IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 8, 2020, accepted December 21, 2020, date of publication January 8, 2021, date of current version January 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049883

Joint Task Offloading and Resource Allocation for Multi-Task Multi-Server NOMA-MEC Networks

JIANBIN XUE AND YANING AN
School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
Corresponding authors: Jianbin Xue (xuejb@lut.edu.cn) and Yaning An (lz_ayn@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61841107 and Grant 61461026.

ABSTRACT By offloading computationally intensive tasks of smart end devices to edge servers deployed at the edge of the network, mobile edge computing (MEC) has become a promising technology to provide computing services for Internet of Things (IoT) devices. In order to further improve the access capability of MEC and increase the spectrum utilization efficiency, in this article, Non-Orthogonal Multiple Access (NOMA) technology is introduced into MEC systems and we study the computing offloading problem of multi-user, multi-task and multi-server through joint optimization of task offloading and resource allocation, we intend to maximize the system's processing capability as an optimization goal. To solve the proposed mixed integer nonlinear programming (MINLP) problem, the objective optimization problem is firstly decoupled into two sub-problems of resource allocation and task allocation. Secondly the resource allocation problem is further decomposed into computation resource optimization and communication resource allocation. For the communication resource allocation, it first fixed power allocation, then the sub-channel allocation problem is regarded as a many-to-one matching problem between sub-channels and users. In addition, we propose a low-complexity sub-optimal matching algorithm for sub-channel allocation to maximize the offloading efficiency. Based on our proposed sub-channel allocation scheme, the transmission power allocation is regarded as a convex optimization problem, which is tackled by Lagrangian multiplier method. Finally, under the condition of resource allocation, the tasks of all end devices (EDs) are allocated. Experimental numerical results show that the proposed scheme can effectively decrease latency and energy consumption of networks, improve system processing capability, and further improve MEC system performance.

INDEX TERMS Mobile edge computing, multi-task multi-server, non-orthogonal multiple access, processing capability, resource allocation, task offloading.

I. INTRODUCTION

The process of social industrialization puts forward high-quality requirements for fast and effective data services and the application of 5G network provides a basic platform for this demand. However, with the popularity of mobile terminals (MTs) such as smart-phones, wearable devices and etc., the increase of massive data and the emergence of diversified services, the development of wireless communication has been affected a lot. Meanwhile the explosive growth of mobile Internet services has generated various emerging mobile applications with huge amount of computation, such as virtual reality (VR), human-computer interaction and big data analysis, which often demand stringent delay and processing requirements. This will bring challenges to MTs with limited battery capacity and computing resources [1]–[3]. Although the cloud center is rich in computing and storage resources, the main problem is resource centralization and the distance between MTs and the cloud is longer, which will lead to the large network delay, high energy consumption and task execution overhead, while sensitive applications (such as electronic medicine) require low delay and small energy consumption [4]. To confront such a real-time challenge, as a distributed computing paradigm, MEC was proposed to solve the above problems by bringing computation and storage close to edge network [5]. MEC can reduce the delay and energy consumption of computing tasks and improve the resource utilization through properly offloading computation-intensive tasks to nearby MEC servers. With its advantages, MEC has been applied

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

in many fields, such as IoT, Internet of Vehicles (IoV) and ultra-dense network *et al.* [6], [7]. In addition, in order to further improve the efficiency and flexibility of computing offloading by MEC, a new multi-access MEC paradigm is proposed, in which MTs can use different wireless access networks to offload computing tasks to multiple edge servers simultaneously [8]–[10]. However, due to data transmission through wireless link, the performance of MEC system mainly depends on the allocation of wireless resources for data transmission and calculation task allocation, which has aroused many scholars' research.

With the rapid development of the IoT, orthogonal Multiple Access (OMA) technology has become difficult to meet the demand of mass MTs for simultaneous access, so how to implement a time-frequency resource block (RB) to carry more MTs has become a new research direction. And NOMA technology [11], [12] comes into being. NOMA allows multiple MTs to use the same RB simultaneously and further apply the successive interference cancellation (SIC) technology to alleviate the MTs' co-channel interference, which can effectively improve resource utilization. A lot of studies have confirmed the potential advantages of NOMA, such as improved system throughput, increased energy spectrum efficiency and reduced latency [13]–[15].

MEC can reduce task execution costs, but MEC server has finite computation capacity relative to cloud center, combined with the advantages of MEC and NOMA, the paper considers a NOMA enabled multi-node MEC system, where EDs utilize NOMA to offload their computing tasks to different edge servers simultaneously. By reasonably allocating computing tasks of EDs and wireless resources in system, the offloading efficiency can be enhanced and the MEC network performance will be further improved. The main contributions of this article are summarized as follows:

(1) We investigate a network scenario with multiple mobile edge server nodes (MSNs) and multiple EDs which have computation-intensive tasks to process, in which each MSN is equipped with MEC server to provide wireless and computing resources, and each ED's task can be divided into parts of any size for local and remote computing.

(2) To improve resource utilization and MEC performance, NOMA technology is adopted for data transmission during task offloading. We consider the constraints of communication resources and wireless resources, the joint optimization problem of task offloading and resource allocation is formulated to maximize the task processing capability of the system.

(3) To cope with the formulated MINLP problem, according to the characteristic of the objective function, we firstly break down original problem into two sub-problems, namely resource allocation (RA) problem and task allocation (TA) problem. Then we can further decompose the RA problem into computation resource and communication resource allocation.

(4) For communication resource allocation, the power allocation among sub-channels is first supposed to be equal, and then the sub-channel allocation problem is regarded as a two-sided matching process between sub-channels and MTs. And we put forward a low complexity sub-optimal matching algorithm for sub-channel allocation. Based on the subchannel allocation result, the transmission power allocation is considered as a convex optimization problem and is solved by using Lagrange multiplier method. Finally, on the basis of resource allocation, the task allocation algorithm is used to solve the TA problem. The computer simulation results indicate that the proposed task offloading and resource allocation scheme improves the MEC system performance.

The remainder of the paper is organized as follows. We introduce the related works in section II and the NOMA-MEC network model in section III. In section IV, we show the formulated optimization problem and decompose the problem. Section V describes the solution for the proposed problem and we show the simulation results in section VI. Finally, the conclusion is given for the paper in section VII.

II. RELATED WORKS

Recently, since MEC has made great breakthroughs in improving quality of experience (QoE), which has aroused many scholars' research in task offloading and resource allocation. In most of the research, they regard delay, energy consumption and system overhead as the important criteria with the constraints of quality of service (QoS) and resources. In a single MEC server scenario, some works focused on decreasing the energy consumption with the constraints of computation resources and delay, the tasks of multiple users are offloaded to an edge server, and the joint optimization problem of offloading decision and resource allocation was studied in binary offloading case [16]–[18]. In [19], the joint sub-channel and power allocation problem in the MEC system based on Orthogonal Frequency Division Multiple Access (OFDMA) was investigated to minimize the delay of each mobile device. Chen *et al.* [20] studied the multi-user offloading problem in a multi-channel wireless environment and regarded the distributed offloading decision problem as a multi-user potential game, and proved the existence of Nash equilibrium. In [21], the authors designed a MEC offloading mechanism to save energy and concurrently meet low latency for a mobile user. Although the mentioned above research about single server has made some achievements in improving the performance of MEC, due to the limited computing capacity of the MEC server, when a large number of terminals request computing offloading, it will cause network congestion and large delay. In order to further improve the offloading efficiency, some researchers have studied the cooperative multi-node resource allocation. Literature [22] proposed an offloading scheme that MTs' additional tasks could be further offloaded to other MEC servers connected to it through the collaboration of multiple MEC servers, to enhance the computing offloading service and improve the revenue of the terminal. Yang *et al.* [23] presented a two-layer architecture consisting of micro base station and macro base station, in which users offload their computing tasks to micro base station (MBS) or MBS relayed them to macro base station to complete task execution, effectively reducing system energy consumption. In [24], K. Cheng *et al.* proposed a computation offloading framework to enable multiple users to offload their computing tasks to multiple MEC servers by jointly optimizing the offloading strategy and radio resource allocation, which assumed the computing resources allocated to each user are fixed. Yang *et al.* [25] proposed a novel offloading framework for the multi-server MEC network assisting mobile users in executing computation-intensive jobs via uplink OFDMA offloading system. A multi-task learning based feedforward neural network (MTFNN) model is designed to resolve the MINLP problem by jointly optimizing offloading decision and computational resource allocation. The simulation results show that the uploading scheme based on the MTFNN model has better performance. It is worth learning from when solving such problems.

The aforementioned research about MEC have had obvious effects in reducing time delay and energy consumption, but the studies of single server and multi-server are all based on OMA-MEC system, the spectrum utilization efficiency is lower, and the user experience cannot be well satisfied while a large number of terminals accessed to request task offloading.

Therefore, facing the deficiencies of previous MEC research, with the advantages of NOMA, many works focused on NOMA-MEC systems to achieve better resource allocation and improve the quality of user experience. M. Zeng *et al.* [26] introduced wireless power transfer (WPT) technology into NOMA-MEC system for energy-efficient computation, and studied task offloading problem to maximize the sum of computing rates of all users. To better improve the ability to access of MEC systems and reduce users' computation overhead, in [27], Zhou *et al.* introduced NOMA into MEC system and investigated a multi-user computation offloading problem, then by fixing the offloading decision iteratively updating the resource allocation and efficiently solve it. In [9] and [10], the authors considered a NOMA-based multi-access MEC IoT system.

The emergence of 5G network brings a huge breakthrough on transmission rate, MEC-enable IoT was proved as a promising solution to reduce the delay of task and save the energy of UEs in some IoT scenarios, such as unmanned aerial vehicle (UAV), autonomous vehicle and Industry IoT *et al.* [28]–[31]. In [29], an edge learning-assisted offloading framework for autonomous driving is proposed to improve the inference accuracy while meeting the latency constraint for autonomous driving. In [30], due to the high inference accuracy and strict delay requirements in the target tracking scenario, and the limited computing resources and energy budget of the UAV, a novel hierarchical deep learning (DL) tasks distribution framework was proposed, where the type of DL task is offloaded to the MEC server, and further improve the accuracy of reasoning. In [31], Z. Zhao *et al.* investigated a communication and computation problem for industrial IoT networks. To enhance the system performance, a three-hierarchical optimization framework is proposed to reduce the latency and energy consumption by jointly optimizing bandwidth allocation, offloading, and relay selection.

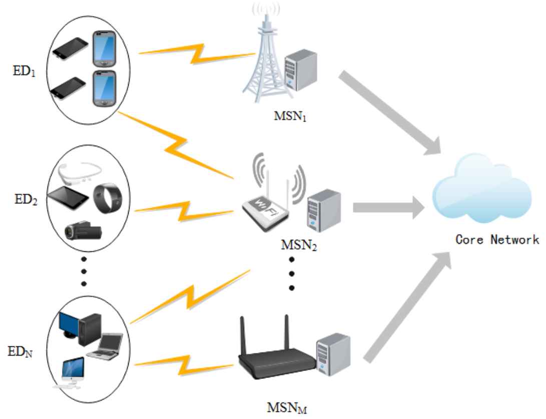


FIGURE 1. Network model of the system.

By summarizing the research of NOMA in MEC, we can conclude that the combination of NOMA and MEC has made progress in meeting user requirements and improving user experience. Different from some studies, we are committed to propose an extensive computation offloading solution for the multi-user multi-task and multiple servers NOMA enabled MEC system by jointly optimizing the task allocation and resources allocation.

III. SYSTEM MODEL

A. NETWORZK MODEL

In this article, we consider a heterogeneous NOMA-MEC network shown as Fig. 1, which consists a number of MSNs with different storage and computing capabilities to provide offloading services for multiple EDs. These nodes are mainly composed of base stations, wireless access points, wireless routers, etc., and each node equipped with a MEC server. To increase spectrum utilization, all MSNs share spectrum resources, the system spectrum equally divides into a set of sub-channels denoted as $\mathcal{SC} = \{1, \ldots, K\}$ and denote $k \in \mathcal{SC}$ as sub-channel k. To facilitate analysis, we assume a quasi-static network, this assumption has been widely used in [27], [36]. Table 1 shows the main notations to be used in the paper.

We denote the set of MSNs by $\mathcal{M} = \{1,\ldots,\mathrm{M}\}$ and denote $m \in \mathcal{M}$ as MSN m, the set of EDs by $\mathcal{N} = \{1, \ldots, N\}$ and $n \in \mathcal{N}$ as ED n, each ED n has a task CT_n, which can be expressed by $< D_n, f_n^l, B_n >$, where D_n denotes the size of the input data, f_n^l denotes the local computing capacity of ED n, and B_n denotes the number of CPU cycles required for computing one bit of task of ED n. We assume the input task can be into sections of any size to execute paralleled at the EDs and MEC servers [32]. We supposed all EDs have J tasks offloaded to MSNs for computation, and denote the set of offloading tasks as $\mathcal{J} = \{1, 2, \ldots J\}, J \leq N \times M$, we denote $D_{nm} \in \mathcal{J}$ as ED n offloads the tasks to MSN m.

TABLE 1. Notations.

Notation	Definition
$\mathcal{N}$	The set of EDs
N	Total number of EDs
$\mathcal{M}$	The set of MSNs
M	Total number of MSNs
$\mathcal{SC}$	The set of sub-channels
K	Total number of sub-channels
D_n	The input data size of the task CT_n
f_n^l	The computing capacity of ED n
B_n	The number of CPU cycles required to compute a bit task
$\mathcal{J}$	The set of offloading tasks of EDs
D_{nm}	The task size ED n offloads to MSN m
s_{nm}^k	Sub-channel allocation variable
J_{max}	Maximum number of offloading tasks of each sub-channel
p_{nm}^k	The transmit power from ED n to MSN m
g_{nm}^k	The channel power gain from ED n to MSN m
σ^2	The power of the additive white Gaussian noise
r_{nm}	Total sum rate from ED n to MSN m
θ_{nn}	The task partial factor of ED n
θ_{nm}	The proportion when ED n offloads tasks to MSN m
T_{nn}^l	The required computing time of task $\theta_{nn}D_n$ on ED n
f_{mn}^e	Computing resources allocated by the MSN m
T_{nm}^{off}	Transmission delay that task $\theta_{nm}D_n$ is offloaded to MSN m
T_{nm}^{exe}	Delay of remote computation on MSN m
T_n	Time to complete the task CT_n of ED n
C_n	Task processing capacity of ED n
C	Overall task processing capability of the whole system
R_{nm}	Achievable rate of ED n offloading task $\theta_{nm}D_n$ to MSN m
R	The total offloading rate of all EDs
F_m	Maximum computing capacity of MSN m

B. COMMUNICATION MODEL

We first introduce the wireless transmission model in this system. In the uplink, each ED sends the signals that are superimposed together by NOMA respectively. we assume SIC receiver is implemented at the MSNs for receiving end. On each channel, according to the order of EDs' channel gains, while the MSN with small channel gain is decoded, the higher channel gain is regarded as interference [33]. Specifically, through continuous decoding and reshaping, the signal with poor channel quality is first demodulated, and we subtract it from the entire superimposed signal interfering signal, and then decode signal with the second poor channel quality, and so on, until all the signals are separated. So the signal with the best channel quality is not interfered by others in the same NOMA cluster, but the MSN with the worst channel quality is interfered by all other MSNs in the cluster. Without loss of generality, s_{nm}^k is denoted as channel allocation variable, if ED n offloads its tasks to MSN m through subchannel k, $s_{nm}^k = 1$, otherwise, $s_{nm}^k = 0$. We assume that a subchannel can be occupied by multiple offloading tasks, and each task occupies at most one sub-channel, so there are the following constraints

$$\sum_{n=1}^{N}\sum_{m=1}^{M} s_{nm}^k \leq J_{\max}, \quad \forall k \tag{1}$$

$$\sum_{k=1}^{K} s_{nm}^k \leq 1, \quad \forall n, m \tag{2}$$

where $J_{\max}$ represents the maximum number of offloading tasks that can be assign to each sub-channel.

The transmission power and channel power gain from ED n to MSN m on channel k are respectively denoted by p_{nm}^k and g_{nm}^k, Generally, we assume the channel gains of ED n on channel k are ordered as

$$\left|g_{n1}^k\right| \leq \left|g_{n2}^k\right| \leq \ldots \leq \left|g_{nm}^k\right| \leq \ldots \leq \left|g_{nM}^k\right|, \quad \forall n \in \mathcal{N},\ m \in \mathcal{M} \tag{3}$$

After the SIC technology, the received signal at MSN m from ED n on subchannel k is

$$y_{nm}^k = g_{nm}^k\sqrt{p_{nm}^k}x_{nm}^k + g_{nm}^k\sum_{l=m+1}^{M}\sqrt{p_{nl}^k}x_{nl}^k + \sum_{m=1}^{M}\sum_{l'=1,l'\neq n}^{N} g_{l'm}^k\sqrt{p_{l'm}^k}x_{l'm}^k + \sigma^2 \tag{4}$$

where x_{nm}^k denotes the modulated symbol of MSN m on subchannel k. The first term in (4) is the received signal transmitted from ED n to MSN m. The second term represents the co-interference when ED n offloads tasks to other MSNs on the same sub-channel. The third term represents the interference that other EDs offload tasks to MSN m through the same channel. The fourth term is white Gaussian noise (AWGN).

The data rate from ED n to MSN m on sub-channel k is

$$r_{nm}^k = W\log_2\left(1 + \gamma_{nm}^k\right) \tag{5}$$

In Eq. (5)

$$\gamma_{nm}^k = \frac{p_{nm}^k\left|g_{nm}^k\right|^2}{\left|g_{nm}^k\right|^2\sum\limits_{l=m+1}^{M}p_{nl}^k + \sum\limits_{m=1}^{M}\sum\limits_{l'=1,l'\neq n}^{N}p_{l'm}^k\left|g_{l'm}^k\right|^2 + \sigma^2}$$

where W is the bandwidth of the sub-channel. σ^2 denotes the power of the additive white Gaussian noise (AWGN). Let $I_{nm}^k = \left|g_{nm}^k\right|^2\sum\limits_{l=m+1}^{M}p_{nl}^k + \sum\limits_{m=1}^{M}\sum\limits_{l'=1,l'\neq n}^{N}p_{l'm}^k\left|g_{l'm}^k\right|^2.$

Therefore, the total sum rate from ED n to MSN m is

$$r_{nm} = \sum\nolimits_{k\in\mathcal{K}} s_{nm}^k r_{nm}^k \tag{6}$$

C. COMPUTION MODEL

In the paper, EDs' tasks are divided according to the proportion, θ_{nn} and θ_{nm} are represented the proportion of local computing and offloading tasks to MSN m of ED n respectively.

When requested task $\theta_{nn}D_n$ of ED n is computed locally, the required computing time T_{nn}^l can be expressed as

$$T_{nn}^l = \frac{\theta_{nn}D_nB_n}{f_n^l} \tag{7}$$

The system generates additional delay when the tasks are executed at the MSNs. For each ED, the latency consists of uplink transmission time, the processing time at the MSNs, and the downloading time of computation results. In this article, it is assumed that the results downloading time are ignored, since the task results are usually much smaller than the size of input data, as in [27], [36].

When computation task $\theta_{nm}D_n$ of ED n is offloaded to MSN m executed remotely, the transmission time can be computed as

$$T_{nm}^{off} = \frac{\theta_{nm}D_n}{r_{nm}} \tag{8}$$

The delay of remote calculation on MSN m is

$$T_{nm}^{exe} = \frac{\theta_{nm}D_nB_n}{f_{mn}^e} \tag{9}$$

where f_{mn}^e is the computing resources allocated by the MSN m to ED n.

The total delay caused by executing task $\theta_{nm}D_n$ of ED n at the MEC server m can be represented as

$$T_{nm}^c = T_{nm}^{off} + T_{nm}^{exe} \tag{10}$$

For the convenience of processing, the total delay of task execution for ED n is expressed as

$$T_{nm} = \begin{cases} \dfrac{\theta_{nn}D_nB_n}{f_n^l}, & n = m \\ \dfrac{\theta_{nm}D_n}{r_{nm}} + \dfrac{\theta_{nm}D_nB_n}{f_{mn}^e}, & n \neq m \end{cases} \tag{11}$$

Since the tasks of ED n are sliced into multiple parts for local computation and remote computation respectively, tasks are transmitted and processed in parallel. Therefore, the time of ED n completing the computation task is the maximum of local and edge computation time, which is presented as follows

$$T_n = \max T_{nm} \tag{12}$$

In addition, the task processing capacity of ED n is defined as [34]

$$C_n = \frac{D_n}{\mathrm{T}_n} \tag{13}$$

Finally, this article defines C as the overall task processing capability of the whole system.

$$\mathrm{C} = \sum_{n=1}^{N} C_n \tag{14}$$

IV. PROBLEM FORMULATION

When the task processing capability of the whole system is higher, namely the C is so higher that the system obtains better the system processing capability. Therefore, our goal is to maximize the task processing capability of the system by jointly optimizing the ratio of tasks $\boldsymbol{\theta}$, sub-channel assignment $\boldsymbol{s}$, transmit power $\boldsymbol{p}$, and computing resources allocation $\boldsymbol{f}$, our optimization problem P1 is described as

$$\max_{\{P,\theta,f,S\}} \ \mathrm{C} \tag{15}$$

$$\text{s.t. c1}: s_{nm}^k \in \{0,1\}, \quad \forall n,m,k \tag{15a}$$

$$\text{c2}: \sum_{n=1}^{N}\sum_{m}^{M} s_{nm}^k \leq J_{\max}, \quad \forall k \tag{15b}$$

$$\text{c3}: \sum_{k=1}^{K} s_{nm}^k \leq 1, \quad \forall n,m \tag{15c}$$

$$\text{c4}: \sum_{m=1}^{M}\sum_{k\in\mathcal{K}} s_{nm}^k p_{nm}^k \leq P_n^{\max}, \quad \forall n \tag{15d}$$

$$\text{c5}: p_{nm}^k \geq 0, \quad \forall n,m,k \tag{15e}$$

$$\text{c6}: r_{nm}^k \geq r_{\min}, \quad \forall n,m,k \tag{15f}$$

$$\text{c7}: f_{mn}^e \geq 0, \quad \forall n,m \tag{15g}$$

$$\text{c8}: \sum_{n=1}^{N} f_{mn}^e \leq F_m, \quad \forall m \tag{15h}$$

$$\text{c9}: \theta_{nn},\theta_{nm} \in [0,1], \quad \forall n,m \tag{15i}$$

$$\text{c10}: \theta_{nn} + \sum_{m=1}^{M}\theta_{nm} = 1, \quad \forall n \tag{15j}$$

where $P_n^{\max}$ is the maximum transmit power of ED n, F_m is the maximum computing capacity of MSN m.

Constraint c1 states that subchannel allocation s_{nm}^k is a binary variable; c2 means that at most one channel serves a task of one ED; c3 implies that each offloading task occupies at most one sub-channel; c4 and c5 represent power constraints; c6 represents the QoS constraint, and the basic communication must guarantee the network data rate; c7 and c8 indicate the computing resources limitation for each ED; c9 and c10 ensure all tasks of each ED are executed both local and on MSNs remotely.

Obviously, since channel decision variable is 0-1, the proposed problem P1 is a MINLP problem and it is difficult to obtain the optimal solution. To make it more tractable, problem P1 can be written as

$$\max\{\mathrm{C}\} = \max\left\{\sum_{n=1}^{N} C_n\right\} = \max\left\{\sum_{n=1}^{N}\min\left\{\frac{D_n}{\mathrm{T}_{nm}}\right\}\right\} \tag{16}$$

We assume that the input data D_n of each mobile device is fixed, the task processing capability of each ED is only related to its delay, and the delay among each ED does not affect each other. So the optimal solution of P1 can be solved when the minimum T_{nm} is obtained [34]. Therefore, the P1 problem can be transformed into

$$\begin{aligned} \text{P2:} \ \min_{\forall n} \ \max T_{nm} \\ \text{s.t.} \ \ \text{c1} - \text{c10} \end{aligned} \tag{17}$$

T_{nm} is mainly a function of task allocation, transmission power and computation resource. Therefore, task allocation and resource allocation are crucial in offloading. And we decomposed the P2 problem into two sub-problems for solution: resource allocation problem (P2$'$) and task offloading problem (P2$''$). Where P2$'$ determines how much transmitting power and computing capacity allocated on MSNs for offloading tasks, and P2$''$ determines how many tasks will be assigned to local EDs and MSNs to compute. The specific description is as follows.

Due to consider the transmission rate and computing capacity, the maximum effective offloading rate R is obtained from resource allocation is

$$\begin{aligned}\mathrm{P2}':\ &\max_{S,P,f} \mathrm{R} = \max \sum_{n=1}^{N}\sum_{m=1}^{M} R_{nm}\\ &\text{s.t. } \mathrm{c1}-\mathrm{c8}\end{aligned} \tag{18}$$

where $R_{nm} = \begin{cases} \dfrac{f_n^l}{B_n}, & n = m \\ \dfrac{r_{nm} f_{mn}^e}{B_n r_{nm} + f_{mn}^e}, & n \neq m \end{cases}$

Through task assignment, the minimum effective delay of task processing is

P2”:

$$\begin{aligned}&\min_{\theta} \max T_{nm} = \min\ \max\left\{\frac{\theta_{nn} B_n D_n}{f_n^l}, \frac{\theta_{nm} D_n}{R_{nm}}\right\}, \quad \forall n, m\\ &\text{s.t. } \mathrm{c9}, \mathrm{c10}\end{aligned} \tag{19}$$

V. PROPOSED ALGORITHM

A. PROBLEM DECOMPOSITION

For the problem (P2’), the effective offloading rate R of the system in (18) is equivalent to

$$\mathrm{R} = \sum_{n=1}^{N}\sum_{m=1}^{M} \frac{\sum_{k=1}^{K} s_{nm}^k r_{nm}^k f_{mn}^e}{B_n \sum_{k=1}^{K} s_{nm}^k r_{nm}^k + f_{mn}^e} \tag{20}$$

By using the idea of divide-and-conquer strategy, the offloading rate of ED n to MSN m for computing tasks was firstly optimized

$$\begin{aligned}&\max_{S,P,f} R_{nm} = \frac{\sum_{k=1}^{K} s_{nm}^k r_{nm}^k f_{mn}^e}{B_n \sum_{k=1}^{K} s_{nm}^k r_{nm}^k + f_{mn}^e}, \quad \forall m, n\\ &\text{s.t. } \mathrm{c1}-\mathrm{c8}\end{aligned} \tag{21}$$

Therefore problem (20) is equivalent to

$$\begin{aligned}\min_{S,P,f} \frac{1}{R_{nm}} &= \frac{B_n \sum_{k=1}^{K} s_{nm}^k r_{nm}^k + f_{mn}^e}{\sum_{k=1}^{K} s_{nm}^k r_{nm}^k f_{mn}^e}, \quad \forall m, n\\ &= \frac{B_n}{f_{mn}^e} + \frac{1}{\sum_{k=1}^{K} s_{nm}^k r_{nm}^k}, \quad \forall m, n\\ \text{s.t. } &\mathrm{c1}-\mathrm{c8}\end{aligned} \tag{22}$$

It can be seen from (22) that computational resource allocation f_{mn}^e and communication allocation s_{nm}^k and p_{nm}^k are decoupled in the objective function and constraints. By using this nature, we can decompose problem (22) into two independent problems, namely communication resource allocation and computational resource allocation, and solve them respectively, as shown in the following sections.

1) COMPUTATION RESOURCES ALLOCATION

Through (22), we describe the computational resource allocation problem as follow

$$\begin{aligned}&\min_{f} \frac{B_n}{f_{mn}^e}, \quad \forall m, n\\ &\text{s.t. } \mathrm{c7}, \mathrm{c8}\end{aligned} \tag{23}$$

It noted that the constraints c7 and c8 are convex, and we denote the objective function of (23) as $\Lambda(f)$, by obtaining the second derivative of $\Lambda(f)$ with respect to f_{mn}^e, we have

$$\frac{\partial^2 \Lambda(f)}{\partial f_{mn}^{e2}} = \frac{2B_n}{\left(f_{mn}^e\right)^3} > 0, \quad \forall m, n \tag{24}$$

Therefore, (22) is a convex optimization problem and can be solved by KKT conditions. We first express the Lagrangian function of (23) as

$$L(\Lambda(f), v) = \frac{B_n}{f_{mn}^e} + \sum_{m=1}^{M} v_m \left(\sum_{n=1}^{N} f_{mn}^e - F_m\right) \tag{25}$$

where $v = [v_1, v_2, \ldots, v_{\mathrm{M}}]$ is the Lagrange multiplier vector. And take the first derivative of a Lagrange function

$$\frac{\partial L(\Lambda(f), v)}{\partial f_{mn}^e} = -\frac{B_n}{\left(f_{mn}^e\right)^2} + v_m, \quad \forall n, m \tag{26}$$

Let $\partial L(\Lambda(f), v)/\partial f_{mn}^e = 0$, the optimal resource allocation of problem (21) can be obtained as follows

$$\left(f_{mn}^e\right)^* = \sqrt{\frac{B_n}{v_m^*}}, \quad \forall n, m \tag{27}$$

When $v_m^* > 0$, meet the following constraints

$$\sum_{n=1}^{N} \left(f_{mn}^e\right)^* = F_m, \quad \forall m \tag{28}$$

Substitute (25) into (26), the Lagrange multiplier is

$$v_m^* = \left(\frac{1}{F_m}\sum_{n=1}^{N}\sqrt{B_n}\right)^2, \quad \forall m \tag{29}$$

Substituting (29) into (27), we can obtain the optimal solution of computation resource as

$$\left(f_{mn}^e\right)^* = \frac{F_m\sqrt{B_n}}{\sum_{n=1}^{N}\sqrt{B_n}}, \quad \forall m \tag{30}$$

2) COMMUNICATION RESOURCES ALLOCATION

Through (22), we describe the communication resource allocation of all subchannels problem as follows

$$\min_{S,P} \frac{1}{\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K} s_{nm}^{k} r_{nm}^{k}} \quad \text{s.t. c1} - \text{c6} \tag{31}$$

Similarly, we equate (31) as

$$\max_{S,P} \xi = \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K} s_{nm}^{k} r_{nm}^{k} \quad \text{s.t. c1} - \text{c6} \tag{32}$$

Since subchannel allocation is a 0-1 decision problem, problem (32) is still a MINLP problem. Moreover, it can be seen that the subchannel allocation and power allocation are coupled of on all subchannels, and it is quite complicated to obtain the global optimal solution, so we firstly decouple the subchannel allocation and power allocation to obtain the solution. It assumes that the transmission power is equal on each subchannel, and we propose a greedy subchannel matching algorithm for channel assignment.

3) SUB-CHANNEL ASSIGNMENT

In order to describe the dynamic matching between EDs and sub-channels, sub-channel allocation is regarded as a two-sided matching process between the set of offloading tasks $\mathcal{J}$ and sub-channel set SC. If ED n uses channel k to offload the task to MSN m in the process of offloading tasks, we deem that sub-channel SC_k and offloaded task D_{nm} are matched each other. According to channel state information, the preference lists of offloading tasks and subchannels can be expressed as

$$\begin{aligned} PF\_ED &= [PF\_ED(D_{11}), \cdots, PF\_ED(D_{nm}), \cdots, \\ &\quad PF\_ED(D_{NM})]^T \\ PF\_SC &= [PF\_SC(1), \cdots, PF\_SC(k), \cdots, PF\_SC(K)]^T \end{aligned}$$

where $PF\_ED(D_{nm})$ and $PF\_\mathrm{SC}(k)$ are the preference lists of D_{nm} and SC_k, respectively.

We use a notation $\succ$ to represent the preference relationship, if ED n offloads task D_{nm} to MSN m has higher channel gain on SC_i than that on SC_j, we say D_{nm} prefers SC_i to SC_j. It can be noted as

$$SC_i(D_{nm}) \succ SC_j(D_{nm}) \tag{33}$$

If the offloading tasks in set q can provide higher ξ than in set q' on SC_k, we say SC_k prefers offloading task set q to task set q', and we describe the case as

$$\xi(q) > \xi\left(q'\right), \quad q,\ q' \subset \mathcal{J} \tag{34}$$

According to the preference list of EDs' offloading tasks and sub-channels, the sub-channel allocation problem is expressed as a two-sided matching problem as [15] and [33]. First, two definitions are considered.

Definition 1: Given offloading task of EDs and subchannels as two disjoint sets $\mathcal{J}$ and SC. A many-to-one, two-sided matching $\mathcal{A}$ is a mapping from all the subsets of $\mathcal{J}$ into SC for $D_{nm} \in \mathcal{J}$ and $SC_k \in \mathcal{SC}$, and satisfies follow conditions

1) $\mathcal{A}(D_{nm}) \in \mathcal{SC}$.
2) $\mathcal{A}^{-1}(SC_k) \subseteq \mathcal{J}$.
3) $|\mathcal{A}(D_{nm})| = 1, \left|\mathcal{A}^{-1}(SC_k)\right| \leq J_{\max}$.
4) $SC_k \in \mathcal{A}(D_{nm}) \Leftrightarrow D_{nm} \in \mathcal{A}^{-1}(SC_k)$.

The above conditions are explained as follows: 1) shows that each offloading task if and only if matches one subchannel; 2) implies each subchannel can be matched with a subset of tasks; 3) represents that the number of tasks of EDs can be allocated on the same subchannel is limited to $J_{\max}$; and 4) expresses offloading task D_{nm} and subchannel SC_k are matched with each other.

Definition 2: Given a matching $\mathcal{A}$, we suppose $D_{nm} \notin \mathcal{A}^{-1}(\mathrm{SC}_k), \mathrm{SC}_k \notin \mathcal{A}(D_{nm})$, if there is $\xi(S_{new}) > \xi\left(\mathcal{A}^{-1}(\mathrm{SC}_k)\right)$, S_{new} becomes the preferred tasks set for subchannel SC_k and (D_{nm}, SC_k) is a preferred matched pair. Where $S_{new} \subseteq \{D_{nm}\} \cup S, S = \mathcal{A}^{-1}(\mathrm{SC}_k)$,and where S is the task set has been assigned to SC_k.

On basis of the above analysis, we will depict the matching process between the offloading tasks of EDs and the subchannels. When EDs offload tasks, if each ED has to select the best subchannel to transfer tasks. Meanwhile, each subchannel has to assign the best subset of tasks. This will lead to high complexity, especially while there are more EDs. Since the optimal solution case is to search all possible matches to maximize overall transmission rate. Therefore, in order to reduce the complexity, a suboptimal subchannel allocation algorithm (SSAA) is proposed. The main idea of the suboptimal matching algorithm is that each ED sends matching request through its offloading tasks' preference list to its preferred channel, but this preferred channel has the right to reject or accept the task based on the offload efficiency provided by all offloading tasks. The algorithm 1 describes as follows.

B. COMPLEXITY ANALYSIS

The optimal subchannel assignment scheme can be obtained by exhaustive searching over all possible combinations of EDs and subchannel and selecting one that maximizes the system offloading efficiency. If we have J offloading tasks of EDs and K subchannels, we suppose there are two offloading tasks that can reuse the same subchannel. The time complexity of exhaustive searching is $O\left((2K)!/2^K\right)$. To compare with the complexity of our proposed algorithms, we take natural logarithm of the complexity. The exhaustive searching logarithm complexity is $O(\ln((2K)!) - K) = O(\ln((2K)!))$. By adopting the Stirling's formula [15], $\ln(x!) = x\ln x - x + O(\ln(x))$, the logarithm complexity of the exhaustive method can be expressed as $O(K\ln K)$. In the proposed suboptimal algorithm, the complexity of the worst case is $O\left(K^2\right)$. And the logarithm complexity is $O(\ln K)$. Since $O(K\ln K) > O(\ln K)$ and the actual complexity of the proposed suboptimal algorithm is much smaller than the worst-case complexity, so the complexity of the proposed algorithm is much smaller than the optimal sub-channel

```
Algorithm 1 SSAA
1: Initialize the power allocation for each ED $p_{nm} = P_n^{\max}/M$.
2: Initialize preference lists $PF\_ED\,(D_{nm})$ for all the offloading tasks of EDs and $PF\_SC\,(k)$ for all the subchannels according to the channel state information, $\forall n, m$.
3: Initialize the matched list $S_{Match}\,(k)$ to record the set of tasks $D_{nm}$ allocated on $SC_k$ for all the subchannels, $\forall k \in \mathcal{SC}$.
4: Initialize $S_{UnMatch}$ to record $D_{nm}$ that has not been allocated any subchannel.
5: while $\{S_{UnMatch}\} \neq \emptyset$ do
6:   for $n = 1$ to $N$ do
7:     for $m = 1$ to $M$ do
8:       Based preference lists $PF\_ED(D_{nm})$, each ED sends matching request to its most preferred subchannel $k^\wedge$.
9:       if $|S_{Match}(k^\wedge)| < J_{\max}$ then
10:        Subchannel $k^\wedge$ adds task $D_{nm}$ of ED n to $S_{Match}(k^\wedge)$, and removes $D_{nm}$ from $S_{UnMatch}$
11:      end if
12:      if $|S_{Match}(k^\wedge)| = J_{\max}$ then
13:        a) Subchannel $k^\wedge$ select the set of $D_{nm}$, which satisfies maximum $\xi$.
           b) Reject other tasks of EDs, update $S_{UnMatch}$ and delete the rejected tasks from subchannel k's preference list.
           c) The unchosen $D_{nm}$ will go to step 8 and repeat this step until it has been allocated on one subchannel.
14:      end if
15:    end for
16:  end for
17: end while
```

allocation scheme. It can be found that for a small number of EDs ($N = 3$), the SSAA will yield the identical results from the exhaustive search.

1) POWER ALLOCATION ON EACH SUBCHANNEL

In this section, we will optimize the transmit power P under given sub-channel allocation. The power allocation problem on each sub-channel is expressed as

$$\begin{aligned} &\max_{P} \sum_{n=1}^{N}\sum_{m=1}^{M} r_{nm}^{k}, \quad \forall k \\ &\text{s.t. } \mathrm{c4} - \mathrm{c6} \end{aligned} \tag{35}$$

Obviously, the objective function of (35) is the logarithmic function of P_{nm}^k, the second derivative of P_{nm}^k is less than 0, so problem (35) is convex. Therefore, it can be solved by the KKT condition. First construct the Lagrangian function

$$\begin{aligned} L\,(P, \alpha, \beta) = {} & \sum_{n=1}^{N}\sum_{m=1}^{M} r_{nm}^{k} + \sum_{n=1}^{N}\sum_{k=1}^{K} \alpha_{nm}^{k} \left(P_n^{\max} - \sum_{m=1}^{M} p_{nm}^{k}\right) \\ & + \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} \beta_{nm}^{k} \left(r_{nm}^{k} - r_{\min}\right) \end{aligned} \tag{36}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} > 0$ represents the Lagrange multiplier vector.

```
Algorithm 2 PAA
1: Set the iteration index $t = 0$, the iteration step sizes $\zeta_1\,(t) > 0$, $\zeta_2\,(t) > 0$
2: Initialize $\varepsilon > 0$, $\alpha_{nm}^k\,(1) > 0$ and $\beta_{nm}^k\,(1) > 0$, according to eq. (38), we calculate $p_{nm}^k$.
3: Update the Lagrange multiplier $\alpha_{nm}^k\,(t)$ and $\beta_{nm}^k\,(t)$ according to (39) and (40).
4: Then update $p_{nm}^k$ according to equation (38)
5: if $\left|p_{nm}^k\,(t+1) - p_{nm}^k\,(t)\right| < e||t > T_{\max}$, the result is the optimal solution, where $T_{\max}$ is the maximum number of iterations.
6: else set $t = t + 1$and return step 3.
7: end if
```

The first derivative of p_{nm}^k is

$$\frac{\partial L\,(P, \alpha, \beta)}{\partial p_{nm}^k} = \frac{W g_{nm}^k}{\left(I_{nm}^k + \sigma^2 + p_{nm}^k g_{nm}^k\right) \ln 2} \left(1 + \beta_{nm}^k\right) - \alpha_{nm}^k \tag{37}$$

Let $\partial L\,(P, \alpha, \beta)\,/\partial p_{nm}^k = 0$, the optimal power allocation is expressed as

$$\left(p_{nm}^k\right)^* = \frac{W\left(1 + \left(\beta_{nm}^k\right)^*\right)}{\left(\alpha_{nm}^k\right)^* \ln 2} - \frac{I_{nm}^k + \sigma^2}{g_{nm}^k} \tag{38}$$

The power allocation solution is obtained in (38), we can apply the sub-gradient method to update the Lagrange multiplier for the objective function is differentiable. Therefore, the Lagrange multiplier can be updated with gradient descent as

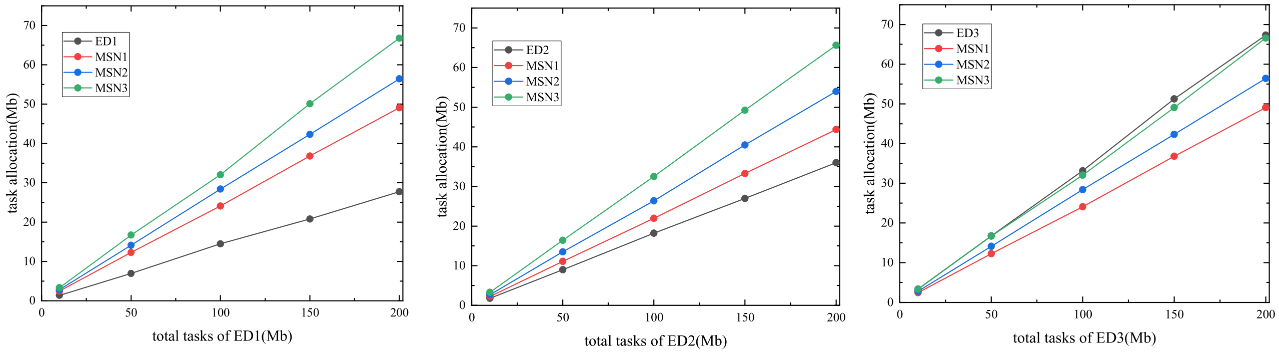
$$\alpha_{nm}^k\,(t+1) = \left[\alpha_{nm}^k\,(t) - \zeta_1\,(t)\left(P_n^{\max} - \sum_{m=1}^{M} p_{nm}^k\right)\right]^+, \quad \forall k \tag{39}$$

$$\beta_{nm}^k\,(t+1) = \left[\beta_{nm}^k\,(t) - \zeta_2\,(t)\left(r_{nm}^k - r_{\min}\right)\right]^+, \quad \forall k \tag{40}$$

where t is the iteration index, $\zeta_1\,(t)$, $\zeta_2\,(t)$ are positive step sizes at iteration t. Power allocation algorithm (PAA) is elaborated in **Algorithm 2**.

C. JOINT SUB-CHANNEL ASSIGNMENT AND TRANSMIT POWER ALLOCATION

In the previous sections, equal power allocation is assumed, the solution for the subchannel allocation was given, then we optimize transmission power under the given conditions of channel allocation. To simultaneously optimize power allocation and subchannel allocation, we propose a joint communication resources allocation algorithm (JCRAA) to solve problem (31). The key idea of JCRAA is to iteratively update sub-channel assignment through Algorithm 1 and transmit power allocation through Algorithm 2. When the subchannel assignment solution can't be changed, the iterations stop. The details of JCRAA is shown in **Algorithm 3**.



(a) (b) (c)

FIGURE 2. The relationship between the total tasks and task allocation for each ED.

```
Algorithm 3 JCRAA
1: Let the iteration index $t_1 = 0$.
2: Initialize power allocation for each ED within the power range $\mathrm{P}^{(0)}$.
3: Let $t_1 = t_1 + 1$.
4: Under power allocation $\mathrm{P}^{(t_1-1)}$, update the sub-channel allocation result according to Algorithm 1.
5: Assign the sub-channel $\mathcal{SC}^{(t_1-1)}$ and update the power allocation results according to Algorithm 2.
6: if $\mathcal{SC}^{(t_1)} = \mathcal{SC}^{(t_1-1)}$ then
7:     the algorithm is terminated
8: else return step 3
9: end if
```

```
Algorithm 4 TAA
Input: Task size of each ED to be calculated, and resource allocation results in P2'.
Output: Local task allocation ratio $\theta_{nn}$ and offloading task ratio $\theta_{nm}$.
For $\forall n$, compute the total offloading rate $R_n$;
For $\forall n$, compute offloading rate $R_{nm}$ from ED $n$ to MSN $m$;
Calculate the offloading ratio $\theta_{nm} = R_{nm}/R_n$;
Calculate the proportion of tasks performed locally $\theta_{nn} = 1 - \sum_{m=1}^{M} \theta_{nm}, \forall n$
```

D. TASK ALLOCATION STRTEGY BASED ON RESOURCE ALLOCATION

Based on the resource allocation, we design a task allocation algorithm (TAA) to determine how many tasks should be allocated to local and edge nodes for problem P2. The main steps of the **Algorithm 4** are as follows.

VI. SIMULATION RESULTS

In this section, the simulation results are used to evaluate the impact of the proposed resource allocation scheme on system performance. We consider the scenario of 3 EDs, 3 MSNs and 5 subchannels. In simulations, the local computing capacity for the 3 EDs is $f_n^l = \{0, 6, 0.8, 1\}$ GHz. Meanwhile for MEC server, the computing capacity of the 3 MSNs is $F_m = \{5, 10, 15\}$ GHz as [9]. For the wireless transmission, the channel gains are characterized by a path-loss model and we set the pass loss model as modeled as $36.7\log(d_{nm}) + 140.7$ [35], d_{nm} is the distance between the ED n and MSN m, Rayleigh fading obeys zero mean and unit variance similar to [37]. Sub-channel bandwidth $W =$ 1MHz, maximum transmission power of the 3 EDs is $P_n^{\max} =$ 27dBm, we set the noise power is $\sigma^2 = -174$dBm/Hz.The minimum transmission rate of each ED is normalized $r_{\min} =$ 1bps/Hz. We compare our proposed scheme in this article (NOMA-MEC) against the following benchmark schemes:

1) All local computing scheme (ALL LOCAL): A All tasks of EDs are executed locally.

2) All MEC computing scheme (ALL MEC): All tasks are offloaded to 3 MSNs by NOMA.

3) One MSN scheme (One MSN): We consider a MSN scheme, that is an edge server, as in [20]. In simulation, Let MSN's computing capability be the sum of the computing capability of the multi-node collaborative edge nodes in this article, which is 30GHz.

4) OMA-MEC scheme (OMA-MEC): We consider an OMA-MEC system as a benchmark, where EDs adopt frequency division multiple access (FDMA) scheme for computation offloading.

A. EVALUATION OF TASK ALLOCATION RESULTS

Firstly, to intuitively and concretely reflect the task distribution of the proposed scheme in this article, the relationship between the total tasks volume of each ED and task assignment is shown respectively in Fig. 2 (a), (b) and (c). It can be seen from the figures that for each ED, the number of tasks assigned to local EDs and MSNs increases with the total number of processed tasks growing. At the same time, it can be seen that for ED3, the amount of tasks assigned to local is basically is equal to the tasks assigned to MSN3, which is due to the larger computing capacity of ED3, when assigned a large amount of tasks, it will lead to a small delay.

In addition, in order to analyze the overall distribution of system computation tasks, Fig. 3 describes the relationship between the total tasks of the system and the tasks allocation of MSNs and local EDs. It can be seen that as the total number

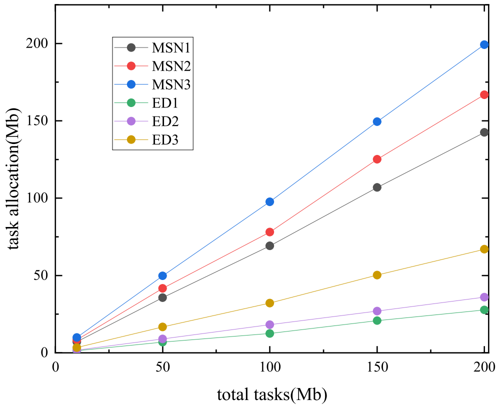


FIGURE 3. The relationship between the total number of tasks in the system and task allocation.

of system tasks increases, the amount of tasks allocated to the MSNs and users themselves increases. At the same time, it can be seen that when the amount of system tasks is small (less than 100Mb), the tasks allocated to the three edge servers is basically the same. However, when the system tasks become large (200Mb), the tasks allocated to MSN3 is much larger than the tasks allocated to MSN1 and MSN2. This is because MSN3 has strong computing capacity and can meet the needs of EDs with large tasks. As the amount of tasks increases, the amount of tasks allocated to ED3 is larger.

B. EVALUATION OF SYSTEM PERFORMANCE

The relationship between total tasks and system performance under the five schemes is shown as Fig. 4. As shown in Figure (e) and (f), the relationship between system delay and energy consumption with the total tasks is pictured. Since it is a task assignment based on resource allocation, the time delay and energy consumption are linear with the total amount of tasks. We can see from the figure that the solution in this article has the most advantages in terms of delay and energy consumption. The case of OMA-MEC is slightly inferior to the NOMA-MEC case proposed in this article. The third best case is all offloading scheme, compared with the case of one MSN, although local computing is not considered, EDs' tasks are offloaded to different MSNs to perform collaborative computing. And the computing capacity of the edge server is the sum of the computing capacity of the 3 MSNs proposed in this article, but it does not carry out collaborative computing, resulting in high delay and energy consumption. And the system performance of all local was the worst compared to the other schemes.

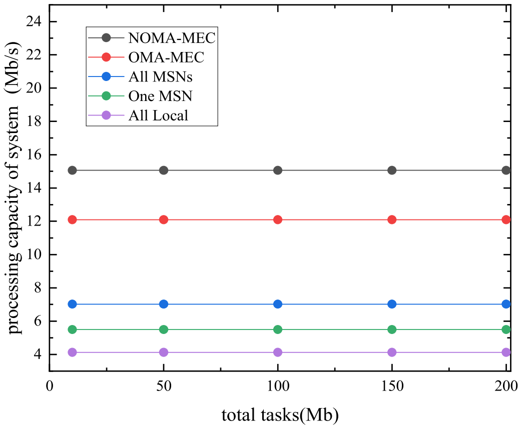
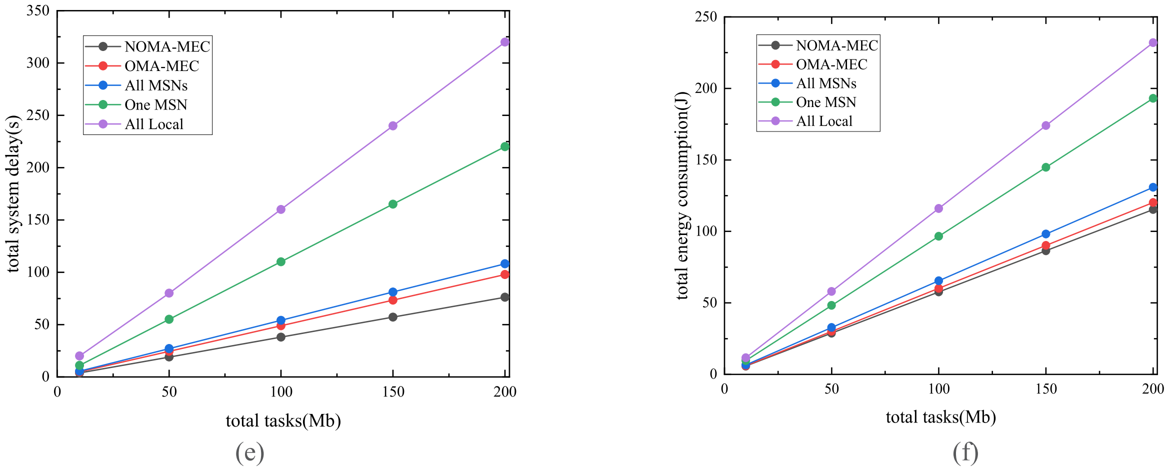


FIGURE 5. The relationship between system tasks and system processing capability.

Fig.5 performs the relationship between the total amount of tasks and the system processing capability. Since the processing capability characterizes the system capacity, for a system, as the task size increases, the system processing capability remains unchanged. However, it can be seen from the figure that when the solution in this article is adopted, the processing capability of the system is the best, followed by the solution for OMA-MEC, and the worst for all local computing. This is because for the five schemes, under the condition of equal tasks, the scheme proposed in this article has the smallest delay so that leads to the processing capability.



(e)

(f)

FIGURE 4. The relationship between total tasks and system performance.

VII. CONCLUSION

To enhance the performance of MEC systems and further improve user experience, we proposed a novel network scenario which we used NOMA to transmit the offloading tasks of EDs to multiple edge servers. In the paper, we presented an optimization framework for a multi-user multi-task and multi-server NOMA-MEC system to maximize system processing capability via jointly optimizing the tasks offloading and resources allocation. By decomposing the formulated problem, an efficient algorithm was proposed to tackle the formed problem under the help of convex optimization theory. Through computer simulation, the effectiveness of the proposed scheme was verified. It showed that our proposed scheme can efficiently reduce the delay and energy consumption and improve the processing capability of MEC systems.

REFERENCES

[1] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.

[2] 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.

[3] 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.

[4] R. Roman, J. Lopez, and M. Mambo, "Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges," *Future Gener. Comput. Syst.*, vol. 78, pp. 680–698, Jan. 2018.

[5] M. Caprolu, R. Di Pietro, F. Lombardi, and S. Raponi, "Edge computing perspectives: Architectures, technologies, and open security issues," in *Proc. IEEE Int. Conf. Edge Comput. (EDGE)*, Milan, Italy, Jul. 2019, pp. 116–123.

[6] X. Lyu, H. Tian, L. Jiang, A. Vinel, S. Maharjan, S. Gjessing, and Y. Zhang, "Selective offloading in mobile edge computing for the green Internet of Things," *IEEE Netw.*, vol. 32, no. 1, pp. 54–60, Jan. 2018.

[7] K. Zhang, Y. Mao, S. Leng, Y. He, and Y. Zhang, "Mobile-edge computing for vehicular networks: A promising network paradigm with predictive offloading," *IEEE Veh. Technol. Mag.*, vol. 12, no. 2, pp. 36–44, Jun. 2017.

[8] T. Taleb, K. Samdanis, B. Mada, H. Flinck, 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.

[9] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. K. Tsang, "NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12244–12258, Dec. 2018.

[10] L. P. Qian, B. Shi, Y. Wu, B. Sun, and D. H. K. Tsang, "NOMA-enabled mobile edge computing for Internet of Things via joint communication and computation resource allocations," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 718–733, Jan. 2020.

[11] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[12] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[13] A. J. Muhammed, Z. Ma, L. Li, P. D. Diamantoulakis, and G. K. Karagiannidis, "Energy efficient power and subcarrier allocation for downlink non-orthogonal multiple access systems," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Honolulu, HI, USA, Sep. 2019, pp. 1–5.

[14] J. Wang, X. Kang, S. Sun, and Y.-C. Liang, "Throughput maximization for peer-assisted wireless powered IoT NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5278–5291, Aug. 2020.

[15] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[16] P. Zhao, H. Tian, C. Qin, and G. Nie, "Energy-saving offloading by jointly allocating radio and computational resources for mobile edge computing," *IEEE Access*, vol. 5, pp. 11255–11268, 2017.

[17] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Mar. 2018.

[18] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4188–4200, Jun. 2019.

[19] M. Li, S. Yang, Z. Zhang, J. Ren, and G. Yu, "Joint subcarrier and power allocation for OFDMA based mobile edge computing system," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–6.

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

[21] Z. Kuang, L. Li, J. Gao, L. Zhao, and A. Liu, "Partial offloading scheduling and power allocation for mobile edge computing systems," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6774–6785, Aug. 2019.

[22] W. Fan, Y. Liu, B. Tang, F. Wu, and Z. Wang, "Computation offloading based on cooperations of mobile edge computing-enabled base stations," *IEEE Access*, vol. 6, pp. 22622–22633, 2018.

[23] L. Yang, H. Zhang, M. Li, J. Guo, and H. Ji, "Mobile edge computing empowered energy efficient task offloading in 5G," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6398–6409, Jul. 2018.

[24] K. Cheng, Y. Teng, W. Sun, A. Liu, and X. Wang, "Energy-efficient joint offloading and wireless resource allocation strategy in multi-MEC server systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[25] B. Yang, X. Cao, J. Bassey, X. Li, and L. Qian, "Computation offloading in multi-access edge computing: A multi-task learning approach," *IEEE Trans. Mobile Comput.*, early access, Apr. 27, 2020, doi: 10.1109/TMC.2020.2990630.

[26] M. Zeng, R. Du, V. Fodor, and C. Fischione, "Computation rate maximization for wireless powered mobile edge computing with NOMA," in *Proc. IEEE 20th Int. Symp.*, Washington, DC, USA, Jun. 2019, pp. 1–9.

[27] W. Zhou, L. Lin, J. Liu, D. Zhang, and Y. Xie, "Joint offloading decision and resource allocation for multiuser NOMA-MEC systems," *IEEE Access*, vol. 7, pp. 181100–181116, 2019.

[28] K. Wang, Z. Hu, Q. Ai, Y. Zhong, J. Yu, P. Zhou, L. Chen, and H. Shin, "Joint offloading and charge cost minimization in mobile edge computing," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 205–216, Feb. 2020.

[29] B. Yang, X. Cao, X. Li, C. Yuen, and L. Qian, "Lessons learned from accident of autonomous vehicle testing: An edge learning-aided offloading framework," *IEEE Wireless Commun. Lett.*, vol. 9, no. 8, pp. 1182–1186, Aug. 2020.

[30] B. Yang, X. Cao, C. Yuen, and L. Qian, "Offloading optimization in edge computing for deep learning enabled target tracking by Internet-of-UAVs," *IEEE Internet Things J.*, early access, Aug. 14, 2020, doi: 10.1109/JIOT.2020.3016694.

[31] Z. Zhao, R. Zhao, J. Xia, X. Lei, D. Li, C. Yue, and L. Fan, "A novel framework of three-hierarchical offloading optimization for MEC in industrial IoT networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5424–5434, Aug. 2020.

[32] Z. Song, Y. Liu, and X. Sun, "Joint radio and computational resource allocation for NOMA-based mobile edge computing in heterogeneous networks," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2559–2562, Dec. 2018.

[33] F. Fang, J. Cheng, and Z. Ding, "Joint energy efficient subchannel and power optimization for a downlink NOMA heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1351–1364, Feb. 2019.

[34] C. Bin, L. Y. Cheng, and L. Lei, "Distributed game offloading strategy in ad hoc cloud environment," (in Chinese), *J. Commun.*, vol. 38, no. 11, pp. 24–34, 2017.

[35] T. X. Tran and D. Pompili, "Joint task offloading and resource allocation for multi-server mobile-edge computing networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 856–868, Jan. 2019.

[36] X. Lyu, H. Tian, C. Sengul, and P. Zhang, "Multiuser joint task offloading and resource optimization in proximate clouds," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3435–3447, Apr. 2017.

[37] Z. Jian, W. Muqing, and Z. Min, "Joint computation offloading and resource allocation in C-RAN with MEC based on spectrum efficiency," *IEEE Access*, vol. 7, pp. 79056–79068, 2019.

JIANBIN XUE was born in Huining, Gansu, China, in 1973. He received the B.S. degree in communication engineering from Sichuan University, in 1997, and the M.S. degree in signal and information system and the Ph.D. degree in control theory and control engineering from the Lanzhou University of Technology, in 2005 and 2009, respectively.

He is currently the Vice President of the Graduate School, Lanzhou University of Technology, the Leader of the Internet of Things engineering with the Lanzhou University of Technology, the Director of the degree point of information and communication engineering, and the Vice President of the Internet of Things Talent Training Maker Alliance. He also served as the Judge for key research and development projects of the Ministry of Science and Technology, and a Letter Reviewer of National Natural Science Foundation projects and National Excellent Master and Doctoral theses. He has completed more than 20 projects, including the National Natural Science Foundation, the National Key Laboratory Project, and the Basic Business Expenses of Colleges and Universities, the research results have published more than 60 articles. His main research interests include the wireless communication theory and technology, the information system modeling and simulation, the communication network and communication systems and the multi-antenna system and technology, mobile edge computing, non-orthogonal multiple access, and D2D technology.

Dr. Xue is a Senior Member of the China Electronics Society. He has won the Science and Technology Award of the Gansu Science and Technology Information Society, the Science and Technology Progress Award of Lanzhou City, and the Supervisor of the Higher Education Cup National College Students Mathematical Modeling Competition National Award. He has won awards and titles of the Lanzhou University of Technology, such as the Hongliu Youth Talent Program, the Three Education, and the Advanced Individual in Science and Technology Innovation.

YANING AN received the B.S. degree in communication engineering from Jishou University, China, in 2018. She is currently pursuing the M.S. degree in electronics and communication engineering with the Lanzhou University of Technology. Her current research interests include resource allocation, mobile edge computing, and non-orthogonal multiple access.