy consumption. The detail contributions are listed as follows:

- In this article, we consider a multi-user multi-BS NOMA-MEC network with imperfect CSI. We aim to minimize the energy consumption of the offloading and computation processes by optimizing the transmit power, target offloading data rate and task partition at users. To reduce the decoding complexity at the receivers in NOMA transmission, we first consider that each user can offload its tasks to at most two BSs. The energy consumption minimization problem is formulated as a probabilistic problem, which is nonconvex. To efficiently solve the problem, we first transform the probabilistic problem into a nonprobabilistic one. Specifically, the outage probability constraint is incorporated into the objective function by using non-central chi-square distribution approximations.

- The transformed problem is still nonconvex and challenging to solve. To efficiently solve this problem, we first focus on one-user two-BS case. Considering offloading phase and computing phase at MEC servers, we propose an optimal solution to minimize the system energy via the bilevel programming method. More particularly, the closed-form expressions of the transmit power allocation and task partition are derived by carefully studying and analyzing the monotonicity and convexity of the problem. The optimal solution is concluded in five cases, which significantly reduces the computation complexity of the proposed scheme.

- We obtain some significant insights from the derived optimal solution, which clearly demonstrate the relationships between the optimal offloading schemes, pure NOMA and OMA, in the multi-BS NOMA-MEC network with imperfect CSI. The energy consumption efficiency (ECE) of communicating and computation phases is proposed to present conditions of pure NOMA offloading and OMA offloading. These insights demonstrate practical applications of the proposed resource allocation scheme.

- Base on the optimal solution derived from the two-BS case, we propose a low-complexity algorithm for the user association in the multi-user multi-BS scenario.

1To avoid the confusion between NOMA downlink transmission in MEC and the traditional downlink transmission, we use NOMA transmission to replace NOMA downlink transmission for the rest of this article.
Specifically, by using the obtained closed-form expressions, we design two-sided matching to group each user with two BSs in the multi-user multi-BS network. The optimal power allocation and task partition can be applied in each group, which significantly reduces the complexity of the matching-based user association algorithm compared to the optimal solution obtained through exhaustive searching. It can be shown that for a small number of BSs (i.e., $M = 3$), the proposed algorithm will yield identical results from the exhaustive search.

C. Organization

The rest of the article is organized as follows: In Section II, the system model with the imperfect channel model and problem formulation are introduced. The optimal solution with closed-form expressions is proposed in Section III. In Section IV, an efficient user association algorithm is introduced. Simulation results are presented in Section V, and Section VI concludes the article.

II. SYSTEM MODEL AND CHANNEL MODEL

A. Multi-BS NOMA-MEC System Model

We consider a multi-user multi-BS NOMA-MEC network shown in Fig. 1, where multiple users such as wearable devices and autonomous vehicles can offload their computation-intensive and latency-critical tasks to multiple BSs equipped with MEC servers by NOMA transmission. Each BS and each user are equipped with single antenna. This scenario is applicable in practice, especially for the cell edge users holding the computation-intensive task and limited computing capacity. Autonomous vehicles can offload their computation-intensive tasks to MEC servers by NOMA transmission. Each BS and each user can be associated with as BSs and users, respectively. Denote the index of $N$ users by $n$ and denote user equipment $n$ by UE$_n$. Define the maximum number of BSs that UE$_n$ can be associated with as $M_n$. The set of BSs serving UE$_n$ is denoted by $B_n = \{B_{n,1}, B_{n,2}, \cdots, B_{n,M_n}\}$ where $B_{n,m}$ is the $m$-th BS that UE$_n$ can offload its tasks to, thus we have $M_n \triangleq |B_n|$ where $M \leq M_1 + M_2 + \cdots + M_N$. Note that “$\leq$” indicates that each BS can serve multiple users due to its considerable computing capacity.

The channel gain from UE$_n$ to BS$_{n,m}$ is denoted by $|g_{n,m}|^2 = |h_{n,m}|^2/d_{n,m}^\alpha$ where $h_{n,m} \sim \mathcal{CN}(0, 1)$ is Rayleigh fading coefficient, and $d_{n,m}$ is the distance from UE$_n$ to BS$_{n,m}$, and $\alpha$ is the path loss exponent, and $\mathcal{CN}(0, 1)$ is the complexed Gaussian distribution with mean zero and variance one. Without loss of generality, the channel gains of $M_n$ BSs for UE$_n$ are sorted as $|g_{n,1}|^2 \leq |g_{n,2}|^2 \leq \cdots \leq |g_{n,M_n}|^2$. Assume that channel gains are constant within each transmission block and vary from different blocks. The SIC technology is applied at BSs with a decoding order of increasing order of the channel gains. Each BS can decode and remove the signals from BSs that have been decoded before. Denote the transmit power from UE$_n$ to BS$_{n,m}$ by $p_{n,m}$. The signal received at BS$_{n,m}$ is

$$y_{n,m} = |g_{n,m}|^2 p_{n,m}s_{n,m} + \sum_{i=m+1}^{M_n} |g_{n,m}|^2 p_{n,i}s_{n,i} + z$$

where $s_{n,m}$ is the transmit message from UE$_n$ to BS$_{n,m}$, and $z \sim \mathcal{CN}(0, \sigma_z^2)$ is zero-mean additive white Gaussian noise (AWGN) with variance $\sigma_z^2$. The second term is the interference from other BSs. Let $G_{n,m} = |g_{n,m}|^2$. Given by perfect CSI at BSs and the bandwidth $B$, the maximum achievable offloading data rate from UE$_n$ to BS$_{n,m}$ can be written by

$$C_{n,m} = B \log_2 \left(1 + \frac{G_{n,m}p_{n,m}}{\sum_{i=m+1}^{M_n} G_{n,m}p_{n,i} + 1} \right).$$

B. Imperfect Channel Model

Most previous works assumed that all BSs know the entire knowledge of CSI. However, the perfect CSI is challenging to obtain in practice due to the high complexity of the backhaul signaling overhead. In this article, we investigate the energy consumption minimization by assuming that the small scale fading channel is estimated at BSs. The BSs forward the estimated CSI to the central unit via high-speed fronthaul links for global decision making. In this section, the minimum mean square error (MMSE) channel estimation error model is adopted to describe the small scale fading coefficients $g_{n,m}$. Thus perfect channel gain can be written by

$$g_{n,m} = \hat{g}_{n,m} + \epsilon$$

where $|\hat{g}_{n,m}|^2 = |\hat{h}_{n,m}|^2/d_{n,m}^\alpha$ is the estimated channel gain including small scale fading estimation $\hat{h}_{n,m}$ and large scale fading $d^\alpha$, and $\epsilon \sim \mathcal{CN}(0, \sigma_\epsilon^2)$ is the channel estimation error with mean zero and variance $\sigma_\epsilon^2$. In this work, we assume that the large scale fading factors are perfectly estimated since the path loss and shadowing are slowly varying. Thus $G_{n,m} = |\hat{g}_{n,m}|^2$ is the estimated channel gain from UE$_n$ to BS$_{n,m}$ normalized by $\sigma_\epsilon^2$.

Under imperfect CSI, a channel outage event happens when the instantaneous data rate with perfect CSI drops below the target rate. Define the target rate from UE$_n$ to BS$_{n,m}$ as $R_{n,m}$. The actual channel gains $G_{n,m}$ are random

![Fig. 1. The multi-BS NOMA-MEC network.](image-url)
variables since the estimate error \( \epsilon \) is unknown. Given the target rate \( R_{n,m} \), the outage probability can be defined as
\[
\Pr \left[ C_{n,m} < R_{n,m} | \hat{G}_{n,m} \right],
\]
which indicates the communication from UE\(_n\) to BS\(_{n,m}\) fails when the instantaneous data rate \( C_{n,m} \) drops below the target rate \( R_{n,m} \) given by the estimated channel gain. To guarantee the quality of service (QoS) requirements, we usually limit the outage probability by \( \varepsilon_o \). To incorporate the outage probability into our system performance measurement, we adopt the average outage data rate [29]
\[
\hat{R}_{n,m} = R_{n,m} \Pr \left[ C_{n,m} \geq R_{n,m} | \hat{G}_{n,m} \right]
\tag{4}
\]
where \( \hat{R}_{n,m} \) indicates the minimum of the average data rate successfully received by BS\(_{n,m}\).

III. PROBLEM FORMULATION AND TRANSFORMATION

A. Problem Formulation

In NOMA-MEC, there are three phases to complete the task computation, i.e., task offloading, task computation at BSs and downloading task results from BSs to the user. In this work, we adopt the fully offloading scheme due to the limited power of the battery-powered devices [5], [14], [18]. The downloading transmission from BSs to the user is not considered for the following two reasons: First, the size of task results are generally small [30]–[32], and each BS has more power to transmit the task result than that of the user. As a result, the downloading time is much shorter than the offloading time. Second, the transmission energy consumption minimization at the BSs is more related to resource optimization from the perspective of BSs, such as transmit power allocation and subchannel allocation at BSs. However, in this work, we mainly focus on the resource optimization from the perspective of users including offloading power, offloading data rate and task partition. Therefore, in this work, the optimization is designed to save the processing energy at the MEC server and the offloading energy consumption at users. In general, the task of each user (UE\(_n\)) can be denoted by \( L_n \), which is the input number of bits for the task. \( C_{n,m} \) denotes the number of CPU cycles required to compute one bit of this task at BS\(_{n,m}\). In this system, we assume the task can be divided into several parts, and each user can offload different parts to different BSs for remote executions. The offloading task assignment ratio from UE\(_n\) to BS\(_{n,m}\) is denoted by \( \beta_{n,m} \in [0,1] \) and \( \sum_{m=1}^{M_n} \beta_{n,m} = 1 \). For example, there are two BSs serving UE\(_n\). If the value of offloading bits to BS\(_{n,1}\) is \( \beta_{n,1} L_n \), then the offloading bits to BS\(_{n,2}\) is \( \beta_{n,2} L_n = (1 - \beta_{n,1}) L_n \).

- **Task offloading:** In this phase, UE\(_n\) offloads its partial task \( \beta_{n,m} L_n \) to BS\(_{n,m}\). Given by the target data rate \( R_{n,m} \), then the offloading time to BS\(_{n,m}\) is
\[
T_{n,m} = \frac{\beta_{n,m} L_n}{R_{n,m}}.
\tag{5}
\]

Given by the transmit power \( p_{n,m} \), the energy consumption of offloading \( \beta_{n,m} L_n \) task to BS\(_{n,m}\) is
\[
E_{n,m} = T_{n,m} p_{n,m}.
\tag{6}
\]

- **Remote computation:** In this phase, the offloaded task will be computed by each BS’s MEC server. Based on CPU frequency \( f_{n,m} \) at BS\(_{n,m}\), the computing time at BS\(_{n,m}\) can be written by
\[
T_{n,m} = \frac{\beta_{n,m} L_n C_{n,m}}{f_{n,m}}.
\tag{7}
\]

The computing energy consumption at BS\(_{n,m}\) is
\[
E_{n,m}^c = \kappa_{n,m} \beta_{n,m} L_n C_{n,m} f_{n,m}^2 = \kappa_{n,m} T_{n,m} f_{n,m}^3.
\tag{8}
\]

where \( \kappa_{n,m} \) denotes the effective capacitance coefficient for each CPU cycle of BS\(_{n,m}\).

Since the decoding complexity of the SIC technology in NOMA will exponentially increase with the number of receivers (BSs), we assume that each user can only offload tasks to at most two BSs (\( M_n = 2 \)).\(^2\) We aim to minimize the energy consumption of offloading and computing at MEC servers by optimizing the user association, task assignment, target offloading rate, transmit power. Then the problem can be formulated as
\[
\min_{\{U, \beta, R, P\}} \sum_{n=1}^{N} \sum_{m=1}^{2} (E_{n,m} + E_{n,m}^c)
\tag{9a}
\]

s.t. \( \Pr \left[ C_{n,m} < R_{n,m} | \hat{G}_{n,m} \right] \leq \varepsilon_o \), \( \forall m \), \( \beta_{n,1} \leq 1 \), \( \forall n \), \( p_{n,m} \geq 0 \), \( \forall n, m \), \( \sum_{m=1}^{2} p_{n,m} \leq P_{\text{max}} \), \( \forall n \), \( \frac{\beta_{n,1} L_n}{R_{n,1}} \leq T_{\text{max}} \), \( \forall n, m \), \( \frac{(1 - \beta_{n,1}) L_n}{R_{n,2}} \leq T_{\text{max}} \), \( \forall n, m \)
\tag{9b\text{--}9g}

where \( U \) is users associate policy, \( \beta = [\beta_{n,1}]_{N \times 1}, R = [R_{n,m}]_{N \times 2} \) and \( P = [p_{n,m}]_{N \times 2} \). Constraint (9b) limits the outage probability by \( \varepsilon_o \); constraint (9c) specifies the range of task assignment ratio; constraint (9d) and constraint (9e) describe the range of transmit power and the limitation of total transmit power; constraint (9f) and (9g) describe the delay limitations.

B. Equivalent Data Rate With Imperfect CSI

Problem (9) is nonconvex problem due to the outage constraint (9b). In this section, we aim to incorporate the outage constraint (9b) into the objective function. In the following, we provide an effective way to transform the probabilistic problem to non-probabilistic one. By using the imperfect channel model, the actual instantaneous data rate can be

\(^2\)It turns out that the closed-form solution cannot be obtained when \( M_n > 2 \). To reduce the decoding complexity at BSs and obtain the significant insights, we consider \( M_n = 2 \). Please refer to Section IV for details.
rewritten as

\[ C_{n,m} = B \log_2 \left( 1 + \frac{(\hat{G}_{n,m} + \epsilon) p_{n,m}}{\frac{1}{M} \sum_{i=m+1}^{\infty} (\hat{G}_{n,m} + \epsilon) p_{n,i} + 1} \right). \]  

(10)

In general, the outage probability requirement is low (i.e., \( \varepsilon_o \leq 0.1 \)). Thus the outage constraint can be satisfied with the equality at the optimal point [29]. Therefore, we replace the “\( \leq \)” sign with a “\( = \)” sign in the following transformation [33]. The approximation is proved to be accurate [34]. We introduce the following proposition to derive the target data rate:

**Proposition 1:** If we have the outage constraint:

\[ \Pr \left[ C_{n,m} < R_{n,m}[\hat{G}_{n,m}] = \varepsilon_o, \right. \]

the target data rate can be derived as

\[ R_{n,m} = B \log_2 \left( 1 + \frac{H_{m,n} p_{n,m}}{H_{m,n} \sum_{i=m+1}^{\infty} p_{n,i} + \sigma_z^2} \right), \]

where \( H_{m,n} = -\ln(1 - \varepsilon_o)2 \left( 1 + \frac{\beta_o p_{m,n}}{\sigma_z^2} \right) / (\sigma_z^2 \sigma_{n,m}^2) \).

**Proof:** The derivative proof can be found in Appendix A.

\[ \textbf{C. Problem Transformation} \]

Based on the target rate obtained from Proposition 1, the minimum average offloading data rate incorporated with outage constraint is \( R_{n,m} = (1 - \varepsilon) R_{n,m} \). Therefore, problem (9) can be transformed as

\[ \min_{\{u, \beta, \rho\}} \left\{ \sum_{n=1}^{N} \left( \frac{\beta_{n,1} L p_{1}}{R_{n,1}} + (1 - \beta_{n,1}) L p_{2} \right) \right\} \right. \]

\[ + \kappa_1 \beta_{n,1} L C_{1} f_{1}^2 + \kappa_2 (1 - \beta_{n,1}) L C_{2} f_{2}^2 \]

\[ \text{s.t. (9c) - (9e), } \]

\[ \frac{\beta_{n,1} L}{R_{n,1}} \leq T_{\text{max}}, \forall n, \]

\[ \frac{(1 - \beta_{n,1}) L}{R_{n,2}} \leq T_{\text{max}}, \forall n \]

(13a)

(13b)

(13c)

(13d)

By incorporating the outage constraint into the target rate, problem (13) is nonprobabilistic problem. To solve problem (13), in the following sections, we first obtain the optimal solution for the one-user two-BS case, shown in Section IV. Based on the derived closed-form solution, we propose a low-complexity algorithm for the user association in the multi-BS scenario, shown as Algorithm 1 in Section V.

\textbf{IV. OPTIMAL SOLUTION TO THE ONE-USER TWO-BS CASE} \]

In this section, we focus the one-user and two-BS case. We aim to derive the optimal closed-form solution, in which some significant insights are obtained to help address the user association policy \( U \) in the multi-user multi-BS scenario. In the following, we omit user index \( n \) for simplicity. Thus the energy minimization problem can be rewritten as

\[ \min_{\{\beta, \rho\}} \left\{ \frac{\beta_1 L p_{1}}{R_{1}} + \frac{(1 - \beta_1) L p_{2}}{R_{2}} \right\} \]

\[ + \kappa_1 \beta_1 L C_{1} f_{1}^2 + \kappa_2 (1 - \beta_1) L C_{2} f_{2}^2 \]

\[ \text{s.t. } 0 \leq \beta_1 \leq 1, \]

\[ p_1 \geq 0, \; p_2 \geq 0, \]

\[ p_1 + p_2 \leq P_{\text{max}}, \]

\[ \frac{\beta_1 L}{R_{1}} \leq T_{\text{max}}, \]

\[ \frac{(1 - \beta_1) L}{R_{2}} \leq T_{\text{max}}, \]

(14a)

(14b)

(14c)

(14d)

(14e)

(14f)

Problem (14) is still nonconvex. In order to obtain the globally optimal solution, we transform problem (14) into a programming problem [35]:

\[ \min_{0 \leq \beta_1 \leq 1} \{ g(\beta_1) \} \triangleq \min_{\{p_1, p_2\}} \left( \frac{\beta_1 L p_{1}}{R_{1}} + \frac{(1 - \beta_1) L p_{2}}{R_{2}} \right) \]

\[ + \kappa_1 \beta_1 L C_{1} f_{1}^2 + \kappa_2 (1 - \beta_1) L C_{2} f_{2}^2 \]

\[ \text{s.t. } (14c) - (14f) \]

(15a)

(15b)

where \( g(\beta_1) \) is the inner problem. Given by fixed \( \beta_1 \), the inner problem \( g(\beta_1) \) is the energy minimization problem with respect to \( p_1 \) and \( p_2 \). Problem \( g(\beta_1) \) is challenging to solve due to its nonconvexity. Thus we first analyze its monotonicity and obtain the following proposition:

**Proposition 2:** The energy consumption function is monotonic increasing with \( p_1 \) and \( p_2 \). Therefore, the minimum energy consumption is only achieved when the offloading power equals the minimum value:

\[ p_1^* = \left( 2 A \beta_1 - 1 \right) \left( \frac{1}{H_2} - \left( 2 A (1 - \beta_1) - 1 \right) \right), \]

\[ p_2^* = \frac{2 A (1 - \beta_1) - 1}{H_2}. \]

(16a)

(16b)

where \( A = \frac{L}{(1 - \varepsilon_o) B T_{\text{max}}} \).

**Proof:** The proof can be found in Appendix B.

\textbf{A. Optimal Task Assignment Ratio Derivation} \]

In the following, we will derive the optimal task assignment ratio. Based on the optimal offloading power obtained from Proposition 2, we have

\[ g(\beta_1) = T_{\text{max}} \left( \frac{1}{H_2} 2^A + 2 A \beta_1 \left( \frac{1}{H_2} - \frac{1}{H_1} \right) - \frac{1}{H_1} \right) + \left( \kappa_1 L C_{1} f_{1}^2 - \kappa_2 L C_{2} f_{2}^2 \right) \beta_1 + \kappa_2 L C_{2} f_{2}^2. \]

(17)

Thus the task assignment ratio optimization problem (the outer problem in (15)) can be rewritten by

\[ \min_{\beta_1} g(\beta_1) \]

\[ \text{s.t. } 0 \leq \beta_1 \leq 1, \]

\[ \beta_1 \leq \frac{1}{A} \log_2 \left( \frac{P_{\text{max}} + \frac{1}{H_1} - \frac{1}{H_2}}{\frac{1}{H_1} - \frac{1}{H_2}} \right), \]

(18a)

(18b)

(18c)
To solve this problem, we have the following observations.

**Proposition 3:** $g(\beta_1)$ is convex function with respective to $\beta_1$.

**Proof:**

$$\frac{\partial^2 g}{\partial \beta_1^2} = (A \ln(2))^2 \left( \frac{1}{H_1} - \frac{1}{H_2} \right) 2^A \beta_1 \geq 0. \quad (19)$$

The optimal solution of Problem (18) relies on the feasible region of $\beta_1 \in [\beta_{1,\text{min}}, \beta_{1,\text{max}}]$, which is

$$0 \leq \beta_1 \leq \max \left\{ \hat{\beta}_{1,\text{max}}, 1 \right\} \quad (20)$$

where $\hat{\beta}_{1,\text{max}} = \frac{1}{A} \log_2 \left( \frac{P_{\text{max}} + \frac{1}{H_1} - \frac{1}{H_2} 2^A}{\frac{1}{H_1} - \frac{1}{H_2}} \right)$.

**Lemma 1:** To guarantee the feasibility of problem (18), we have

$$2^A \leq 1 + H_2 P_{\text{max}}. \quad (21)$$

This indicates that the offloading time is less than or equal to the maximum time delay $T_{\text{max}}$ if all the tasks are transmitted to BS1. **Proof:** To guarantee the feasible set, we must have

$$\frac{1}{A} \log_2 \left( \frac{P_{\text{max}} + \frac{1}{H_1} - \frac{1}{H_2} 2^A}{\frac{1}{H_1} - \frac{1}{H_2}} \right) \geq 0. \quad (22)$$

Thus we have $2^A \leq 1 + H_2 P_{\text{max}}$. **■**

To obtain the optimal solution, we have the following theorem for the analysis.

**Theorem 1:** When problem (18) is feasible, due to its convexity, the energy consumption $g(\beta_1)$

(a) strictly decreases with $\beta_1$, when $\frac{\partial g}{\partial \beta_1} < 0$ within the feasible region.

(b) firstly strictly decreases and then strictly increases within feasible region, $\frac{\partial g}{\partial \beta_1} < 0$ when $\beta_1 < \hat{\beta}_1$, and $\frac{\partial g}{\partial \beta_1} > 0$ when $\beta_1 > \hat{\beta}_1$ where

$$\hat{\beta}_1 = \frac{1}{A} \log_2 \left( \frac{\kappa_2 L C_2 f_2^2 - \kappa_1 L C_1 f_1^2}{A \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right)} \right), \quad (23)$$

which is achieved when $\frac{\partial g}{\partial \beta_1} = 0$.

(c) strictly increases with $\beta_1$ when $\frac{\partial g}{\partial \beta_1} > 0$ within the feasible region.

Theorem 1 demonstrates the convexity of the energy consumption on the task assignment and guarantees the uniqueness of the globally optimal energy consumption, which can be illustrated by Fig. 2. In cases (a) and (b) of Theorem 1, the upper bound of $\beta_1$ can be 1 or $\beta_{1,\text{max}}$, then we have two optimal solutions for each case. In case (c) of Theorem 1, there is only one optimal solution since only the lower bound value can affect the optimal solution. As a result, the optimal task assignment ratio can be concluded in the following five cases:

- **Case 1:** $\beta_1^* = 1$, based on the conditions

$$\begin{cases} 2^A \leq 1 + H_1 P_{\text{max}} \\ 2^A \leq \frac{\kappa_2 L C_2 f_2^2 - \kappa_1 L C_1 f_1^2}{A \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right)} \end{cases}. \quad (24)$$

In this case, the energy consumption is decreasing with $\beta_1$, which is illustrated by Fig. 2(a). Thus the minimum energy consumption is achieved when $\beta_1^* = 1$, which is OMA system, where the user only transmits its tasks to the BS1.

- **Case 2:** $\beta_1^* = \hat{\beta}_{1,\text{max}}$, based on the conditions

$$\begin{cases} 1 + H_1 P_{\text{max}} \leq 2^A \leq 1 + H_2 P_{\text{max}} \\ \frac{\kappa_2 L C_2 f_2^2 - \kappa_1 L C_1 f_1^2}{A \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right)} \geq \frac{P_{\text{max}} + \frac{1}{H_1} - \frac{1}{H_2} 2^A}{\frac{1}{H_1} - \frac{1}{H_2}}. \end{cases} \quad (25)$$

In this case, the energy consumption is decreasing with $\beta_1$. Since the upper bound of the feasible region is $\beta_{1,\text{max}}$, the minimum energy consumption can be achieved when $\beta_1^* = \hat{\beta}_{1,\text{max}}$. This is pure NOMA offloading scheme, where the user offloads its tasks to two BSs simultaneously.

- **Case 3:** $\beta_1^* = \hat{\beta}_1$, based on the conditions

$$\begin{cases} 2^A \leq 1 + H_1 P_{\text{max}} \\ 1 \leq \frac{\kappa_2 L C_2 f_2^2 - \kappa_1 L C_1 f_1^2}{A \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right)} \leq 2^A. \end{cases} \quad (26)$$

In this case, the energy consumption first decreases until $\beta_1$ reaches $\hat{\beta}_1$, and then increases with $\beta_1$ until $\beta_1$ reaches its maximum. Since the upper bound of the feasible region is 1, the minimum energy consumption is achieved when $\beta_1^* = \hat{\beta}_1$. This is pure NOMA offloading scheme.
Fig. 3. Optimal task assignment ratio cases.

- **Case 4:** $\beta_1^* = \hat{\beta}_1$, based on the conditions
  \[
  \left\{ \begin{array}{l}
  1 + H_1 P_{\text{max}} \leq 2^A \leq 1 + H_2 P_{\text{max}} \\
  1 \leq \frac{\kappa_2 L C_2 f_2^2 - \kappa_1 L C_1 f_1^2}{A \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right)} \leq \frac{P_{\text{max}} + \frac{1}{H_1} - \frac{1}{H_2}}{\frac{1}{H_1} - \frac{1}{H_2}}. \\
  \end{array} \right.
  \]
  (27)

  Similar to Case 3, the upper bound of the feasible region is $\hat{\beta}_{1,}\text{max}$. Thus the minimum energy consumption is achieved when $\beta_1^* = \hat{\beta}_1$ with the above conditions. This is also pure NOMA offloading scheme.

- **Case 5:** $\beta_1^* = 0$, based on the conditions
  \[
  \left\{ \begin{array}{l}
  2^A \leq 1 + H_2 P_{\text{max}} \\
  \frac{\kappa_2 L C_2 f_2^2 - \kappa_1 L C_1 f_1^2}{A \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right)} \leq 1. \\
  \end{array} \right.
  \]
  (28)

  In this case, the energy consumption increases with $\beta_1$, which is illustrated by Fig. 2(c). Thus the minimum energy consumption is achieved when $\beta_1^* = 0$. This is an OMA system, where the user only transmits its tasks to the BS$_2$.

  The detail derivation can be found in Appendix C. Based on the optimal task assignment ratio, the optimal power allocation scheme can be achieved by closed-form expressions (16).

**B. Remarks and Discussions**

In this section, we present some analysis of the optimal solution for the energy minimization in NOMA transmission assisted MEC networks. Let us first define the energy consumption of each link (Link $m$ denotes the link from the user to BS$_m$) with the task assignment ratio $\beta_1$ as

\[
E_1(\beta_1) = T_{\text{max}} \left( 2^{A\beta_1} - 1 \right) \left( \frac{1}{H_2} \right) - \frac{\kappa_1 L C_1 (f_1)^2}{\ln(2)} + \frac{1}{H_1} - 1 \\
+ \kappa_1 \beta_1 L C_1 (f_1)^2
\]

\[
E_2(\beta_1) = T_{\text{max}} \frac{2^{A\beta_1} - 1}{H_2} + \kappa_2 \beta_1 L C_2 (f_2)^2.
\]

(29)

Then we define energy consumption efficiency (ECE$_m$) of each link ($m = 1, 2$) by the derivatives, which includes energy consumption efficiency of offloading via the link to BS$_m$ and computing at BS$_m$. If ECE$_1 > ECE_2$, then Link 1 will consume more energy than Link 2 given by offloading task assignment ratio $\beta_1$.

\[
ECE_1 = \frac{\partial E_1}{\partial \beta_1} = A \ln(2) 2^{A(1-\beta_1)} + A \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right) 2^{A\beta_1}
\]

\[
+ \kappa_1 LC_1 f_1^2 - \frac{ECE_1}{ECE_2} \leq 1
\]

(30)

where $\bar{A} = T_{\text{max}} A$. ECE$_{m,\text{opt}}$ is the ECE of the offloading link to BS$_m$. ECE$_{c,\text{opt}}$ is the ECE of the computing phase at BS$_m$. To compare the ECE of these two links, we let

\[
ECE_1 - ECE_2 = \bar{A} \ln(2) \left( \frac{1}{H_1} - \frac{1}{H_2} \right) 2^{A\beta_1} - \frac{ECE_1}{ECE_2} \leq \frac{ECE_1}{ECE_2} + \frac{\kappa_2 L C_2 f_2^2 - \kappa_1 L C_1 f_1^2}{ECE_2 - ECE_1}
\]

(31)

where the first term, ECE$_{o}$ = ECE$_{m,\text{opt}}$ - ECE$_{c,\text{opt}}$, is the ECE difference for the offloading phase. The second term, ECE$_c = ECE_1 - ECE_2$, is the ECE difference for the computing phase. From the definition of ECE, we can obtain the following observations: ECE$_o$ is an increasing function of $\beta_1$.

**Remark 1:** The ECE of offloading phase, ECE$_o$, is always positive due to $\frac{1}{H_1} - \frac{1}{H_2} \geq 0$. Thus, ECE$_o$ is a increasing function of $\beta_1$ due to $2^{A\beta_1} \geq 0$.

**Remark 2:** The ECE of computing phase, ECE$_c$, is a constant. ECE$_c$ is positive when BS$_1$ has lower computing ECE than BS$_2$. Otherwise, ECE$_c$ is negative.

Based on Remarks 1 and 2, the optimal solution can be concluded by two offloading schemes OMA system and pure NOMA system. The optimal solution can be interpreted by the ECE concept:
(1) $ECE_1 \leq ECE_2 (\max \{ECEo\}) \leq ECEc$: this case corresponds to the scenario where the computing ECE difference $ECEc$ is higher than the maximum of offloading ECE difference, $ECEo$. This scenario is illustrated in Fig. 3(a). The upper bound of $\beta_1$ can be 1 or $\beta_1_{\text{max}}$, which correspond to the optimal solutions Case 1 or Case 2, respectively. This scenario indicates that the ECE of the link to BS1 is far less than that of the link to BS2. In this case, the user prefers to offload its tasks to BS1 for remote executions to achieve the minimum energy consumption.

(2) $ECE_1 = ECE_2 (\min \{ECEo\}) < ECE < \max \{ECEo\}$: this case corresponds to the scenario in which the computing ECE difference, $ECEc$, is higher than the minimum of offloading ECE difference, $\min \{ECEo\}$ and lower than the maximum of offloading ECE, $\max \{ECEo\}$. This scenario is illustrated in Fig. 3(b). The upper bound of $\beta_1$ can be 1 or $\beta_1_{\text{max}}$, which correspond to the optimal solution Case 3 or Case 4, respectively. This scenario is a pure NOMA offloading system, in which the user offloads its partial task $\beta_1 L$ to BS1 and offload the remaining task $(1 - \beta_1') L$ to BS2.

(3) $ECE_1 \geq ECE_2 (\min \{ECEo\}) \geq ECEc$: this case corresponds to the scenario in which the computing ECE difference, $ECEc$, is less than the minimum of offloading ECE difference, $\min \{ECEo\}$. This is illustrated in Fig. 3(c). This case corresponds to the optimal solution Case 5. This scenario is an OMA system, in which the user prefers to offload all its tasks to BS2 for remote executions to achieve the minimum energy consumption. This also indicates that the ECE of the link to BS2 is far less than that of the link to BS1.

V. USER ASSOCIATION FOR MULTI-BS AND MULTI-USER VIA MATCHING

In Section IV, we obtain the closed-form solution for the one-user two-BS case. In this section, we focus on user association for the multi-user multi-BS scenario. Specifically, we design a two-sided matching scheme to group each user with two BSs in the multi-user and multi-BS network. For each group, the obtained closed-form solution for the two-BS case can be directly and effectively applied to obtain the optimal power allocation, task partition and energy consumption. The computation complexity is significantly reduced due to the derived closed-form expressions. The obtained significant insights from the derived optimal solution clearly demonstrate the conditions for optimal offloading schemes, i.e., pure NOMA and OMA. In the following, to avoid the extremely high complexity of the exhaustive search method, we proposed a low-complexity algorithm for user association via matching theory.

A. Design of User Association Algorithm

Assume that each user associated with two BSs occupies one subchannel. Thus the interference between users can be ignored due to different

### Algorithm 1: Matching Based User Association Algorithm

1. Initialize $\mathcal{M}$ by randomly matching each user to the BS groups.

2. **Swap Matching Phase:**

3. Each UEi searches other user UEj, $\forall j \neq i$ to form the user pair $(UEi, UEj)$.

4. If the user pair $(UEi, UEj)$ is a swap blocking pair, then

5. Swap the matching pair.

6. Update $\mathcal{M} = \mathcal{M}^j$.

7. **end if**

8. Until there is no swap-blocking pair in $\mathcal{M}$.

**resource blocks. Let $\mathcal{B} = \{S_1, S_2, \cdots, S_{CM}\} = \{\{S_1, S_2\}, \{S_1, S_3\}, \cdots, \{S_{M-1}, S_M\}\}$ denote the set of all the subsets of two distinct BSs and $K = \|\mathcal{B}\|$. Let $\mathcal{U} = \{UE1, UE2, \cdots, UEK\}$ denote a set of users. We consider user association as a two-sided matching process between a set of $N$ users and a set $K$ of BS pairs. Therefore, the user association problem via matching ($N \leq K$) can be defined as:

**Definition 1:** A two-sided matching $\mathcal{M}$ is a mapping between the user set $\mathcal{U}$ and the BS pair set $\mathcal{B}$, satisfying the following conditions

1. $\mathcal{M}(UE_n) \in \mathcal{B}$, $\mathcal{M}(S_k) \in \mathcal{U}, \forall n, k$;
2. $|\mathcal{M}(UE_n)| = 1$, $|\mathcal{M}(S_k)| = 1$, $\forall n, k$;
3. $S_k = \mathcal{M}(UE_n) \Leftrightarrow UE_n = \mathcal{M}(S_k)$, $\forall n, k$.

In Definition 1, condition (1) indicates that each user in set $\mathcal{U}$ can be matched with a BS pair in set $\mathcal{B}$, and each BS pair in $\mathcal{B}$ is matched with a user in $\mathcal{U}$; Condition (2) states that each user can be matched with only one BS pair in $\mathcal{B}$ and vice versa; Property (3) implies that if $UE_n$ is matched with $S_k$, then $S_k$ should be matched with $UE_n$.

According to Definition 1, user association optimization is formulated as a two-sided matching problem. We aim to minimize the total energy consumption of the system. We first establish a preference list of users. For any $UE_n \in \mathcal{U}$, $UE_n$ prefers the BS pair $S_k$ rather than $S_k'$ can be expressed as

\[(S_k, \mathcal{M}) \succ_{UE_n} (S_k', \mathcal{M}) \Leftrightarrow EC_{UE_n}(\mathcal{M}) < EC_{UE_n}(\mathcal{M}')\]

(32)

where $EC_{UE_n}$ is the energy consumption for $UE_n$ associated with BSs in $S_k$. In terms of BS pairs, $S_k$ prefers to match with $UE_n$ rather than $UE_{n'}$ is described as

\[(UE_n, \mathcal{M}) \succ_{S_k} (UE_{n'}, \mathcal{M}') \Leftrightarrow EC_{S_k}(\mathcal{M}) < EC_{S_k}(\mathcal{M}')\]

(33)

where $EC_{S_k}(\mathcal{M})$ is the energy consumption of the BS pair $S_k$ matched with user $UE_{n'} = \mathcal{M}(S_k)$.

To guarantee all the users are well matched with BSs, we develop a matching algorithm with low complexity to

---

$^3$Generally, the number of users $N$ is larger than the number of BSs $M$. Therefore, the number of the possible subset including two BSs is $C_M^2$, where $C_M^2$ is the number of all the possible subsets of two distinct elements of $M$ BSs.
achieve a stable solution. We adopt swap matching, which is mathematically described as

**Definition 2:** A swap matching is denoted by

$$\mathcal{M}_n^{"} = \mathcal{M}\setminus\{(\text{UE}_n, S_k)\} \cup \{(\text{UE}_{n'}, S_k), (\text{UE}_{n'}, S_k')\},$$

where $\text{UE}_n = \mathcal{M}(S_k)$, $\text{UE}_{n'} = \mathcal{M}(S_{k'})$, $\text{UE}_n = \mathcal{M}_{n'}^{"}(S_{k})$, and $\text{UE}_{n'} = \mathcal{M}_{n'}^{"}(S_{k'})$.

where $\text{UE}_n$ and $\text{UE}_{n'}$ switch their matched BS pairs while keeping other matched pair in the matching scheme invariant.

In a swap operation, considering their own interests, the user might not be approved by other users. Thus we introduce the concept of swap-blocking pair and then we evaluate the conditions under which the swap operations can be approved.

**Definition 3:** Given matching $\mathcal{M}$ and two users $(\text{UE}_n, \text{UE}_{n'})$ with $\text{UE}_n = \mathcal{M}(S_k)$ and $\text{UE}_{n'} = \mathcal{M}(S_{k'})$, if there exists a swap matching $\mathcal{M}_n^{"}$ such that the energy consumption of these two users gets a decrease, then the swap operation is approved, and $(\text{UE}_n, \text{UE}_{n'})$ is called a swap-blocking pair.

Definition 3 implies that there is a benefit by exchanging the matching user pair $(\text{UE}_n, \text{UE}_{n'})$ and this operation will not hurt the benefit of the other users’ energy consumption. In the matching process, a potential swap blocking pair might be arranged by the scheduler, the scheduler will check if these two users can benefit from exchanging their matched BSs. The users will keep performing approved swap operations until they reach a stable status, which is known as two-sided exchange stable matching, which is defined as

**Definition 4:** $\mathcal{M}$ is a two-sided exchange stable matching if $\mathcal{M}$ is not blocked by any swap blocking pair.

Based on the above definitions, we propose Algorithm 1 to solve the user association problem. In this algorithm, we first initialize the matching scheme $\mathcal{M}$. In the swap matching process, each user will iteratively check whether there are swap blocking pairs in the current matching scheme $\mathcal{M}$. If so, we swap the user pair and update the current matching scheme. The matching process will terminate when there is no swap blocking pair in the current matching.

**B. Complexity Analysis**

For each $\text{UE}_n$, there exist $N-1$ possible $\text{UE}_{n'}$, $n' \neq n$ to do swapping, thus the complexity order is given by $O\left(\frac{N(N-1)}{2}\right)$. Therefore, the total complexity is $O(K^2)$. Compared to the optimal strategy using the exhaustive search, which has a complexity order of $O(K!)$, the computational complexity of the proposed swap matching based algorithm is dramatically decreased.

**VI. SIMULATIONS**

In this section, we evaluate the performance of our proposed scheme for the energy minimization in the downlink NOMA-MEC network. In the simulation setup, users and MEC servers are randomly distributed in a single disc with a radius of 500 m. The bandwidth is $B = 1$ GHz. According to the 3GPP urban path loss model, we set the path loss factor $\alpha = 3.76$ [36]. The AWGN power is $\sigma_n^2 = B N_0$ where $N_0 = -174$ dBm/Hz is the AWGN spectral density. In order to guarantee the communication quality, the outage probability is set to $\varepsilon_o = 0.1$. For simplicity, we use $L$ to represent the length of task bits for each user.

In Fig. 4, the energy consumption comparison of different offloading schemes is provided. We adopt the two benchmarks for the performance comparison: 1. BS$_1$ has the priority for offloading; 2. BS$_2$ has the priority for offloading. We can observe that the system energy consumption decreases when the maximum time $T_{\text{max}}$ increases. The offloading energy consumption of Link 1, including the offloading energy consumption on Link 1 to BS$_1$ and the computing energy consumption at BS$_1$, decreases with $T_{\text{max}}$. In the offloading scheme with the priority of BS$_1$, the user will first offload its tasks to BS$_1$. The link will be used when the task cannot be computed by BS$_1$ within the $T_{\text{max}}$. In this case, the fractional power allocation scheme [37] is adopted to allocate different power to BSs. In other words, the link will be allocated more power when it has better channel gain than the other one.

Fig. 5 describes the energy consumption versus the maximum power with different variances of channel estimation.
error. The energy consumption decreases with the maximum power. As the maximum power grows larger, the energy consumption continues to decrease, but the decreasing slope becomes smaller. This is because the feasible set of the optimization problem increases when the maximum power will increase. However, when it reaches a level, the optimal solution can be achieved without the effects of the maximum power. We call this maximum value the maximum power required to achieve the optimum. Moreover, Fig. 5 shows that the scheme with higher error variance will consume more energy than the scheme with lower error variance.

Fig. 6 depicts the task assignment ratio to BS1 versus the maximum power. The task assignment ratio increases the maximum power, but the increasing rate becomes slower. Moreover, Fig. 6 shows that the user prefers to assign more tasks to BS1 compared with the schemes with higher error variances.

Fig. 7 shows the energy consumption versus the CPU frequency of BS2. We can observe that the offloading task assignment ratio is zero before the value of $f_2$, which corresponds to $ECEc = 0$. Before this point, the best solution is the OMA system. The user only offloads its tasks to BS1. After this point, the user starts to offload the partial task to BS2 as well. The task assignment ratio to BS1 increases with the CPU frequency of BS2 and stays stable after it reaches its optimal value. This corresponds to the pure NOMA offloading scheme. It can also be observed that the scheme with lower error variance has a higher task assignment ratio than the scheme with a higher estimated error variance.

Fig. 8 shows the average task assignment ratio to BS1 versus the CPU frequency $f_2$ at BS2 with different error variances. Radius: 500 m, $L = 3.2 \times 10^7$ bits, $C = 10^3 \times [1, 1]$, $\kappa = 10^{-28} \times [0.8, 1.2]$ and $f_1 = 0.8 \times 10^9$.
scheme in this work can provide lower energy consumption than the proposed power allocation scheme in [28]. This is because the proposed scheme in this article is an optimal solution, while the bi-section search power allocation scheme in [28] is a suboptimal solution by using an approximation to transfer the nonconvex solution into a convex problem. It also shows that the proposed scheme in this work is more energy efficient than that in [28] since the NOMA transmission is applied in this work.

VII. CONCLUSION

In this article, considering the imperfect CSI, we proposed an energy-efficient resource allocation scheme for a multi-user multi-cell NOMA-MEC network. Specifically, based on the estimated channel model, we transformed the probabilistic problem into a nonprobabilistic one by incorporating the outage constraint into the objective function. Subsequently, we derived the closed-form expressions of the task and power allocation to achieve the minimum energy consumption for the two-BS case. The analysis of the optimal solution provides the conditions and optimal solutions of two transmission models: 1. OMA offloading transmission to BSs; 2. Pure NOMA offloading transmission to both BSs. In addition, a low complexity user association algorithm was proposed for the multi-user multi-BS case. Simulation results showed that the proposed solution can achieve better performance than the conventional OMA system, and the proposed user association algorithm can achieve lower complexity than the exhaustive search method. Since the closed-form solution can largely reduce the computation complexity, and the proposed scheme can be efficiently implemented in practice.

APPENDIX

A. Proof of Proposition 1

In this proof, we omit index \( n \) for simplicity. By using (2) and (11), we have

\[
\Pr \left[ B \log_2 \left( 1 + \frac{G_m p_m}{G_m \sum_{i=m+1}^M p_i + 1} \right) < R_m | \tilde{G}_m \right] \\
= \Pr \left[ G_m p_m < \gamma_s \left( G_m \sum_{i=m+1}^M p_i + 1 \right) | \tilde{G}_m \right] \\
= \Pr \left[ G_m < \frac{\gamma_s}{p_m - \gamma_s} \sum_{i=m+1}^M p_i | \tilde{G}_m \right] \\
= F_{G_m} \left( \frac{\gamma_s}{p_m - \gamma_s} \sum_{i=m+1}^M p_i \right)
\]

where \( F_{G_m} (x) \) denotes the cumulative distribution function (CDF) of \( G_m \) and \( \gamma_s = \frac{2^{R_m}}{\gamma} - 1 \). Since \( G_m \) is a non-central Chi-square distributed random variable with degrees of freedom 2. The non-centrality parameter is \( \lambda_m = \frac{G_m \sigma^2 d_m}{\sigma^2 / 2} \).
Thus we have

\[ F_G(x) = 1 - Q_1 \left( \frac{\sigma^2 d_m^2 x}{\sigma^2 / 2} \right) = \varepsilon_\alpha \]  

where \( Q_1(\alpha, b) = \exp \left( -\frac{\alpha^2 + b^2}{2} \sum_{k=0}^{\infty} \frac{(\frac{\alpha}{b})^k I_k(ab)}{k!} \right) \) is the Marcum-Q-function \( Q_1 \) with modified Bessel function \( I_k \) of order \( k \) [33]. The inverse CDF of \( G_m, F_G^{-1}(\varepsilon_\alpha) \), can be used to obtain \( R_m \). The inverse CDF can be evaluated by using a lookup table [29]. To obtain a closed-form result, we approximate non-central chi-square distribution, \( \chi_{G_m}^2(\lambda_m) \), by central chi-square distribution expression, \( \chi_{G_m}^2(0) \), [33]:

\[ \Pr \left[ \chi_{G_m}^2(\lambda_m) < x \right] \approx \Pr \left[ \chi_{G_m}^2(0) < \frac{x}{1 + \lambda_m^2 / 2} \right] \]  

This approximation can be proved accurate when the ratio between the non-central parameter \( \lambda_m^2 \) and degrees of freedom 2 is less than 0.2, which is \( \lambda_m^2 / 2 \leq 0.2 \). It has been shown in [38] that this approximation works well even the estimated error is large. Since central chi-square distribution with freedom 2 is exponential distribution, this approximation can transform the probabilistic constraint into a deterministic closed-form

\[ \Pr \left[ \chi_{G_m}^2(0) < \frac{x}{1 + \lambda_m^2 / 2} \right] = 1 - \exp \left( -\frac{x}{2(1 + \lambda_m^2 / 2)} \right). \]  

According to (36) and (37), we have

\[ 1 - \exp \left( -\sum_{i=m+1}^{\infty} \frac{p_i}{2(1 + \lambda_m^2 / 2)} \right) = \varepsilon_\alpha. \]  

Finally, we derive the target SINR as

\[ \gamma_s = \frac{H_m p_m}{1 + H_m \sum_{i=m+1}^{\infty} p_i} \]  

where \( H_m = -\ln(1-\varepsilon_\alpha)2(1 + \lambda_m^2 / 2)/(\sigma^2 d_m^2) \).

**B. The Optimal Power Derivation**

The energy consumption function of \( p_1 \) and \( p_2 \) can be written by

\[ E(p_1, p_2) = \frac{\beta_1 L p_1}{(1-\varepsilon_\alpha)B \log_2 (1 + \frac{H_1 p_1}{\lambda_1 p_2 + 1})} \]  

\[ + \frac{(1-\varepsilon_\alpha)B \log_2 (1 + H_2 p_2)}{(1-\beta_1)L p_2} \]  

\[ + \kappa_1 \beta_1 L \gamma_1 f_1^2 + \kappa_2 (1-\beta_1)L \gamma_2 f_2^2. \]  

It is difficult to find the optimal solution of problem \( g(\beta_1) \) due to the nonconvexity of the objective function. However, the problem can be equivalently transformed into a bilevel programming problem, with the upper-level variable \( p_2 \), which is given by

\[ \min_{0 \leq p_1 \leq P_{\max}} h(p_1) \triangleq \min_{p_2} E(p_1, p_2) \]  

s.t. \( 0 \leq p_2 \leq P_{\max} - p_1 \),

\[ \frac{(1-\beta_1) L}{(1-\varepsilon_\alpha)B \log_2 (1 + H_2 p_2)} \leq T_{\max}, \]  

\[ \frac{\beta_1 L}{(1-\varepsilon_\alpha)B \log_2 (1 + H_1 p_2)} \leq T_{\max}. \]  

where \( h(p_1) \) is the inner optimization problem with respect to \( p_1 \). Constraints (41c) and (41d) can be rewritten by

\[ \frac{2^{A(1-\beta_1)} - 1}{H_2} \leq p_2 \leq \frac{1}{H_1} \left( \frac{H_1 p_1}{2 A \beta_1 - 1} - 1 \right). \]  

Take the partial derivative of (40) with respect to \( p_2 \), we have

\[ \frac{\partial E}{\partial p_2} = \frac{\beta_1 L p_1}{(1-\varepsilon_\alpha)B \log_2 (1 + \frac{H_1 p_1}{\lambda_1 p_2 + 1})^2} \]  

\[ + \frac{\beta_1 L}{(1-\varepsilon_\alpha)B \log_2 (1 + H_2 p_2) \log_2 (1 + \frac{H_1 p_1}{\lambda_1 p_2 + 1})^{1 \geq 0.2}}. \]  

It can be observed that the first term and the second term in (43) are positive. By using the inequality \( x \ln x \geq x - 1, \forall x > 0 \), we have \( 1 + H_2 p_2 \ln(1 + H_2 p_2) - H_2 p_2 > 0 \). Therefore, the partial derivative of (43), \( \frac{\partial E}{\partial p_2} > 0 \). Therefore, it can be concluded that the energy consumption is monotonically increasing with \( p_2 \), and the optimal \( p_2^{\star} \) can be obtained at its minimum, which is \( p_2^{\star} = \frac{2^{A(1-\beta_1)} - 1}{H_2} \). Then \( h(p_1) = E(p_1, p_2^{\star}) \). Since \( p_2^{\star} \) is not a function of \( p_1 \), we can treat \( p_2^{\star} \) as a constant. The outer optimization problem can be written by

\[ \min_{0 \leq p_1 \leq P_{\max}} h(p_1) \]  

s.t. \( 0 \leq p_1 \leq P_{\max} - p_2^{\star} \)

\[ \frac{\beta_1 L}{(1-\varepsilon_\alpha)B \log_2 (1 + \frac{H_1 p_1}{\lambda_1 p_2^{\star} + 1})} \leq T_{\max}. \]  

Take the partial derivative of (44a) with respect to \( p_1 \), we have

\[ \frac{\partial h}{\partial p_1} = \frac{\beta_1 L}{(1-\varepsilon_\alpha)B \log_2 (1 + \frac{H_1 p_1}{\lambda_1 p_2^{\star} + 1})^2}. \]  

By using the inequality \( x \ln x \geq x - 1, \forall x > 0 \), we have

\[ \left( 1 + \frac{H_1 p_1}{H_1 p_2^{\star} + 1} \right) \ln \left( 1 + \frac{H_1 p_1}{H_1 p_2^{\star} + 1} \right) \geq \frac{H_1 p_1}{H_1 p_2^{\star} + 1}. \]  

It can be verified that \( \frac{\partial h}{\partial p_1} \geq 0 \). Therefore, it can be concluded that the energy consumption is monotonically increasing with \( p_1 \). We have the optimal solution of \( p_1^{\star} \), which is

\[ p_1^{\star} = \left( \frac{2^{A(1-\beta_1)} - 1}{H_2} \left( 2^{A(1-\beta_1)} - 1 \right) + \frac{1}{H_1} \right) . \]
Above all, the energy consumption is monotonically increasing with $p_1$ and $p_2$. Then the minimum energy consumption can be achieved when transmit power equals the minimum required power by the constraints in (15).

### C. Optimal Solution Derivation

According to Lemma 1, the feasible range requirement of problem (14) is $2^A \leq 1 + H_2 P_{\text{max}}$.

Case 1: $g(\beta_1)$ decreases within its feasible region with its upper bound 1, which requires $\beta_{1,\text{max}} \geq 1$ and $\beta_1 \geq 1$. Then we have the conditions (24) for $\beta_1^* = 1$.

Case 2: $g(\beta_1)$ decreases within its feasible region with its upper bound $\beta_{1,\text{max}}$, which requires $0 < \beta_{1,\text{max}} < 1$ and $\beta_1 > \beta_{1,\text{max}}$. Then we have conditions (25) for $\beta_1^* = \beta_{1,\text{max}}$.

Case 3: $g(\beta_1)$ first decreases until $\beta_1$ and then increases. Its feasible region is upper bounded by 1, which requires $\beta_{1,\text{max}} \geq 1$ and $0 < \beta_1 < 1$. Then we have conditions (26) for $\beta_1^* = \beta_{1,\text{max}}$.

Case 4: $g(\beta_1)$ first decreases until $\beta_1$ and then increases. Its feasible region is upper bounded by $\beta_{1,\text{max}}$, which requires $\beta_{1,\text{max}} < 1$ and $0 < \beta_1 < \beta_{1,\text{max}}$. Then we have conditions (27) for $\beta_1^* = \beta_{1,\text{max}}$.

Case 5: $g(\beta_1)$ increases within its feasible region with its lower bound 0, which requires $\beta_1 \geq 0$. Combined with the feasible condition (21), then we have the conditions (24) for $\beta_1^* = 1$.

### References

[1] A. C. Baktir, A. Ozgovde, and C. Ersoy, “How can edge computing benefit from software-defined networking: A survey, use cases, and future directions,” IEEE Commun. Surveys Tuts., vol. 19, no. 4, pp. 2359–2391, 4th Quart., 2017.

[2] T. Taleb, K. Samdanis, B. Mada, H. Finck, S. Dutta, and D. Sabella, “On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration,” IEEE Commun. Surveys Tuts., vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

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

[4] Q.-V. Pham et al., “A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art,” IEEE Access, vol. 8, pp. 116974–117017, 2020.

[5] 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.

[6] Z. Ding, X. Wang, and J. Huang, “Mobile edge computing-enabled 5G vehicular networks: Toward the integration of communication and computing,” IEEE Veh. Technol. Mag., vol. 14, no. 1, pp. 54–61, Mar. 2019.

[7] M. Vaezi, G. A. Aruna Baduge, Y. Liu, A. Arafia, F. Fang, and Z. Ding, “Interplay between NOMA and other emerging technologies: A survey,” IEEE Trans. Cognit. Commun. Netw., vol. 5, no. 4, pp. 900–919, Dec. 2019.

[8] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, “Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks,” IEEE Signal Process. Mag., vol. 31, no. 6, pp. 45–55, Nov. 2014.

[9] P. Mach and Z. Becvar, “Mobile edge computing: A survey on architecture and computation offloading,” IEEE Commun. Surveys Tuts., vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

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

[11] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, “Mobile-edge computing: Partial computation offloading using dynamic voltage scaling,” IEEE Trans. Commun., vol. 64, no. 10, pp. 4268–4282, Oct. 2016.
[35] Y. Xu et al., “Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems,” IEEE Trans. Signal Process., vol. 65, no. 18, pp. 4874–4886, Sep. 2017.
[36] Study on Downlink Multiuser Superposition Transmission (MUST) for LTE (Release 13), document TR 36.859, 3GPP, Mar. 2015.
[37] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, “Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI,” IEEE J. Sel. Areas Commun., vol. 35, no. 12, pp. 2874–2885, Dec. 2017.
[38] P. Chung, H. Du, and J. Gondzio, “A probabilistic constraint approach for robust transmit beamforming with imperfect channel information,” IEEE Trans. Signal Process., vol. 59, no. 6, pp. 2773–2782, Jun. 2011.

Fang Fang (Member, IEEE) received the Ph.D. degree in electrical engineering from The University of British Columbia (UBC), Canada, in 2017. From 2018 to 2020, she was a Research Associate with the Department of Electrical and Electronic Engineering, The University of Manchester, U.K. Since August 2020, she has been an Assistant Professor in wireless communications with the Department of Engineering, Durham University, Durham, U.K. Her current research interests include machine learning for intelligent wireless communications, non-orthogonal multiple access (NOMA), intelligent reflecting surface (IRS), multi-access edge computing (MEC), and green communications.

Dr. Fang has been serving as a technical program committee (TPC) member for IEEE flagship conferences, e.g., IEEE GLOBECOM, IEEE ICC, and IEEE VTC. She received the Exemplary Reviewer Certificate of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2017. She is currently an Associate Editor of IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.

Kaidi Wang (Member, IEEE) received the M.S. degree in communications and signal processing from Newcastle University, U.K., in 2014. He is currently pursuing the Ph.D. degree in wireless communication with The University of Manchester, U.K. From 2015 to 2018, he was with Lancaster University. His current research interests include game theory, convex optimization, machine learning, non-orthogonal multiple access, and mobile edge computing.

Zhiguo Ding (Fellow, IEEE) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications in 2000 and the Ph.D. degree in electrical engineering from Imperial College London in 2005. From July 2005 to April 2018, he was with Queen’s University Belfast, Imperial College, Newcastle University, and Lancaster University. Since April 2018, he has been a Professor in communications with The University of Manchester. From October 2012 to September 2020, he has also been an Academic Visitor with Princeton University. His research interests include 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He received the Best Paper Award in IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship 2012-2014, the Top IEEE TVT Editor 2017, the IEEE Heinrich Hertz Award 2018, the IEEE Jack Neubauer Memorial Award 2018, the IEEE Best Signal Processing Letter Award 2018, and the Web of Science Highly Cited Researcher 2019. He is serving as an Area Editor for IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and Journal of Wireless Communications and Mobile Computing, and was an Editor for IEEE WIRELESS COMMUNICATION LETTERS and IEEE COMMUNICATION LETTERS from 2013 to 2016.

Victor C. M. Leung (Life Fellow, IEEE) is currently a Distinguished Professor of computer science and software engineering with Shenzhen University, China. He is also an Emeritus Professor of electrical and computer engineering and also the Director of the Laboratory for Wireless Networks and Mobile Systems, The University of British Columbia (UBC), Canada. He has coauthored more than 1300 journal/conference articles and book chapters. His research interests include wireless networks and mobile systems.

Dr. Leung is a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. He received the 1977 APEBC Gold Medal, the 1977-1981 NSERC Postgraduate Scholarships, the IEEE Vancouver Section Centennial Award, the 2011 UBC Killam Research Prize, the 2017 Canadian Award for Telecommunications Research, the 2018 IEEE TCSCC Distinguished Technical Achievement Recognition Award, and the 2018 ACM MSWiM Reginald Fessenden Award. His coauthored articles received the 2017 IEEE ComSoc Fred W. Ellersick Prize, the 2017 IEEE SYSTEMS JOURNAL Best Paper Award, the 2018 IEEE CSIM Best Journal Paper Award, and the 2019 IEEE TCSCC Best Journal Paper Award. He is serving on the Editorial Boards for IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE ACCESS, IEEE NETWORK, and several other journals. He is named in the current Clarivate Analytics list of “Highly Cited Researchers.”