IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2019, accepted June 6, 2019, date of publication June 13, 2019, date of current version June 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922702

Joint Computation Offloading and Resource Allocation in C-RAN With MEC Based on Spectrum Efficiency

ZHANG JIAN, WU MUQING, AND ZHAO MIN
School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
Corresponding author: Zhang Jian (zhang_jian@bupt.edu.cn)

This work was supported by the Beijing Laboratory of Advanced Information Networks.

ABSTRACT The cloud radio access network (C-RAN) with mobile edge computing (MEC) structure which consists of a baseband unit (BBU) pool integrating with an MEC server and several remote radio heads (RRHs) beside the mobile terminals can help users with computational resource-intensive tasks and bring extra profits to network operator at the same time. This paper presents a novel task-aware C-RAN with MEC structure and formulates a profit maximization problem by jointly optimizing offloading strategy, radio and computational resources allocation under the constraints of offloading latency, fronthaul capacity along with limited bandwidth and computational resource. To solve this NP-hard optimization problem in a distributed and efficient way, we propose a spectrum efficiency (SE)-based joint optimization for offloading and resource allocation (SJOORA) scheme which decomposes the original problem into two sub-problems. A SE-based offloading strategy is proposed with confirmed resource allocation, and on the other hand, bandwidth and computational resource allocation problem is solved by using a Lagrangian multiplier method with predetermined offloading strategy. Finally, by solving these two sub-problems iteratively, a suboptimal solution is obtained for the original problem. The simulation results show that the proposed SJOORA scheme can effectively increase the profit of network operator with relative lower complexity.

INDEX TERMS Offloading strategy, resource allocation, cloud radio access network, mobile edge computing, spectrum efficiency.

I. INTRODUCTION

With the exponentially increasing of the mobile devices such as smart phones, tablets and hand-held terminals, the corresponding mobile traffic is also growing rapidly which is predicted doubling every year [1]. The mobile service providers, wireless network operators, and even mobile users are facing not only opportunities but also tough challenges. Obviously, traditional wireless cellular networks are becoming incapable to meet the exponentially growing demand in high data rate. In addition, the hugely increasing of computational resource intensive tasks, such as multimedia applications, high definition video playing and gaming that appear in our daily life, causes a heavy load to the mobile terminals (MTs) with limited computing capability and radio resource [2].

To tackle the data rate issue, both novel network architectures (e.g., heterogeneous network (HetNet) and ultra-dense network (UDN)) and techniques (e.g., mm-Wave, massive-MIMO etc.) are applied to promote the network throughput [3]. In the process of increasing data rate, overall spectrum efficiency (SE) and energy efficiency (EE) are also considered with limited radio resource constraints [4]–[8]. Reference [4] jointly optimizes the subchannels and resource allocation problem aiming to maximize the EE of the HetNet with multi-homed users. References [5] and [6] concentrate on the resource allocation scheme to maximize both EE and SE. Reference [7] considers the effect of interference in formulating a global EE maximization resource allocation metric. The authors in [8] investigate the subchannels and power allocation in a multiple downlink NOMA system with considering queue stability, and the formulated problem is solved by using a Lyapunov optimization.

Especially, a new promising network infrastructure, i.e., a cloud radio access network (C-RAN) architecture is proposed which soon draws a lot of attention in both academia and industry these years [9]. The C-RAN divides a traditional

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

base station (BS) into three parts: a baseband unit (BBU) pool, several remote radio heads (RRHs) and the fronthaul links between them. Concretely, most of the baseband signal processing techniques are implemented by the virtual machines (VMs) in BBU, thus RRHs can be distributed close to users easily due to the light design with only limited radio functions such as A/D, D/A conversion, frequency conversion and signal amplification [10]. Thus, the C-RAN can facilitate costs saving, better SE and interference management due to the feature of the centralization of BBU pool and the distribution of RRHs [11]. Reference [12] considers a multitenant heterogeneous C-RAN structure and jointly optimizes user association, bandwidth allocation, power allocation and virtual BBU capacity allocation to maximize the weighted network throughput where tenants' priorities, baseband resource, fronthaul and backhaul capacities, quality of service (QoS), as well as interference are taken into account. Reference [13] jointly optimizes the RRH selection, user-RRH association and beamforming vectors to minimize the total network power consumption while the channel state information (CSI) is incomplete.

Then, to deal with the computational capability issue, a mobile cloud computing (MCC) system is proposed in which a centralized cloud can help dealing with the computational intensive tasks for the MTs [14], [15]. However, the MCC system also causes huge additional transmit loads on radio and backhaul of mobile networks and gives rise to high latency as well, since the cloud servers which help computing the offloading tasks are far away from the MTs [16]. To address the problem of the long latency, the cloud servers should be placed close to the MTs, that motivates the development of the mobile edge computing (MEC) system which combines the mobile cloud computing and wireless network service so that the MTs can get nearby distributed computational resource with much lower latency [17]. Compared with an MCC system, the MEC system has several advantages such as lower latency, saving energy for mobile devices, supporting context-aware computing, and enhancing privacy and security for mobile users [18]. The computation offloading strategy is one of the key challenges and research focuses for mobile computing which always couples with computational and radio resources allocation problem.

The main objective of the extensive works that focus on the offloading decision is to minimize offloading latency, energy consumption, or to maximize the utility while users' QoS and resource constraints are satisfied. In [19], the authors consider the case that the computational resource allocated to each user is stationary and then jointly optimize the offloading strategy and radio resource allocation to minimize the total network energy consumption. In [20], the allocation of computational resource, spectrum resource and cache resource are jointly optimized with considering the offloading strategy, and the aim is to maximize the utility. Authors in [21] propose a MEC system with multiple users and tasks where not only users but also small BSs can offload tasks to a particular MEC server or to a small BS and/or the macro BS respectively, and a total energy consumption minimization problem is formulated with jointly optimizing computation offloading and users association. In [22], not only mobile cloud users but also ordinary communication users without offloading demand are considered and the authors jointly optimize radio and computational resources to minimize the overall energy consumption at users' side with the constraints of transmit power and offloading latency, however, only radio and computational resources are optimized while the MEC users are fixed and no offloading strategies are concerned. Reference [23] considers a single-user single-MEC system with multiple independent computation tasks offloading requests, in this scene, tasks' offloading order and corresponding transmit power are jointly optimized on purpose of minimizing the execution delay and devices energy consumption. Reference [24] considers the maximization of users' total utility in the heterogeneous MCC network. Specifically, bandwidth and computational resource allocation are jointly optimized and the combinatorial optimization problem is solved by the proposed evolutionary approaches.

Typically, MEC servers are data centers integrated within the BSs that distributed at the edge of mobile network, and they are normally accessible by nearby mobile users via one-hop wireless connection [25]. Considering the computational and storage resources in the BBU pool and the distribution of the RRHs that close to mobile users, a C-RAN may facilitate the implementation of the MEC system. Actually, several researches have been dedicated to the combination of the MEC system and the C-RAN structure [11], [26], [27]. Reference [11] proposes a novel mobile cloud radio access network (MC-RAN) structure, where in the BBU pool VMs are divided into two kinds: virtualized BBU (vBBU) as the communication computing providing unit (CCPU) and virtualized mobile clone (vMC) as the service computing providing unit (SCPU), and then the VMs computational resource allocation between CCPU and SCPU is considered to minimize the overall computation power consumption. In reference [26] a mobile clone structure is proposed where users' task information and data are on board in advance so that only indication signal and configuration information are needed for user task execution instead of uploading bulk data, then a total energy minimization problem is formulated which is solved by jointly optimizing the computational resource and transmit power allocation under the constraints of task latency, and fronthaul capacity. Reference [27] integrates the C-RAN with an MEC system and concentrates on the power-performance tradeoff of mobile service provider (MSP) by jointly allocating the radio resource and the computational resource in the MEC server to maximize the revenue of the MSP, however, only electricity costs are considered in this article while radio resource and computational resource costs are omitted.

Different from the previous works, this paper jointly optimizes the offloading strategy, bandwidth allocation, and computational resource assignment in the C-RAN with MEC structure to maximize the profit of network operator.

We consider the computational resource consumption and bandwidth occupation as the cost of a task and the network operator charges an MT for the task computing and output transmission. As for the MTs selection of the offloading strategy, the impact of MTs SE on the network operator profit is analyzed. In particular, the latency constraint is also analyzed in respect to the maximizing of the network operator profit. The main contributions of this paper are listed as follows.

1) A task-aware C-RAN with MEC structure is proposed where a cacheable MEC server is integrated in the BBU pool of the C-RAN. With caching capability in the edge cloud, task data can be prepared in the MEC server in advance by storing data from the Internet purposely or receiving MTs' uploading files. In addition, the MTs that sharing the same MEC server are always closely located and they may prefer similar applications and caching data, hence we can deduce that the MTs of the same MEC server may have similar offloading tasks within a specified period of time which makes the caching more efficient.

2) Communication model and computation model are analyzed respectively, and then the expression of profit function is presented. After that, an optimization problem is formulated by jointly considering the computation offloading strategy policy, bandwidth allocation and computational resource assignment. The objective of the optimization problem is to maximize the economic profit of the network operator while satisfying MTs' QoS (i.e., a minimum time limit for total execution time of task offloading) and kinds of resources constraints. Concretely, network bandwidth, computational resource in the MEC server, and fronthaul capacity of the RRH are limited which will be discussed in more detail below.

3) The optimization problem is decomposed into two sub-problems due to the NP-hard property. On one hand, we propose an SE based offloading strategy policy after analyzing the impact of MTs SE on network operator profit. On the other hand, we use a Lagrangian multiplier method to jointly allocate wireless bandwidth and computational resource for the offloading permitted MTs assuming the offloading strategy is definite. A distributed SE based Joint Optimization for Offloading and Resource Allocation (SJOORA) scheme is proposed to solve the optimization problem by the mutual iteration of the two sub-problems. In addition, the simulation results show the effectiveness of the proposed algorithm.

The remainder of this paper is organized as follows. In Section II, system model and optimization problem are presented. Section III introduces a distributed SJOORA scheme to solve the NP-hard problem efficiently. In Section IV, the simulation results are shown and discussed. Finally, conclusions are given in Section V.

II. SYSTEM MODEL

In this section, a task-aware C-RAN with MEC structure is presented. Specifically, a C-RAN model consisting of a BBU pool and several RRHs is considered as the radio access network and the MEC server is integrated in the BBU pool. First of all, we introduce the network structure and system design, then communication model and computation model are described, and the profit optimization problem is formulated at last. Table 1 shows the main notations to be used in the following sections.

TABLE 1. Notations.

Notation	Definition
N	Total number of RRHs
M	Total number of MTs
$a_{m,n}$	Offloading decision factor
$b_{m,n}$	Spectrum resource allocation factor
c_m	Computational resource allocation factor
B	Total available bandwidth
F	Total computational resource in the MEC server
W_m	Computational intensive task for MT m
D_m	Output data size of task W_m
F_m	Amount of computation for task W_m
$T_{m,max}$	Latency constraint of task W_m
p_n	Transmit power of RRH n
$g_{m,n}$	Channel gain from RRH n to MT m
$e_{m,n}$	Spectrum efficiency of MT m accessing RRH n
$R_{m,n}$	Achievable rate of MT m accessing RRH n
L_n	Fronthaul capacity of RRH n
T_m^{Tr}	Time cost of output data transmission
T_m^{exe}	Time cost of executing task W_m in the MEC server
f_m^{local}	Local computation ability
S_m	Storage cost
p_f	Unit price of charge for computation
p_t	Unit price of charge for transmission
q_b	Unit cost for bandwidth
q_c	Unit cost for computation
κ	Impact factor of bandwidth cost
ω	Impact factor of computation cost

A. NETWORK MODEL

In this paper, a C-RAN with MEC structure is considered. Motivated by [11], the VMs in the BBU pool are divided into two kinds: conventional vBBUs that are responsible for communication and an MEC server that takes charge of task offloading. Here we equip the MEC server with a limited storage pool for caching task data. As mentioned before, the MTs of the same MEC server that located closely may have similar offloading tasks. For example, the MTs located within the areas like museums, shopping malls, airports or railway stations, may have the similar tasks such as Virtual Reality tasks, real-time local information querying and processing. Normally, the same applications are used in these cases. Attribute to the same applications and similar tasks, the caching efficiency and storage space saving are appreciable. Generally, the MTs get these task data from the Internet, and obviously, these data are transmitted from the core network to the BBU pool and then transmitted to the users through RRHs. By the aid of cache capacity of the BBU pool and some previously indication signals or user requests, these task data may be stored in the MEC system temporarily in the transmission process. Besides these task data from the

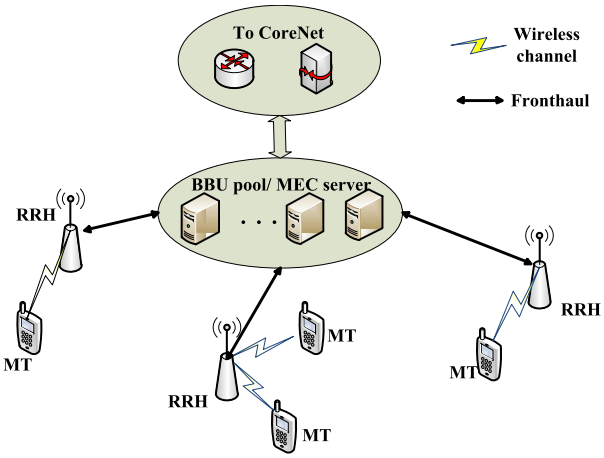

FIGURE 1. System structure of C-RAN with MEC.

Internet, the task computing in the MEC server may need some personal configuration files that come from the MTs themselves, we think that these personal files are small in size and could be uploaded to the edge cloud in idle time purposefully or be settled by a full-duplex mode. In this paper, the cost of personal files' uploading is omitted and only task execution in MEC and computing result transmission are considered. The system structure is illustrated in Fig. 1.

In the C-RAN structure, we denote $\mathcal{N} = \{1, 2, \ldots, N\}$ as the RRHs connecting to the BBU pool through high-speed fiber fronthaul link, which are deployed as Poisson Point Process (PPP) and equipped with single antennas for simplicity. The MTs are also equipped with single antennas and are randomly distributed in the geographical region of the C-RAN, similarly we denote the MTs set as $\mathcal{M} = \{1, 2, \ldots, M\}$. The offloading strategy matrix of MTs is defined as $\mathcal{A} = \{a_{m,n}\}_{M\times\mathcal{N}}$, $m \in \mathcal{M}$, $n \in \mathcal{N}$, where $a_{m,n}$ is a binary variable, i.e., $a_{m,n} = 1$ implies that MT m is permitted to draw support from MEC server to execute its task by accessing RRH n and otherwise $a_{m,n} = 0$. Here we consider a task-awareness case in which the task data of the MTs have already been cached in the MEC server. Similar to [26], we assume that the computational intensive task W_m expected to be accomplished in the MEC server for an MT m is denoted as follows:

$$W_m = (F_m, D_m, T_{m,max}), \quad m = 1, 2, \ldots, M \tag{1}$$

where F_m denotes the total number of the CPU cycles needed to accomplish task W_m in the MEC server; D_m is the whole size of the task's output data transmitted to MT m through the C-RAN after task execution in the MEC server, and $T_{m,max}$ is the latency constraint of task W_m, which is the baseline that the MT m can tolerate for its task W_m.

In addition to the neglect of the cost of personal files' uploading, the time cost in the fronthaul link is also omitted. However the fronthaul capacity is considered as a constraint of the RRH's total data rate which will be discussed later.

B. COMMUNICATION MODEL

In this paper, we consider the full frequency reuse case where the spectrum used by the RRHs is overlaid, thus there exists inter-RRH interference. However, spectrum is orthogonally assigned to each MT that accesses the same RRH, hence intra-RRH interference is ignored here. We assume that all RRHs have the full CSI of all downlinks from any RRH to any MT. The signal-to-interference-plus-noise ratio (SINR) for RRH n to transmit to MT m can be expressed as:

$$SINR_{m,n} = \frac{p_n g_{m,n}}{\sigma^2 + \sum_{j=1, j\neq n}^{N} p_j g_{m,j}} \tag{2}$$

where p_n is the transmit power of RRH n; $g_{m,n}$ represents the channel gain from RRH n to MT m; σ^2 is defined as the power of the additive white Gaussian noise(AWGN) which is assumed to be distributed as $X\mathcal{N}(0, \sigma^2)$. According to Shannon bound, the spectrum efficiency of MT m is given by:

$$e_{m,n} = \log_2(1 + \frac{p_n g_{m,n}}{\sigma^2 + \sum_{j=1, j\neq n}^{N} p_j g_{m,j}}) \tag{3}$$

The total spectrum bandwidth available is B Hz and we denote $b_{m,n} \in [0, 1]$, $\forall m, n$, as the percentage of radio spectrum allocated to MT m by RRH n, thus we have $\boldsymbol{b} = \{b_{m,n}\}$, $m \in \mathcal{M}$, $n \in \mathcal{N}$, as the bandwidth allocation set. Obviously, $\sum_{\mathrm{m}\in\mathrm{M}} \mathrm{b_{m,n}} \leq 1$, $\forall \mathrm{n} \in \mathrm{N}$. Then the achievable instantaneous rate of MT m accessing RRH n, i.e., $R_{m,n}$ is calculated as

$$R_{m,n} = a_{m,n} b_{m,n} B e_{m,n} \tag{4}$$

Consider the fronthaul capacity constraint, we have

$$\sum_{m\in\mathcal{M}} R_{m,n} \leq L_n, \quad \forall n \in \mathcal{N} \tag{5}$$

where L_n is the fronthaul capacity of RRH n.

Finally, the time cost of output data transmission back to MT m from RRH n is given by

$$T_m^{Tr} = a_{m,n} \frac{D_m}{R_{m,n}} \tag{6}$$

C. COMPUTATION MODEL

Assuming the offloading strategy has been decided and an MT m is permitted to have task executed in the cloud, the MEC server who receives the request signal will find the corresponding cached task data and allocate computational resource to start the execution.

We denote F as the total computational resource (i.e., CPU cycles per second) in the MEC server. Here we think all of the computational resource are shared and available for all MTs who are permitted for offloading. Thus we define $c_m \in [0, 1]$, $\forall m$, as the percentage of the computational resource assigned to MT m with the constraint of $\sum_{m\in M} c_m \leq 1$. Similarly, we have $\boldsymbol{c} = \{c_m\}$, $m \in \mathcal{M}$, as the computational resource

allocation set. Then the execution time for task W_m in the MEC server is given by

$$T_m^{exe} = \sum_{n\in\mathcal{N}} a_{m,n}\frac{F_m}{c_m F} \tag{7}$$

Thus the total time cost of the task offloading progress for MT m is given by

$$T_m = T_m^{Tr} + T_m^{exe} \tag{8}$$

In addition, MT m has his locally computation ability which we define as f_m^{local}.Obviously, the computational capability assigned to MT m need to be larger than its local computation ability, or else the offloading will be meaningless. Hence we have the constraint:

$$c_m F \geq a_{m,n} f_m^{local} \tag{9}$$

D. PROFIT FUNCTION

In this paper, we set the maximization of the economic profit of network operator as our optimization target. Concretely, we figure out the revenue and cost of the offloading tasks respectively, and then the profit is their difference.

Firstly, we consider the task revenue. In the MEC server, network operator will charge an MT for the task computing and data caching. As for task computing charge, we consider that the price is related with the task computational complexity and the unit price being charged is defined as p_f per CPU cycle; similarly we define the fee of original task data caching as S_m which is related with the task data size. Then the network operator will charge the MTs for the wireless transmission of output data. Concretely, referring to mobile operators like China Mobile, Vodafone, etc., users are charged by transmission data size rather than user data rate, thus the unit price being charged is defined as p_t per bit. Then the whole revenue for offloading task W_m that involves these three parts is given as.

$$\Omega_m = \sum_{n\in\mathcal{N}} a_{m,n}(p_f F_m + p_t D_m + S_m) \tag{10}$$

Actually, Ω_m is just related with the task itself instead of allocated spectrum and computational resource, which is reasonable and practical.

As for the cost, we consider the spectrum and computational resource occupation for completing the offloading task. Concretely, spectrum resource cost is related to the bandwidth assigned to MT m with the unit price of q_b per Hz and computational resource cost is related to the computational resource assigned to task W_m with the unit price of q_c per CPU cycle/second.

Finally, with the revenue and cost analyzed above, we formulate the profit function of the network operator as:

$$\begin{aligned} U &= \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} (a_{m,n}\Omega_m - a_{m,n}\kappa c_m F q_c - a_{m,n}\omega b_{m,n} B q_b) \\ &= \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}(\Omega_m - \kappa c_m F q_c - \omega b_{m,n} B q_b) \end{aligned} \tag{11}$$

where ω and κ denote the impact factors which represent the tradeoff of scarcity and price fluctuation between the spectrum resource and computational resource respectively. For simplicity, we define $\Gamma_m^C = \kappa F q_c$ and $\Gamma_m^T = \omega B q_b$, then (11) can be rewritten as

$$U = \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}(\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T) \tag{12}$$

Then we have the objective function of the optimization problem

$$\begin{aligned} \max_{a_{m,n},b_{m,n},c_m} & \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}(\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T) \\ \text{s.t. } C1: & \sum_{m\in\mathcal{M}} a_{m,n} b_{m,n} \leq 1, \quad \forall n \in \mathcal{N}, \\ C2: & \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n} c_m \leq 1 \\ C3: & T_m^{exe} + T_m^{Tr} \leq T_{m,\max}, \quad \forall m \in \mathcal{M} \\ C4: & c_m \geq a_{m,n}\frac{f_m^{local}}{F}, \quad \forall m \in \mathcal{M},\ \forall n \in \mathcal{N} \\ C5: & \sum_{m\in\mathcal{M}} R_{m,n} \leq L_n, \quad \forall n \in \mathcal{N} \\ C6: & a_{m,n} \in \{0, 1\}, \quad \forall m \in \mathcal{M},\ \forall n \in \mathcal{N} \end{aligned} \tag{13}$$

Constraint $C1$ guarantees that in each RRH, the sum of bandwidth allocated to all the offloading MTs cannot exceed the total available bandwidth of that RRH. Similarly, $C2$ guarantees that all computational resource allocated to execute tasks cannot exceed total computational capability of the MEC server. $C3$ is the latency constraint which means that the total execution time of the task offloading should satisfy the MT's QoS requirement. $C4$ guarantees that the assigned computational resource to a permitted MT should be more than its local computation capability. Constraint $C5$ means the sum data rate of all the MTs that access a RRH n cannot exceed the fronthaul capacity of the RRH.

For the problem (13), we see that $\{a_{m,n}\}$ are binary variables, and noticing the product relationships between $\{a_{m,n}\}$ and $\{b_{m,n}\}$ as well as $\{c_m\}$, all of this make (13) a mixed integer programming problem which is not convex. With the radically increasing of MTs number, it's extremely complex to solve this non-convex and NP-hard problem directly, thus we have to find an efficient and simplified solution.

III. SE BASED JOINTLY OPTIMIZATION FOR OFFLOADING STRATEGY AND RESOURCE ALLOCATION

In this section, a distributed SJOORA scheme is proposed to solve the optimization problem. Concretely, we firstly analyze the offloading strategy with considering the SE of the MTs and decouple it with the rest resource allocation problem temporarily; then we concentrate on the spectrum and computational resource allocation problem which is solved by Lagrangian multiplier method with specially constraints relaxing; and finally an iterative algorithm is applied to achieve the eventual solution.

A. OFFLOADING STRATEGY ANALYSIS WITH CONSIDERING MTS SE

Considering that $\mathcal{A} = \{a_{m,n}\}$, m $\in \mathcal{M}$, $n \in \mathcal{N}$, is a set of discrete variables, we try to decouple the offloading strategy $\mathcal{A}$ with the spectrum resource allocation factor $\boldsymbol{b}$ and the computational resource allocation factor $\boldsymbol{c}$. From the network operator's perspective, the purpose of handling an MEC system is to win more profit from tasks offloading, hence we can roughly conclude that tasks which consume less resource and/or produce more profit will be preferred when resource shortage situation occurs.

Proposition 1: MTs with higher SE tend to be permitted to offload tasks to the MEC server.

Proof: We define the profit rate for an offloading permitted task as the ratio of task profit and its cost.

$$\gamma_m = \frac{\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T}{c_m\Gamma_m^C + b_{m,n}\Gamma_m^T} \tag{14}$$

For a given task $W_m = (F_m, D_m, T_m)$, Ω_m, Γ_m^C, and Γ_m^T are constant, hence γ_m just varies with c_m and $b_{m,n}$. Here we concentrate on the relationship between SE and profit rate. Since the task computing environment is just the same inside the MEC server, computational resource allocation does not impact the computing cost and latency, for simplicity, we set $c_m = c$ with assuming that $C2$ and $C4$ are satisfied in (13). Then we concentrate on the spectrum resource allocation factor $b_{m,n}$. For a constant revenue Ω_m, the profit of task W_m reaches a maximum when $b_{m,n}$ reaches a minimum and thus is its profit rate γ_m. Ignore $C1$ and $C5$, and transform $C3$, according to (6), (7), and (8) we can get:

$$b_{m,n} \geq \frac{D_m}{Be_{m,n}(T_{m,\max} - \frac{F_m}{cF})} \tag{15}$$

From (14) and (15), it is easy to get

$$\gamma_m \leq \frac{\Omega_m}{c\Gamma_m^C + \frac{D_m\Gamma_m^T}{Be_{m,n}(T_{m,\max} - \frac{F_m}{cF})}} - 1 \tag{16}$$

And then we have

$$\gamma_m^{\max} = \frac{\Omega_m}{c\Gamma_m^C + \frac{D_m\Gamma_m^T}{Be_{m,n}(T_{m,\max} - \frac{F_m}{cF})}} - 1 \tag{17}$$

Fig.2 visually shows the relationship between the upper bound of profit rate and SE. Obviously, the MT with higher SE has higher profit rate. When the MTs are sparsely distributed or in other words, the resource is sufficient, an MT is offloading permitted as long as its task profit is positive. Here we consider a resource limited situation, where numerous MTs compete for task offloading and the network operator makes use of the finite resource to maximize its profit. We analyze (14) from another perspective like this: Γ_m^C, and Γ_m^T can be regarded as the impact factors for c_m and $b_{m,n}$ respectively, then γ_m can be considered as the profit for a weighted unit resource consumption. In other words, higher

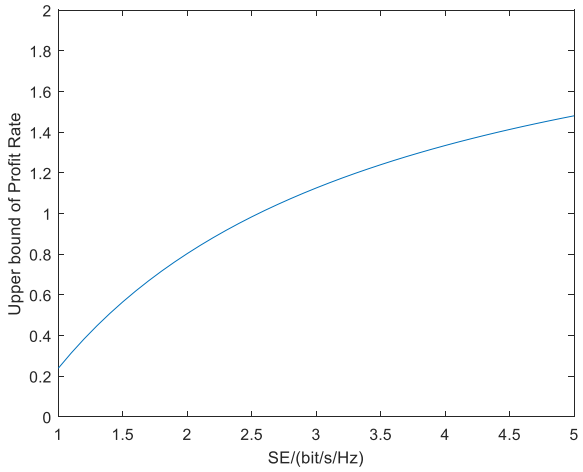

FIGURE 2. Profit rate versus SE for an MT with a given task.

γ_m means higher profit per unit resource. Thus in a resource limited situation, the MTs with higher SE contribute to a lager profit for network operator and *proposition 1* is proved. Note that, this conclusion is not applicable for MTs associated with different RRHs.

Normally, the offloading of computational intensive tasks will cause some additional bandwidth and energy costs, and even unacceptable transmit latency for MTs with poor wireless channel. In some cases, the MTs prefer to execute those tasks locally when additional offloading costs are hard to take, and in these cases the network operator always gains very little profit from the offloading tasks. Based on *proposition 1*, we set preferentially selecting MTs with higher SE as the main idea of our offloading strategy. Note that, the computational resource in the cloud are shared by the whole offloading permitted MTs, in this section we assume computational capability allocation factor $\{c_m\}$ and spectrum resource allocation factor $\{b_{m,n}\}$ are known and constant where $\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T > 0$ is satisfied. For simplicity, the MTs access RRHs that offering the maximum SE in this paper. We iteratively choose part of the MTs with higher SE as the candidates and check whether the constraints in (13) are satisfied. In order to converge faster, half of the unchecked MTs are chosen in each iteration. If the constraints are satisfied, these candidates are permitted for task offloading, or else the candidate with minimum profit is rejected and so are other MTs who access the same RRH with MT m while having lower SE. The detail is shown in Algorithm 1.

After all MTs have been checked, a temporary maximum of profit can be reached with the predefined resource allocation. Next we will introduce the optimization of the spectrum and computational resource allocation under a given offloading strategy in section III-*B*, and the overall algorithm is described in detail in section III-*C*.

B. RESOURCE ALLOCATION USING LAGRANGIAN MULTIPLIER METHOD

In this section, we will jointly solve the spectrum and computational resource allocation problem to maximize the profit

of network operator in the premise that offloading strategy $\mathcal{A}$ is known. Hence only $\{b_{m,n}\}$ and $\{c_m\}$ are concentrated and the optimization problem (13) will become

$$\begin{aligned}
&\max_{b_{m,n},c_m} \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}(\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T)\\
&\text{s.t. } C1: \sum_{m\in\mathcal{M}} a_{m,n}b_{m,n} \le 1, \quad \forall n\in\mathcal{N},\\
&\quad C2: \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}c_m \le 1\\
&\quad C3: T_m^{exe} + T_m^{Tr} \le T_{m,\max}, \quad \forall m\in\mathcal{M}\\
&\quad C4: c_m \ge a_{m,n}\frac{f_m^{local}}{F}, \quad \forall m\in\mathcal{M},\ \forall n\in\mathcal{N}\\
&\quad C5: \sum_{m\in\mathcal{M}} R_{m,n} \le L_n, \quad \forall n\in\mathcal{N}
\end{aligned} \tag{18}$$

Since $\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T$ are linear with respect to $b_{m,n}$ and c_m, hence the objective function is concave. In addition, constraint $C3$ of problem (18) is concave on its domain and the rest constraints are linear, thus problem (18) is a convex optimization problem [28].

Algorithm 1 SE Based Offloading Strategy Optimization With Confirmed Resource Allocation

Input: $W_m, F, B, S_m, p_f, p_t, q_c, q_b, \omega, \kappa, \boldsymbol{b}, \boldsymbol{c}$
Output: $\mathcal{A}$
1. **Initialize:** $\boldsymbol{S_1} = \emptyset; \boldsymbol{S_2} = \emptyset; \boldsymbol{S_3} = \{m|\forall m \in \mathcal{M}\}; K = ||\boldsymbol{S_3}||_0; \boldsymbol{X} = \{x_{m,n} = 0\}_{M\times N}; \mathcal{A} = \{a_{m,n} = 0\}_{M\times N};$
2. **For** $m = 1: M$
3. $e_{m,n} = \max_{i\in\mathcal{N}} e_{m,i}; x_{m,n} = 1;$
4. **End for**
5. **Repeat:**
6. Select $[K/2]$ MTs with higher SE in $\boldsymbol{S_3}$ to be the under testing set, which is denoted as $\boldsymbol{S^*}$, and let $\{a_{m,n} = 1|m \in \boldsymbol{S^*}, x_{m,n} = 1\}$
7. Check whether constraints in (13) are satisfied
8. **If** constraints are satisfied
9. The $[K/2]$ MTs that in $\boldsymbol{S^*}$ will be transferred to $\boldsymbol{S_1}$ and excluded from $\boldsymbol{S_3}$
10. **Else**
11. Calculate U_m and reset $\{a_{m,n} = 0|m \in \boldsymbol{S^*}, x_{m,n} = 1\}$
12. The MT m with minimum profit in $\boldsymbol{S^*}$ and other MTs who access the same RRH with MT m while having lower SE will be transferred to $\boldsymbol{S_2}$ and excluded from $\boldsymbol{S_3}$; The rest MTs in S^* will be transferred back to S_3;
13. **End if**
14. Update $\boldsymbol{S_1}; \boldsymbol{S_2}; \boldsymbol{S_3}; K = ||\boldsymbol{S_3}||_0; \boldsymbol{S^*} = \emptyset;$
15. **Until** $K = 0$
16. The offloading strategy $\mathcal{A}$ is obtained.

Proposition 2: The establishment of latency constraint equation in $C3$ is a necessary condition for achieving the maximum profit when the offloading strategy is known.

Proof: When the offloading strategy $\mathcal{A}$ is known, the revenue Ω_m in (18) is constant, and so are Γ_m^C and Γ_m^T. Obviously, $b_{m,n}$ and c_m should be small enough to reach a maximum profit, however $C1$, $C2$, and $C5$ are the maximum constraint conditions for $b_{m,n}$ and c_m. With the known $\mathcal{A}$, we consider the offloading permitted MTs, i.e., $a_{m,n} = 1$. Assuming $b_{m,n}^*$ and c_m^* are the optimal resource allocation scheme for an MT m where constraints $C1$, $C2$, $C4$, and $C5$ are satisfied.

If we have

$$a_{m,n}\frac{F_m}{c_m^*F} + a_{m,n}\frac{D_m}{b_{m,n}^*Be_{m,n}} < a_{m,n}T_{m,\max} \tag{19}$$

then exist $b_{m,n}^{'} < b_{m,n}^*$ that makes

$$a_{m,n}\frac{F_m}{c_m^*F} + a_{m,n}\frac{D_m}{b_{m,n}^{'}Be_{m,n}} = a_{m,n}T_{m,\max} \tag{20}$$

where constraints $C1$-$C5$ are also satisfied.

According to (12), we can easily get $U(b_{m,n}^{'}, c_m^*) > U(b_{m,n}^*, c_m^*)$. Hence $b_{m,n}^*$ and c_m^* are not the optimal solution which stands in direct contradiction with the hypothesis, and then *proposition 2* is proved.

According to *proposition 2*, $C3$ can be relaxed as

$$a_{m,n}\frac{F_m}{c_mF} + a_{m,n}\frac{D_m}{b_{m,n}Be_{m,n}} = a_{m,n}T_{m,\max}, \quad \forall m\in\mathcal{M} \tag{21}$$

So the optimization problem (18) will become

$$\begin{aligned}
&\max_{b_{m,n},c_m} \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}(\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T)\\
&\text{s.t. } C1: \sum_{m\in\mathcal{M}} a_{m,n}b_{m,n} \le 1, \quad \forall n\in\mathcal{N},\\
&\quad C2: \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}c_m \le 1\\
&\quad C3: a_{m,n}\frac{F_m}{c_mF} + a_{m,n}\frac{D_m}{b_{m,n}Be_{m,n}}\\
&\qquad = a_{m,n}T_{m,\max}, \quad \forall m\in\mathcal{M}\\
&\quad C4: c_m \ge a_{m,n}\frac{f_m^{local}}{F}, \quad \forall m\in\mathcal{M},\ \forall n\in\mathcal{N}\\
&\quad C5: \sum_{m\in\mathcal{M}} R_{m,n} \le L_n, \quad \forall n\in\mathcal{N}
\end{aligned} \tag{22}$$

Considering the given offloading strategy $\mathcal{A}$, the optimization problem may have no valid solution. To make sure we can get a usable solution, we omit constraint $C2$ temporarily which will be checked separately later. And the optimization problem will be transformed as:

$$\begin{aligned}
&\max_{b_{m,n},c_m} \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}(\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T)\\
&\text{s.t. } C1: \sum_{m\in\mathcal{M}} a_{m,n}b_{m,n} \le 1, \quad \forall n\in\mathcal{N},\\
&\quad C3: a_{m,n}\frac{F_m}{c_mF} + a_{m,n}\frac{D_m}{b_{m,n}Be_{m,n}}\\
&\qquad = a_{m,n}T_{m,\max}, \quad \forall m\in\mathcal{M}
\end{aligned}$$

$$C4: c_m \geq a_{m,n}\frac{f_m^{local}}{F}, \quad \forall m \in \mathcal{M}, \ \forall n \in \mathcal{N}$$
$$C5: \sum_{m\in\mathcal{M}} R_{m,n} \leq L_n, \quad \forall n \in \mathcal{N} \tag{23}$$

Then we can write the Lagrangian function of (23) as:

$$\begin{aligned} &L(\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\varphi}) \\ &= \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} a_{m,n}(\Omega_m - c_m\Gamma_m^C - b_{m,n}\Gamma_m^T) \\ &+ \sum_{n\in\mathcal{N}} \beta_n(1 - \sum_{m\in\mathcal{M}} a_{m,n}b_{m,n}) \\ &+ \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} \mu_{m,n}(a_{m,n}(T_{m,\max} - \frac{F_m}{c_mF} - \frac{D_m}{b_{m,n}Be_{m,n}})) \\ &+ \sum_{m\in\mathcal{M}} \eta_m(c_m - \sum_{n\in\mathcal{N}} a_{m,n}\frac{f_m^{local}}{F}) \\ &+ \sum_{n\in\mathcal{N}} \varphi_n(L_n - \sum_{m\in\mathcal{M}} a_{m,n}b_{m,n}Be_{m,n}) \end{aligned} \tag{24}$$

where β_n, $\mu_{m,n}$, η_m and φ_n are the nonnegative Lagrange multipliers corresponding to constraints $C1$, $C3$, $C4$, and $C5$, respectively.

Then, we can write the Lagrangian dual function of (23) as

$$D(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\varphi}) = \max_{b_{m,n}, c_m} L(\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\varphi}) \tag{25}$$

and thus, the dual optimization problem can be formulated as

$$\min_{\beta_n, \mu_{m,n}, \eta_m, \varphi_n} D(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\varphi}) \tag{26}$$

Next we suppose $\{c_m\}$ have been allocated to the MTs and the partial derivative of (24) with respect to $b_{n,m}$ gives the following Karush-Kuhn-Tucker (KKT) condition:

$$\frac{\partial L}{\partial b_{m,n}} = -a_{m,n}\Gamma_m^T - a_{m,n}\beta_n + \frac{a_{m,n}\mu_{m,n}D_m}{b_{m,n}^2 Be_{m,n}} - a_{m,n}\varphi_n Be_{m,n} = 0 \tag{27}$$

And the radio spectrum resource allocation can be performed as

$$b_{m,n} = \sqrt{\frac{a_{m,n}\mu_{m,n}D_m}{(\Gamma_m^T + \beta_n + \varphi_n Be_{m,n})Be_{m,n}}} \tag{28}$$

Again, we suppose $\{b_{n,m}\}$ have been allocated to the MTs and the same as (27), taking the partial derivative of (24) with respect to c_m yields another KKT condition:

$$\frac{\partial L}{\partial c_m} = -a_{m,n}\Gamma_m^C + \frac{a_{m,n}\mu_{m,n}F_m}{c_m^2 F} - a_{m,n}\eta_m = 0 \tag{29}$$

Similarly, the computational resource allocation can be acquired as

$$c_m = \sqrt{\frac{a_{m,n}\mu_{m,n}F_m}{(\Gamma_m^C + \eta_m)F}} \tag{30}$$

At last, together with the third KKT condition i.e., formula (21), we can get $b_{m,n}$, c_m and $\mu_{m,n}$ by jointly solving (21), (28), and (30).

According to reference [29], we can solve the dual problem in (26) by using an iterative subgradient method with updating the Lagrange multipliers like following:

$$\beta_n^{t+1} = [\beta_n^t - \delta_1(1 - \sum_{m\in\mathcal{M}} a_{m,n}b_{m,n})]^+ \tag{31}$$
$$\eta_m^{t+1} = [\eta_m^t - \delta_2(c_m - \sum_{n\in\mathcal{N}} a_{m,n}\frac{f_m^{local}}{F})]^+ \tag{32}$$
$$\varphi_n^{t+1} = [\varphi_n^t - \delta_3(L_n - \sum_{m\in\mathcal{M}} a_{m,n}b_{m,n}Be_{m,n})]^+ \tag{33}$$

where β_n^t, η_m^t and φ_n^t are the values of β_n, η_m and φ_n at the t-th iteration respectively, and δ_1, δ_2 and δ_3 are the positive step sizes that satisfy the infinite travel conditions. The process of updating the spectrum and computational resource allocation, as well as Lagrange multipliers is repeated until convergence or a predefined maximum number of iterations i.e., I_{max} is reached.

C. OVERALL ALGORITHM

In this section, an overall optimization algorithm for the maximization of the network operator's profit is described which jointly optimizes the offloading strategy and spectrum and computational resource allocation.

As shown in Algorithm 2, we divide the MTs into three sets: offloading permitted set $\boldsymbol{S_1}$, unacceptable set $\boldsymbol{S_2}$ and undefined set $\boldsymbol{S_3}$. Just the same with Algorithm 1, an MT accesses the RRH that provides the maximal SE, and we initialize the offloading strategy $\mathcal{A}$ like this: all M MTs are set into $\boldsymbol{S_3}$, so $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are empty. At the beginning of each iteration, the MTs belong to $\boldsymbol{S_3}$ are ranked by their SE in the BBU pool and a bisection method is used when choosing the MTs of $\boldsymbol{S_3}$ as candidates to improve algorithm efficiency, namely, we choose the half of MTs in $\boldsymbol{S_3}$ with higher SE which is denoted as $\boldsymbol{S^*}$. Then we solve problem (23) to get corresponding suboptimal $\boldsymbol{b^*}$ and $\boldsymbol{c^*}$ and the operator profit is obtained at the same time. After that we check whether the operator profit is increased and constraint $C2$ is satisfied or not. If they are, these MTs with higher SE will be transferred to $\boldsymbol{S_1}$ and excluded from $\boldsymbol{S_3}$, else reset all MTs in $\boldsymbol{S^*}$, then the one with minimum profit is rejected and so are other MTs who access the same RRH with this MT while having lower SE, and all of them will be transferred to $\boldsymbol{S_2}$ and excluded from $\boldsymbol{S_3}$. For each iteration, some MTs are transferred to $\boldsymbol{S_1}$ or $\boldsymbol{S_2}$ from $\boldsymbol{S_3}$, and $\boldsymbol{S^*}$ will be emptied, hence we update $\boldsymbol{S_1}$, $\boldsymbol{S_2}$, $\boldsymbol{S_3}$, K and, $\boldsymbol{S^*}$ in the end of each loop. When $\boldsymbol{S}_3$ is empty, we get our expected $\mathcal{A}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ and the algorithm comes to an end.

At last, we give a roughly analysis of the complexity of our proposed algorithm. By using Lagrangian dual method, the asymptotic computational complexity of jointly solving bandwidth and computational resource allocation problem i.e., (23) can be shown as $O(I_{max}((N + 2M)\ N) + 2N))$. Bisection method is used to update the optimal $\mathcal{A}$, so either half of $\boldsymbol{S_3}$ are transferred to $\boldsymbol{S_1}$ or one of the RRHs' offloading strategy is confirmed in each iteration, thus computational

complexity is relatively low. The entire problem will be solved in less than$O(M + logN)$ times. Thus the worst-case complexity of overall algorithm is

$$O((M + logN)I_{max}((N + 2M)N) + 2N)).$$

IV. SIMULATION RESULTS

In this section, we use Monte Carlo simulations to evaluate the performance of our proposed SJOORA scheme. In the following, we firstly introduce the parameter settings, and then simulation results and analysis are presented one by one.

Algorithm 2 SE Based Joint Optimization for Offloading and Resource Allocation

Input: $W_m, F, B, S_m, p_f, p_t, q_c, q_b, \omega, \kappa$
Output: $\{\mathcal{A}\}$ = offloading strategy,
$\{\boldsymbol{b}, \boldsymbol{c}\}$ = resource allocation factors
1. **Initialize**: $\boldsymbol{S_1} = \emptyset; \boldsymbol{S_2} = \emptyset; \boldsymbol{S_3} = \{m | \forall i \in \mathcal{M}\}$;
$K = ||\boldsymbol{S_3}||_0$;
$\boldsymbol{X} = \{x_{m,n} = 0\}_{M\times N}$; $\mathcal{A} = \{a_{m,n} = 0\}_{M\times N}$;
2. **For** $m = 1: M$
3. $e_{m,n} = \max_{i\in\mathcal{N}} e_{m,i}$; $x_{m,n} = 1$;
4. **End for**
5. **Repeat**:
6. Select $[K/2]$ MTs with higher SE in $\boldsymbol{S_3}$ which is denoted as $\boldsymbol{S^*}$, let $\{a_{m,n} = 1 \,|m \in \boldsymbol{S^*}, x_{m,n} = 1\}$;
7. Solve problem (23) to get optimal resource allocation factors $\{\boldsymbol{b^*}, \boldsymbol{c^*}\}$ and calculate $U = \sum_{m\in\mathcal{M}} U_m$
8. Check whether $C2$ in problem (22) is satisfied.
9. **If** $C2$ is satisfied
10. The $[K/2]$ MTs that in $\boldsymbol{S^*}$ will be transferred to $\boldsymbol{S_1}$ and excluded from $\boldsymbol{S_3}$
11. **Else**
12. Reset $\{a_{m,n} = 0 \,|m \in \boldsymbol{S^*}, x_{m,n} = 1\}$
13. The MT m with minimum profit U_m in $\boldsymbol{S^*}$ and other MTs who access the same RRH with MT m while having lower SE will be transferred to $\boldsymbol{S_2}$ and excluded from $\boldsymbol{S_3}$;
The rest MTs in $\boldsymbol{S^*}$ will be transferred back to $\boldsymbol{S_3}$;
14. **End if**
15. Update $\boldsymbol{S_1}; \boldsymbol{S_2}; \boldsymbol{S_3}; K = ||\boldsymbol{S_3}||_0; \boldsymbol{S^*} = \emptyset$;
16. **Until** $K = 0$
17. The offloading strategy $\mathcal{A}$ and optimal resource allocation $\boldsymbol{b}, \boldsymbol{c}$ are obtained.

A. PARAMETER SETTINGS

We consider a simulation scenario like following. The coverage radius of the C-RAN is set to 500m and $N = 10$ RRHs are deployed as PPP within the geographical region. The MTs are also randomly distributed in the same area where the number of the MTs varies from 50 to 120 according to the simulation requirements. For the wireless accessing, the channel bandwidth B is set as 10MHz, the transmit power of each RRH is set as 30dBm and the fronthaul capacity for each RRH is set as 10Mbps. For the wireless channel condition, we set the pass loss model as $37.6 \times \log(dist) + 148.1$ similar to [30], the shadowing factor is given by a log-normal function with standard deviation of 8dB and small scale fading model is independently and identically distributed (i.i.d.) Rayleigh fading with zero mean and unit variance. As for the *AWGN*, we set the noise power as $\sigma^2 = -174$ dBm/Hz. For the MEC server, we set the maximum computation capability F as 100GHz and local computation ability f_m^{local} is set to 0.7GHz [31]. At last, other task parameters as well as price of charging and cost are summarized in Table 2.

TABLE 2. Partial parameters.

Parameter	Value
Output data size D_m	100KB
Latency constraint $T_{m,max}$	600ms
Amount of computation for task F_m	1000× D_m cycle
Storage cost S_m	0.1 $
Impact factor of bandwidth cost κ	0.5
Impact factor of computation cost ω	10
Unit price of charge for computation p_f	0. 03 $/Mega cycle
Unit price of charge for transmission p_t	0.3 $/*Mbit*
Unit cost for bandwidth q_b	0.5 $/*MHz*
Unit cost for computation q_c	0.005 $/Mega cycle/s

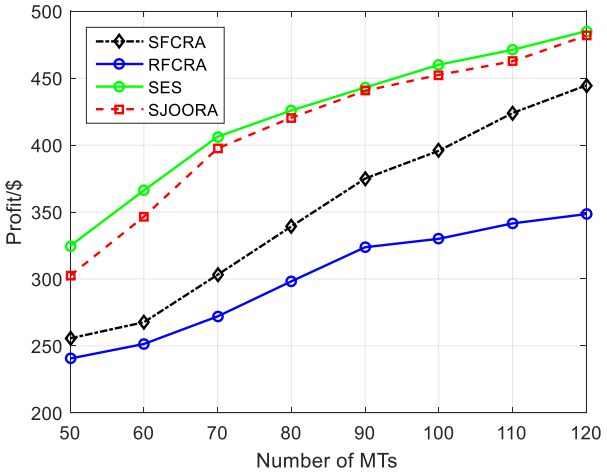

FIGURE 3. Total profit versus MTs number under different schemes.

B. PERFORMANCE EVALUATION OF SJOORA SCHEME

1) ALGORITHM COMPARISON WITH COMPARATIVE METHODS

We evaluate the SJOORA scheme performance to verify its effectiveness by comparing with several baseline algorithms:

Baseline 1: SE based Fixed Computational Resource Allocation (SFCRA). The MEC server selects the MT with maximal SE each time until one of the resource constraints is triggered. The computational resource allocation factor of the offloading permitted MT m i.e. c_m is fixed, which is related to the number of MTs, and then corresponding spectrum resource allocation factor $b_{m,n}$ can be obtained from equation (21).

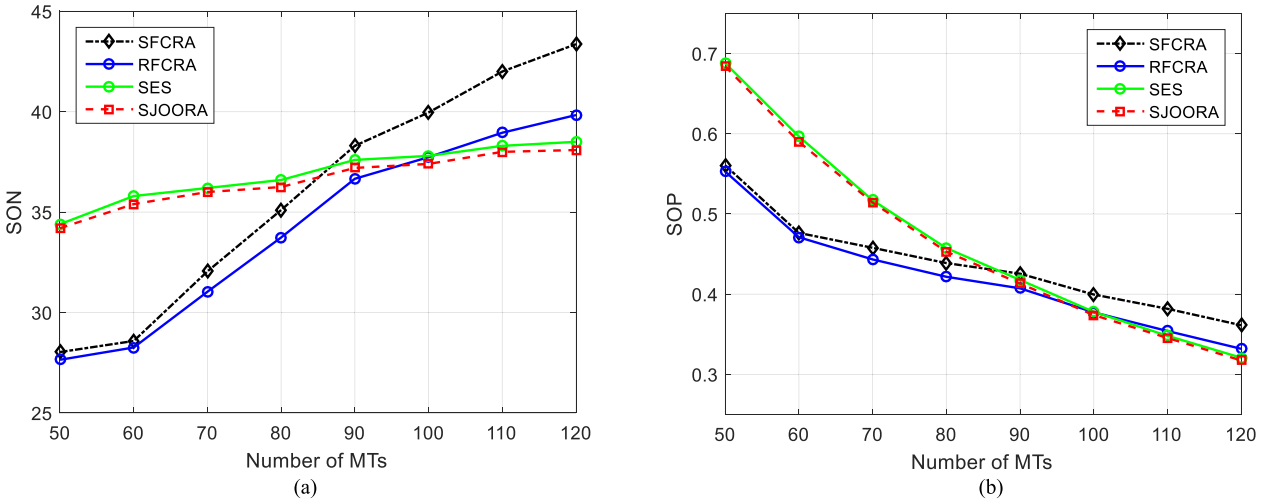

FIGURE 4. SONs (a) and SOPs (b) of different schemes.

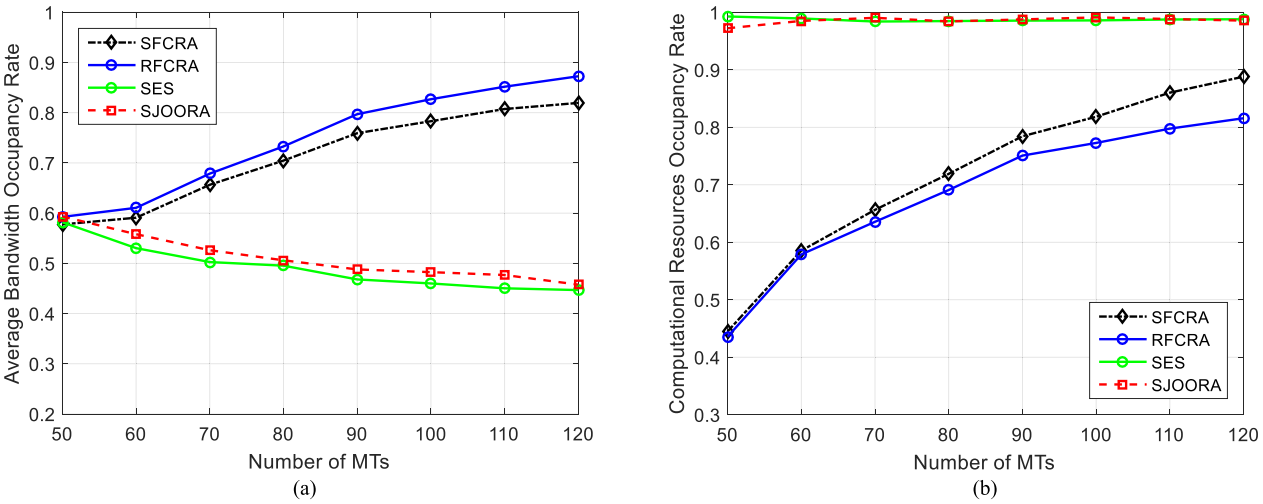

FIGURE 5. The usage of wireless (a) and computational (b) resources under different schemes.

Baseline 2: Random Offloading with Fixed Computational Resource Allocation (RFCRA). In this scheme, the computational resource allocation factor is fixed too, which is the same with SFCRA. The difference with SFCRA is that, in RFCRA, the MEC server randomly selects an MT to offload each time until one of the resource constraints is triggered.

Baseline 3: SE based Exhaustive Search method (SES). Considering the large number of MTs, simple exhaustive method is almost impossible to achieve under our existing condition, hence we use the exhaustive method based on the MTs SE. Concretely, the MEC server selects MTs in order of their SE values and solves (22) to check whether they are suitable for tasks offloading until all MTs have been traversed.

Fig. 3 shows the comparison of the network operator profit versus the number of MTs under different schemes. All of the total profits increase with the growth of the number of MTs under the four candidate schemes, this is because the more MTs means the more MTs with higher SE which make the MEC server have better and more choices to take full advantage of the wireless and computational resources. The gap between SJOORA and SES is quite narrow, which shows that our proposed scheme is quite effective although the SES scheme cannot reach the optimal profit in theory. Moreover, the comparison of SFCRA and RFCRA show the effectiveness of our SE based offloading strategy, and the comparison of SJOORA and SFCRA show the effectiveness of our resource allocation solution by using the Lagrangian multiplier method.

In Fig. 4, we compare the Successful Offloading Numbers (SONs) and Successful Offloading Probabilities (SOPs) of four candidate schemes. SON here means the number of offloading permitted MTs in strategy $\mathcal{A}$ and SOP is the ratio of SON and the number of all candidate MTs. The initial SONs of SFCRA and RFCRA are quite few but grow much

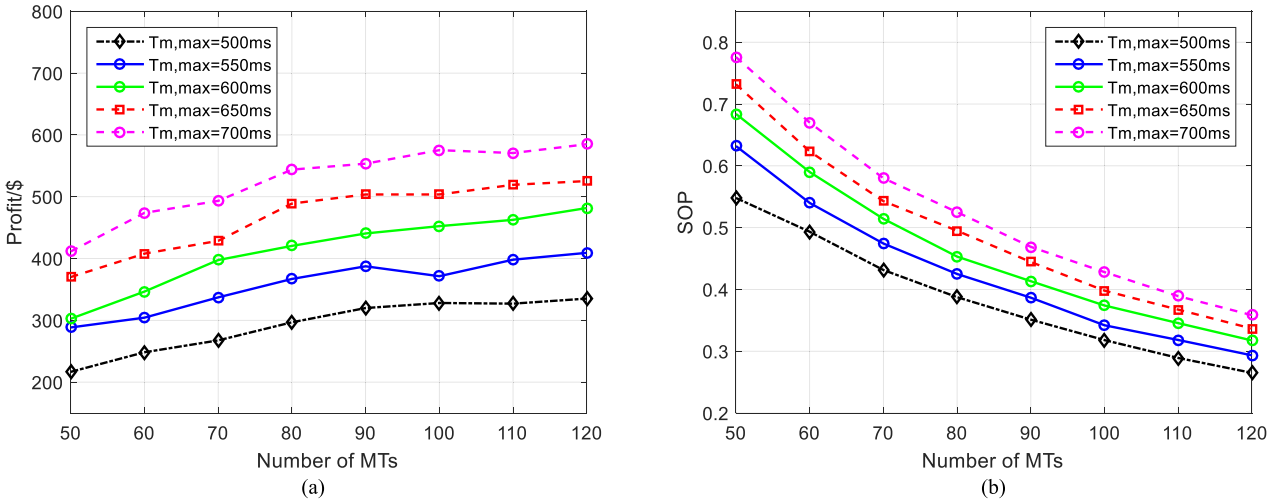

FIGURE 6. Impact of different latency constraints on profit (a) and SOP (b).

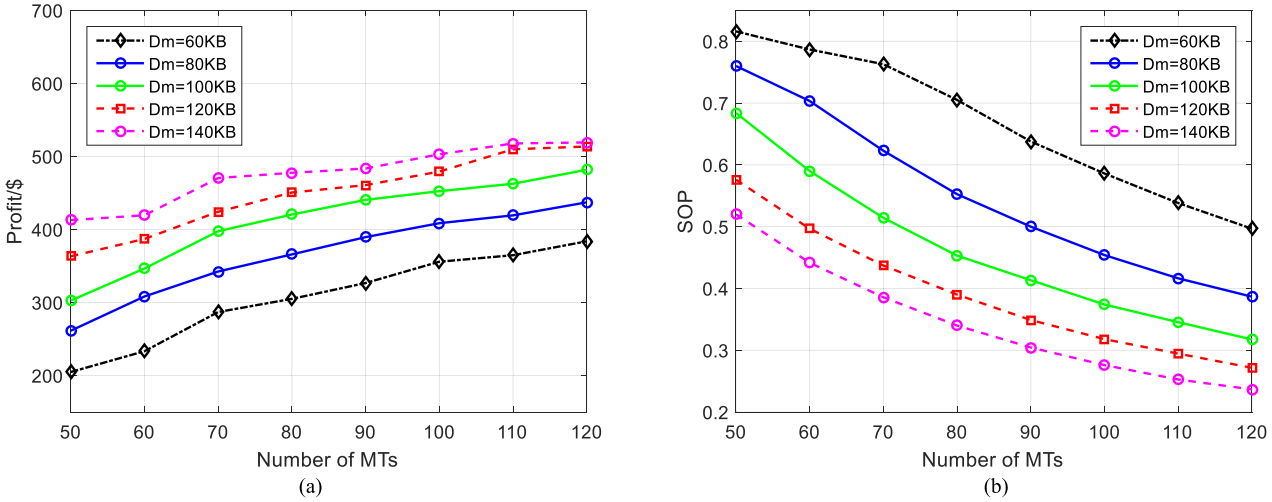

FIGURE 7. Impact of task sizes on profit (a) and SOP (b).

faster than that of SJOORA and SES, this is because the fixed computational resource allocation factors of SFCRA and RFCRA decrease with the growing of the MTs number which help with more MTs to offload. All SOPs decrease with the growing of the MTs number because of the limited wireless and computational resources. SFCRA and RFCRA have higher SOPs than SJOORA and SES due to the fact that the former two schemes fix their computational resource allocation factors, in other words, their purpose is not to maximize their profit which is related to the cost prices of c_m and $b_{m,n}$. In addition, because of the more effective usage of wireless resource, SOP and SON of SFCRA are higher than RFCRA's.

Next, we concentrate on the usage of wireless and computational resources of four candidate schemes as shown in Fig. 5. Obviously, SJOORA and SES who make better use of wireless resource expend less wireless resource than SFCRA and RFCRA, and logically SJOORA and SES consume more computational resource. Concretely, the trend of computational resource usage for SFCRA and RFCRA are similar for the reason of fixed computational resource assignment, however the wireless resource spent in SFCRA is significantly less than RFCRA while the profit of SFCRA is higher on the contrary. Thus the effectiveness of our SE based offloading strategy is verified again. Different with the wireless resource whose utilization efficiency is closely related to MTs' channel condition, the computational resource is fair to all MTs. SJOORA and SES make the best of computational resource and try to avoid overusing the wireless resource under bad channel condition, which efficaciously promote the total profit as we wish.

2) THE IMPACT OF LATENCY CONSTRAINT AND TASK SIZE

After the comparison with comparative methods, we study the impact of different latency constraints and tasks sizes on the total profit in this subsection. Note that, the following simulations are executed under the SJOORA scheme.

Firstly, the profits under different latency constraints with the increase of the MTs number are shown in Fig. 6a. Simulation result shows that profit under bigger $T_{m,max}$ is higher than that of smaller $T_{m,max}$. It is due to the fact that smaller $T_{m,max}$ means stricter constraint, so less MTs are suitable to offload tasks, which will naturally reduce the profit. In addition, Fig. 6b which shows the SOPs under different latency constraints directly verifies the fact that smaller $T_{m,max}$ leads to less offloading MTs.

Then we examine the impact of task size D_m on the total profit. Fig. 7a shows the profits for different tasks sizes D_m and Fig. 7b shows the corresponding SOPs. Obviously, smaller D_m makes more MTs permitted to offload tasks which can be seen intuitively from Fig. 7b. However, more offloading MTs does not mean higher profit for the operator due to the fact that task size is closely related to the cost and revenue. On the contrary, bigger D_m makes profit higher, which is shown in Fig. 7a, although in this case fewer MTs participate in tasks offloading. After comparing and analyzing Fig. 7a and Fig. 7b, we find the fact that bigger D_m can make better use of the MTs with higher SE. For example, there are two MTs where MT *1* has higher SE than MT*2*, MT*1* offloads a $2 \times D$ size task will contribute more profit than each of them offloads a D size task.

V. CONCLUSION

In this paper we proposed a novel task-aware C-RAN with MEC architecture and formulated a profit maximization problem with jointly optimizing offloading strategy and wireless and computational resources allocation. Constraints of offloading task latency, fronthaul capacity, wireless as well as computational resources limitation were considered when solving the optimization problem. To solve the NP-hard problem efficiently, we decoupled the original problem into two sub-problems, and on one hand, an SE based offloading strategy was proposed under the assuming of fixed resource allocation, on the other hand, resource allocation problem with known offloading factors was solved by using the Lagrangian multiplier method. At last, by solving these two sub-problems iteratively, we got a suboptimal solution for the original problem. Simulation results were presented to show the effectiveness of our proposed SJOORA scheme, and the impact of latency constraint and task size were also analyzed.

REFERENCES

[1] S. Barbarossa, S. Sardellitti, and P. D. 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.

[2] B. P. Rimal, D. P. Van, and M. Maier, "Mobile-edge computing vs. centralized cloud computing in fiber-wireless access networks," in *Proc. IEEE INFOCOM WKSHPS*, San Francisco, CA, USA, Apr. 2016, pp. 991–996.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[4] R. Liu, M. Sheng, and W. Wu, "Energy-efficient resource allocation for heterogeneous wireless network with multi-homed user equipments," *IEEE Access*, vol. 6, pp. 14591–14601, 2018.

[5] C. C. Coskun and E. Ayanoglu, "Energy- and spectral-efficient resource allocation algorithm for heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 590–603, Jan. 2018.

[6] A. M. Abdelhady, O. Amin, and M.-S. Alouini, "Resource allocation for phantom cellular networks: Energy efficiency vs spectral efficiency," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[7] A. M. Abdelhady, O. Amin, and M.-S. Alouini, "Energy-efficient resource allocation for phantom cellular networks with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3799–3813, Jun. 2017.

[8] H. Zhang, B. Wang, C. Jiang, K. Long, A. Nallanathan,V. C. M. Leung, and H. V. Poor, "Energy efficient dynamic resource optimization in NOMA system," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5671–5683, Sep. 2018.

[9] China Mobile Research Institute. (Jun. 2014). *C-RAN White Paper: The Road Towards Green RAN*. [Online]. Available: http://labs.chinamobile.com/cran

[10] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

[11] K. Wang and K. Yang, "Power-minimization computing resource allocation in mobile cloud-radio access network," in *Proc. IEEE Int. Conf. Comput. Inf. Technol. (CIT)*, Nadi, Fiji, Dec. 2016, pp. 667–672.

[12] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[13] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint user selection and energy minimization for ultra-dense multi-channel C-RAN with incomplete CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1809–1824, Aug. 2017.

[14] A. U. R. Khan, M. Othman, S. A. Madani, and S. U. Khan, "A survey of mobile cloud computing application models," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 393–413, 1st Quart., 2013.

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

[16] 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.

[17] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, and A. Neal, "Mobile-edge computing introductory technical white paper," ETSI, Sophia Antipolis, France, White Paper, 2014. [Online]. Available: https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/Mobile-edge_Computing_-_Introductory_Technical_White_Paper_V1%2018-09-14.pdf

[18] 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.

[19] 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 ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.

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

[21] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint computation offloading and user association in multi-task mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12313–12325, Dec. 2018.

[22] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[23] Y. Mao, J. Zhang, and K. B. Letaief, "Joint task offloading scheduling and transmit power allocation for mobile-edge computing system," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[24] W. Xia and L. Shen, "Joint resource allocation using evolutionary algorithms in heterogeneous mobile cloud computing networks," *China Commun.*, vol. 15, no. 8, pp. 189–204, Aug. 2018.

[25] J. Plachy, Z. Becvar, and E. C. Strinati, "Dynamic resource allocation exploiting mobility prediction in mobile edge computing," in *Proc. IEEE PIMRC*, Valencia, Spain, Sep. 2016, pp. 1–6.

[26] K. Wang, K. Yang, and C. S. Magurawalage, "Joint energy minimization and resource allocation in C-RAN with mobile cloud," *IEEE Trans. Cloud Comput.*, vol. 6, no. 3, pp. 760–770, Jul. 2018.

[27] X. Wang, K. Wang, S. Wu, S. Di, H. Jin, S. Ou, and K. Yang, "Dynamic resource scheduling in mobile edge cloud with cloud radio access network," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 11, pp. 2429–2445, Nov. 2018.
[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge univ. Press, 2009.
[29] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, Appl. Note EE364b, 2006.
[30] *Physical Layer Aspect for Evolved Universal Terrestrial Radio Access (UTRA) (Release 7)*, document 3GPP TR 25.814, V7.1.0, 2006.
[31] 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. 2015.

ZHANG JIAN was born in Shanxi, China, in 1989. He received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests include the optimization of distributed resource allocation for heterogeneous cellular networks and cloud radio access networks.

WU MUQING was born in 1963. He received the Ph.D. degree in 2001. He is currently a Professor with the Beijing University of Posts and Telecommunications and a Senior Member of the China Institute of Communications. His current research interests include mobile ad hoc networks, UWB, high-speed network traffic control and performance analysis, and GPS locating and services.

ZHAO MIN received the Ph.D. degree in information and telecommunication systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014, where she is currently a Lecturer with the Laboratory of Network System Architecture and Convergence. Her research interest includes wireless communication systems.