Multi-Slice Joint Task Offloading and Resource Allocation Scheme for Massive MIMO Enabled Network

Yin Ren[1], Aihuang Guo[1], and Chunlin Song[1*]
[1] Department of Information and Communication Engineering, Tongji University
Shanghai, 201804 China
[e-mail: renyin@tongji.edu.cn, tigah@tongji.edu.cn, songchunlin@tongji.edu.cn]
[*]Corresponding author: Chunlin Song

Received August 7, 2022; revised January 19, 2023; accepted March 3, 2023; published March 31, 2023

Abstract

The rapid development of mobile communication not only has made the industry gradually diversified, but also has enhanced the service quality requirements of users. In this regard, it is imperative to consider jointly network slicing and mobile edge computing. The former mainly ensures the requirements of varied vertical services preferably, and the latter solves the conflict between the user's own energy and harsh latency. At present, the integration of the two faces many challenges and need to carry out at different levels. The main target of the paper is to minimize the energy consumption of the system, and introduce a multi-slice joint task offloading and resource allocation scheme for massive multiple input multiple output enabled heterogeneous networks. The problem is formulated by collaborative optimizing offloading ratios, user association, transmission power and resource slicing, while being limited by the dissimilar latency and rate of multi-slice. To solve it, assign the optimal problem to two sub-problems of offloading decision and resource allocation, then solve them separately by exploiting the alternative optimization technique and Karush-Kuhn-Tucker conditions. Finally, a novel slices task offloading and resource allocation algorithm is proposed to get the offloading and resource allocation strategies. Numerous simulation results manifest that the proposed scheme has certain feasibility and effectiveness, and its performance is better than the other baseline scheme.

Keywords: mobile communication, massive multiple input multiple output, network slicing, task offloading, resource allocation.

1. Introduction

Due to the diversity of applications and the large-scale growth of traffic, traditional network models and computing methods cannot achieve the unique service quality needs of emerging industries. For guaranteeing the different demands in terms of rate, delay, and terminal scale and solving the contradiction between computing-intensive and delay-intensive applications and limited terminal energy, network slicing (NS) and mobile edge computing (MEC) are both widely used [1]. NS is an indispensable pivotal technology to support the emerging and complex heterogeneous wireless network, such as the Internet of Things (IoT), optical fiber communication networks, and drone networks [2]. Using network function virtualization (NFV) and virtual network function (VNF), network elements can be transformed from the traditional fixed hardware to the software form on the generic hardware [3]. Virtual resources can be dynamically allocated to various vertical applications with specific requirements to build multiple independent logical slices on the same infrastructure [4]. Only the user completes the task, the allocated resources can be dismantled and returned to the infrastructure for the next allocation. Each type of slices, including mobile devices, access, transport, and core network in the network, can logically independently serve customized use cases with unique quality of service (QoS) [5].

An important note, with the explosive growth of sophisticated applications and traffic, the terminal may face insufficient capabilities when processing computing-intensive tasks and time-sensitive applications [6]. However, traditional mobile center cloud (MCC) consumes much energy consumption, because of the long backhaul links, unpredictable delay, and congestion characteristics, which is contrary to the limited terminal energy [7]. Aiming at the shortcomings of traditional MCC, MEC is proposed as a promising computing method, which deploys the server on the access network side, and takes the advantage of short distance to assist users in task processing [8]. As one of the crucial technologies in MEC, computation offloading refers to transferring part or all user tasks to edge servers for computing using the supplied resources by servers. It mainly overcome the problem of poor terminal computing performance and low energy efficiency [9]. Therefore, how to obtain reasonable offloading decisions of tasks is particularly important. Moreover, massive multiple input multiple output (MIMO) is very popular as the preferred technology for current 5G and subsequent network development since it can effectively improve spectrum utilization and reduce transmission delay [10].

In this study, NS and massive MIMO techniques are integrated into MEC to address the problem of coexistence of different vertical industries and make a certain contribution to prolonging the life cycle of terminals. Users can perform partial offloaded tasks and remaining tasks on the server and locally in parallel for slicing delay and rate requirements. This study consists of two levels, for slice level, it clarifies the association between base stations and slice users by comparing processing energy consumption in different servers. For the user level, jointly optimize offloading ratio, user power, and computing resource for partial offloaded tasks. At the same time, the mobile devices can adjust their CPU frequencies to handle non-offloaded tasks. Unlike the previously research, this work aims to minimize the total energy consumption of slice and servers under the conditions of strictly satisfying the latency and rate. For achieving the goal, a massive MIMO enabled multi-slice MEC offloading model is introduced and the two-level problem is formulated as a nonlinear joint optimization problem with multiple equality and inequalities constraints. Given the non-convex properties of the problem, design a new iterative optimization algorithm relying on the alternating optimization and KKT conditions to solve it. Simulation results verify that this scheme has greater

superiority over some baseline offloading algorithms and modes.

Recently, the research on minimizing delay, energy consumption, and the balance between them in MEC-based networks has been solved through wireless and computing resource allocation, without considering network slicing [11]-[14]. In [11], without considering the delay, the authors investigated an energy-saving calculation offloading management scheme, which is suitable for small cell networks and mainly optimizes the energy consumption of all users in the MEC system. In [12] the authors aim to overcome dynamic service requirements of users, jointly optimize user association, D2D mode conversion and spectrum resource in an underlying 5G-HCN with network slicing to minimize the server computing time. The authors in [13] designed a partial offloading framework to minimize unit bit energy consumption. Once the data transmission is compressed, which can save up to 35% energy compared to general transmission. Also, the authors in [14] optimized offloading strategies, bandwidths, and computing resources for minimizing the computation overhead in a heterogeneous MEC network with wireless backhaul. Specifically, time and energy consumption of users are both considered in it.

Currently, some work has been implemented on the fusion of massive MIMO technology and MEC. The authors in [15] designed an offloading algorithm, which is based on the internal primitive dual algorithm and the external delay perception descent algorithm, to study the energy minimization of the system with considering downlink transmission. Additionally, the authors in [16] presented a sequence optimization framework and design a computational offloading scheme to optimize system energy consumption where imperfect channel state information is considered. However, it does not mention the offloading problem. By using binary search and convex optimization methods [17], the authors in [18] given the careful consideration of the different CSI estimation for reducing the maximum energy consumption, which is subject to meeting resources and latency requirements. In [19], the researchers addressed the resource deployment issue by dealing with the offloading decision making and multi-user MIMO precoding. The authors in [20] exploited an algorithm with fractional programming and augmented Lagrangian method to minimize the sum of the deviations between the actual and required latency by an appropriate proportional. Nevertheless, none of the previous works simultaneously consider the offloading decisions and resource optimization in multi-requirement vertical services, and the fusion is crucial for network deployment and low-latency, high-rate communications in the future.

Although the aforementioned research is already sufficient, there are currently a few studies on the sliced network with MEC [21]-[24]. In all researches, the authors in [21] proposed an end-to-end slice and computing resource allocation algorithm to optimize the service delay for URLLC slices. Infrastructure shares wireless and computing resources to multiple virtual network operators (MVNO) in [22] and the authors obtained the optimal resource allocation strategy by minimizing energy consumption and delay. Similarity to [22], the authors in [23] jointly sliced mobile network and resources to minimize the latency of transport, outsourcing, and traffic processing under different types of traffic, network topology, and resources. Furthermore, task offloading, resources allocation, and slices reuse were all considered in a cellular network in [24], while time delay is not involved.

To the best of knowledge, the fusion work of network slicing, MEC, and massive MIMO in the heterogeneous cellular networks has not been discussed of previous studies. So as to ensure the respective delay and rate requirements of vertical industries and reduce the system energy consumption, this study proposes a multi-slice joint offloading and resource allocation scheme with the purpose of jointly optimizing resources on the user side and the server side. The main contributions are introduced in the following

1) Consider the diversified service coexistence scenario with multiple MEC servers in the massive MIMO enabled heterogeneous network, a multi-slice joint task offloading and resource allocation scheme is proposed to achieve the goal of minimizing the system energy consumption. This scheme mainly optimizes the computing offloading and resource allocation of slicing tasks under the premise of satisfying the QoS requirements of different services.

2) Since the joint optimization problem is a mixed integer nonlinear nonconvex problem (MINNP), which cannot be solved directly. Therefore, the original problem is decomposed into two sub-problems, i.e., resource allocation and computation offloading sub-problems, and then propose an effective algorithm to solve each sub-problem iteratively for joint optimization. Specifically, when the offloading coefficient is fixed, the local CPU frequency is first optimized according to the delay requirements. Next, user associations with BSs are determined, which remain unchanged in subsequent problem solving. Meanwhile, the transmission power and server computing resource allocation are optimally by means of D.C programming method and KKT conditions. Then, the resource allocation results can be used as known iteration values to help optimize offloading decisions for slicing users. Finally, the two subproblems are continuously alternately iterated to obtain the optimal solution of the joint optimization problem.

3) Numerous simulation results prove the performance of the proposed scheme with different conditions. Compared with some existing offloading methods and algorithms, the results show that the proposed scheme is superior to the existing methods in terms of energy consumption and convergence speed. The partial offloading can also reduce the energy consumption and better meet the delay requirement.

The rest arrangements of the paper are arranged as follows. Section 2 gives the detailed description of the system model. Section 3 presents the joint formulation and the specific algorithmic solution. Some simulation experiments are performed in Section 4 and the summary is given in Section 5.

2. System Model

As shown in **Fig. 1**, a massive MIMO enabled multi-slice MEC offloading model is described to reflect the heterogeneous feature of the network layout in real environment. It is mainly composed of several base stations (BSs), i.e., $\mathcal{N}=\{0,1,2,...,n,...,N-1\}$, $n=0$ is the macro base station (MBS) with M antennas and $n>0$ denotes the n-th single antenna small base station (SBS). Especially, each BS is equipped with a proximity sever connected by fiber to execute offloading computing for the relevant users. Due to the different QoS requirements of diverse services, for example, some require low delay and high rate, while others require high delay and high rate. In order to better meet the requirements, it can be achieved through network slicing technology. Using the NFV and VFN technologies, resources are abstractly converted into independent virtual network function modules leased to tenants, creating independent slices with specific QoS requirements. In the system, multiple vertical applications are concerned, such as environmental monitoring, autonomous driving, etc. Denote $\mathcal{K}=\{1,2,...,k,...K\}$ as the set of multiple slices, and slice k is represented as $slice_k=(R_k,\tau_k)$, where R_k,τ_k are the minimum rate and maximum delay required of slice k, respectively. The set of users for each slice is $\mathcal{U}=\{1,2,...,u,...,U\}$. Computational task

generated by user u in slice k is $D_{k,u}(t)=\left(d_{k,u},\varphi_{k,u}\right)$, where $d_{k,u}$ is the data size and $\varphi_{k,u}$ is the number of CPU cycles required to process one bit. To accurately guarantee the performance of slices, tasks must be completed at the required rate and within the specified delay. Therefore, it is very necessary to adopt the flexible offloading method, which means that one part of the task is processed locally, and the other part is transferred to the edge server for processing. Besides, the task can be computed independently by the mobile device, or all offloaded to the edge server for computing. This implementation can reduce the energy consumption while meeting latency and rate requirements.

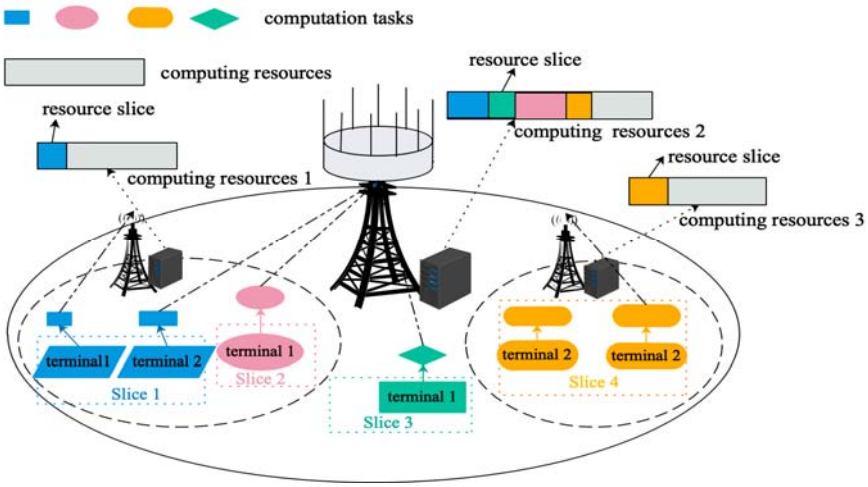

Fig. 1. A massive MIMO enabled multi-slice MEC offloading model in the heterogeneous networks.

2.1 Communication Model

The MBS and SBS occupy different orthogonal frequency bands to avoid cross-layer interference and the spectrum is multiplexed among small base stations (SBSs), where the inter-cell interference is considered. The channel gain is $h_{k,u}^{n}$, which accounts for the shadowing and path loss fluctuating over time. When slice user $u\in\mathcal{U}$ is associated with the MBS, according to Shannon's formula, the uplink rate with the maximum ratio combining can be computed as

$$r_{k,u}^{n}=w_m\log_2\left(1+\left(M-1\right)p_{k,u}h_{k,u}^{n\ 2}\Big/\sigma^2\right),n=0 \tag{1}$$

where w_m, $p_{k,u}$ are the bandwidth and transmission power of the user u, M is the number of antennas installed on the MBS, and σ^2 is the receiver noise. In contrast, when the user u is associated with SBS n, the uplink rate will be defined as

$$r_{k,u}^{n}=w_s\log_2\left(1+p_{k,u}h_{k,u}^{n\ 2}\Big/\left(\sigma^2+Y^2\right)\right),n\neq 0 \tag{2}$$

where $n\in\mathcal{N}$, $Y^2=\sum_{k'\in\{\mathcal{K}\setminus k\}}\sum_{u'\in\{\mathcal{U}\setminus u\}}\sum_{n'\in\{\mathcal{N}\setminus n\}}p_{k',u'}h_{k',u'}^{n'\ 2}$ is the interference in the cell generated by multiple base stations. w_s is the bandwidth allocated by the SBS associated to the user.

In view of the customization requirements of slices and the remaining resources, users can select one server from the base stations to perform offloaded tasks. Introduce

$c_{k,u} \in \{0,1,2,...,n,...,N-1\}$ to indicate the user's offloading location selection. $I(c_{k,u}=n)$ is equal to 1 if the BS n is selected, otherwise $I(c_{k,u}=n)$ is equal to 0. Since base stations can completely cover all types of users, the achievable transmission rate is uniformly expressed as

$$r_{k,u} = \sum_{n=0}^{N-1} I\left(c_{k,u}=n\right) r_{k,u}^{n},\ n \in \mathcal{N} \tag{3}$$

2.2 Task Processing Model

Using the flexible offloading method, this section introduces the task processing model consisting of local computing and offloading computing to ensure the slice delay and rate.

1) Local Computing: Let $\rho_{k,u}$ indicate the proportion of the offloaded tasks to original tasks of user u in slices k, i.e., $\left(1-\rho_{k,u}\right)d_{k,u}$ for local computing, and $\rho_{k,u}d_{k,u}$ for offloading computing. The local delay of executing the fractional task is given by

$$t_{k,u}^{l}\left(\rho_{k,u}, f_{k,u}\right) = \left(1-\rho_{k,u}\right)d_{k,u}\varphi_{k,u} \Big/ f_{k,u} \tag{4}$$

where the values of the parameters $d_{k,u}$ and $\varphi_{k,u}$ depend on the application type, and $f_{k,u}$ is the processing capacity of the terminal, which is further interpreted as the CPU-cycle frequency (cycles/per second). It can be adjusted according to the amount of different tasks, and the maximum is $f_{k,u}^{\max}$. Accordingly, the power consumption of terminal during the local calculation is $p_{k,u} = \omega_{k,u} f_{k,u}^{3}$. Therefore, the local energy consumption is expressed as

$$e_{k,u}^{l}\left(\rho_{k,u}, f_{k,u}\right) = p_{k,u}^{l} t_{k,u}^{l} = \omega_{k,u}\varphi_{k,u}\left(1-\rho_{k,u}\right)d_{k,u} f_{k,u}^{\ 2} \tag{5}$$

where $\omega_{k,u}$ is effective capacitance coefficient depending on different chip structures.

2) Offloading Computing: As for offloading computing, the task is first transmitted to the associated BS through the wireless channel, and then processed by configured server near the BS. Finally, the processed result is transferred back to current user. Owing to the high power of BS and lower processed result, causing the return time is usually ignored [25]. Hence, whether the user is correlated to the MBS or one SBS in this phase, the task processing delay consists of two parts, that is, the task transmitting time and computing time. According to the formula (3), it is easy to get the transmission time and energy consumption of the offloaded task, which can be expressed as follows, respectively

$$t_{k,u}^{tr}\left(c_{k,u}, \rho_{k,u}, p_{k,u}\right) = \frac{\rho_{k,u}d_{k,u}}{r_{k,u}} \tag{6}$$

and

$$e_{k,u}^{tr}\left(c_{k,u}, \rho_{k,u}, p_{k,u}\right) = p_{k,u}\frac{\rho_{k,u}d_{k,u}}{r_{k,u}} \tag{7}$$

where $p_{k,u}$ is the transmission power of the user, and $r_{k,u}^{n}$ is the uplink transmission rate at which the task is delivered to a specific BS. Once the edge server receives a task that need to be computed, it will allocate a portion of CPU frequencies for task calculation. Due to the differences in the computing capacities of MBS and SBS, define $F_{n}^{\max}, \forall n \in \mathcal{N}$ as the

computing resource for BS n. Analogously, the computing time and energy consumption on the MEC server are given by

$$t_{k,u}^{s}\left(c_{k,u},\rho_{k,u},F_{k,u}^{n}\right)=\frac{\rho_{k,u}d_{k,u}\psi_{n}}{\sum_{n=0}^{N-1}I\left(c_{k,u}=n\right)F_{k,u}^{n}},n\in\mathcal{N} \tag{8}$$

and

$$e_{k,u}^{s}\left(c_{k,u},\rho_{k,u},F_{k,u}^{n}\right)=\sum_{n=0}^{N-1}I\left(c_{k,u}=n\right)\omega_{n}\left(F_{k,u}^{n}\right)^{2}\rho_{k,u}d_{k,u}\psi_{n},n\in\mathcal{N} \tag{9}$$

where $F_{k,u}^{n},\psi_{n}$ are the resource allocated for the current user and conversion factor of the server n (i.e., CPU cycles per bit), respectively, and ω_{n} is the hardware coefficient of the server and it is a constant. Consequently, from the above analysis, the time and energy consumption of the entire offloading calculation process can be expressed as

$$t_{k,u}^{o}\left(c_{k,u},\rho_{k,u},p_{k,u},F_{k,u}^{n}\right)=t_{k,u}^{tr}\left(c_{k,u},\rho_{k,u},p_{k,u}\right)+t_{k,u}^{s}\left(c_{k,u},\rho_{k,u},F_{k,u}^{n}\right) \tag{10}$$

and

$$e_{k,u}^{o}\left(c_{k,u},\rho_{k,u},p_{k,u},F_{k,u}^{n}\right)=e_{k,u}^{tr}\left(c_{k,u},\rho_{k,u},p_{k,u}\right)+e_{k,u}^{s}\left(c_{k,u},\rho_{k,u},F_{k,u}^{n}\right) \tag{11}$$

respectively. During the task processing, the local and MEC server computing can be regarded as being performed in parallel. The actual delay is the longer time consuming in local and offloading computing. The specific formula can be described as

$$T_{k,u}=\max\left(t_{k,u}^{l}\left(\rho_{k,u},f_{k,u}\right),t_{k,u}^{o}\left(c_{k,u},\rho_{k,u},p_{k,u},F_{k,u}^{n}\right)\right),\forall u\in\mathcal{U},k\in\mathcal{K},n\in\mathcal{N} \tag{12}$$

3. Problem Formulation and Joint Optimization

The optimization goal is to minimize the overall system energy consumption while maintaining the delay and rate requirements for various slices, and allocate resources and offload tasks for users. Based on the partial offloading feature, it needs to minimize the both the both mobile device and server energy consumption. Notably, the mobile device's energy consumption is concerned with the amount of non-offloaded tasks. The less non-offloaded data sizes, the less energy the terminal consumes, at which point severs will undertake more tasks computing and increase energy consumption. For a given τ_{k} and R_{k} for slice k, it is necessary to comprehensively coordinate the task offloading ratio and resource allocation. Specifically, the system jointly optimizes the offloading ratio, transmission power, user association, and CPU frequencies of terminals and MEC servers, and proposes an effective algorithm to find the most suitable parameter combination to minimize energy consumption and ensure the QoS requirements. The formulation and solution of the problem are given in the following.

3.1 Objective Problem Formulation

The best task offloading ratio and resource allocation strategies are get through optimizing offloading decision $\rho_{k,u}$, user association $c_{k,u}$, transmission power $p_{k,u}$, and CPU frequency

of terminals and servers $f_{k,u}$ and $F_{k,u}^{n}$, respectively. Based on (5)-(11), the energy minimization problem is modeled as

$$(\mathcal{P}): \min_{\substack{f_{k,u},c_{k,u},p_{k,u},\\ F_{k,u}^{n},\rho_{k,u}}} \sum_{k=1}^{K}\sum_{u=1}^{U}\left[e_{k,u}^{l}\left(\rho_{k,u},f_{k,u}\right)+e_{k,u}^{o}\left(c_{k,u},\rho_{k,u},p_{k,u},F_{k,u}^{n}\right)\right] \tag{13}$$

$$\text{s.t.} \quad \text{C1}: 0\le\rho_{k,u}\le 1,\ \ \forall u\in\mathcal{U},k\in\mathcal{K}$$

$$\text{C2}: 0\le p_{k,u}\le p_{k,u}^{\max},\ \ \forall u\in\mathcal{U},k\in\mathcal{K}$$

$$\text{C3}: 0\le f_{k,u}\le f_{k,u}^{\max},\ \ \forall u\in\mathcal{U},k\in\mathcal{K}$$

$$\text{C4}: 0\le F_{k,u}^{n}\le F_{n}^{\max},\forall u\in\mathcal{U},k\in\mathcal{K},n\in\mathcal{N}$$

$$\text{C5}: c_{k,u}\in\left\{0,1,2,...,n,...,N-1\right\}$$

$$\text{C6}: T_{k,u}\le\tau_{k},\ \forall u\in\mathcal{U},k\in\mathcal{K}$$

$$\text{C7}: \sum_{n=0}^{N-1}I\left(c_{k,u}=n\right)r_{k,u}^{n}\ge R_{k},\forall u\in\mathcal{U},k\in\mathcal{K},n\in\mathcal{N}$$

$$\text{C8}: \sum_{k=1}^{K}\sum_{u=1}^{U}\sum_{n=0}^{N-1}I\left(c_{k,u}=n\right)F_{k,u}^{n}\le F_{n}^{\max},\forall u\in\mathcal{U},k\in\mathcal{K},n\in\mathcal{N}$$

$$\text{C9}: I\left(c_{k,u}=0\right)t_{k,u}^{MBS}\le\alpha t_{k,u}^{SBS},\forall u\in\mathcal{U},k\in\mathcal{K}$$

To express more clearly, the BS selection problem is visualized, then the full expression of the problem ($\mathcal{P}$) is

$$\begin{aligned}(\mathcal{P}'): \min_{\substack{f_{k,u},c_{k,u},p_{k,u},\\ F_{k,u}^{n},\rho_{k,u}}} \sum_{k=1}^{K}\sum_{u=1}^{U}\Bigg(& \omega_{k,u}\varphi_{k,u}\left(1-\rho_{k,u}\right)d_{k,u}{f_{k,u}}^{2}+p_{k,u}\frac{\rho_{k,u}d_{k,u}}{\sum_{n=0}^{N-1}I\left(c_{k,u}=n\right)r_{k,u}^{n}} \\ & +\sum_{n=0}^{N-1}I\left(c_{k,u}=n\right)\omega_{n}\left(F_{k,u}^{n}\right)^{2}\rho_{k,u}d_{k,u}\psi_{n}\Bigg),n\in\mathcal{N}\end{aligned} \tag{14}$$

In ($\mathcal{P}$), constraint C1-C4 denote the range of each variable, C5 lists all possible offloading locations, C6 and C7 specify the delay and rate requirements of different slices, and C8 means that the used resources should be within the range of total resources. C9 refers to the comparison of task processing time between the MBS and other base stations. Only when the processing time of the MBS is less than a certain percentage of the SBS processing, the MBS will be selected. Considering the strong processing capability of the MBS, it can support more users. This constraint is more realistic and avoids overloading. According to the different environment, setting appropriate $\alpha\in[0,1]$ in advance. According to the different environment, setting appropriate $\alpha\in[0,1]$ in advance. In special cases, $\alpha=0$, it means that the MBS is unavailable, and $c_{k,u}$ will be randomly selected from small base stations. Otherwise, if $0<\alpha\le 1$ while it is less than the set value of the current environment, the MBS will be selected.

3.2 Joint Optimization

From the above formula, it is noticed that the problem ($\mathcal{P}$) is quite challenging to settle

directly because it is non-convex. By using the alternative optimization and fixing some variables as constants, the problem is separated into two interrelated issues, namely resource allocation and offloading decisions sub-problems. Firstly, the task offloading ratio is given, and the local CPU frequency and user association with BS are optimized according to the delay constraints. Then, considering the different transmission models of heterogeneous networks, it is discussed the optimization problem of the transmission power and computing resources in association with MBS or SBS respectively. Finally, task offloading ratio is optimized using the above optimization parameters. Concrete solutions to two sub-problems are given in the following:

1) Resource Allocation Sub-problem

Given $\rho_{k,u}$, the optimization problem is expressed as

$$(\mathcal{P}_1): \min_{f_{k,u},c_{k,u},p_{k,u},F_{k,u}^n} \sum_{k=1}^{K}\sum_{u=1}^{U}\left[e_{k,u}^l\left(\rho_{k,u},f_{k,u}\right)+e_{k,u}^o\left(c_{k,u},\rho_{k,u},p_{k,u},F_{k,u}^n\right)\right] \tag{15}$$

$$\text{s.t. C2-C9}$$

Obviously, the constraint C6 can be extended the following

$$\text{C6}': t_{k,u}^l\left(\rho_{k,u},f_{k,u}\right)\le\tau_k,\forall u\in\mathcal{U},k\in\mathcal{K}$$

$$\text{C6}'': t_{k,u}^{tr}\left(c_{k,u},\rho_{k,u},p_{k,u}\right)+t_{k,u}^s\left(c_{k,u},\rho_{k,u},F_{k,u}^n\right)\le\tau_k,\forall u\in\mathcal{U},k\in\mathcal{K},n\in\mathcal{N}$$

Problem ($\mathcal{P}_1$) consists of three parts, only the first part is related to the user's CPU frequency and has no embedding relationship with the other parts. According to the C6', it gets $f_{k,u}\ge\left(1-\rho_{k,u}\right)d_{k,u}\varphi_{k,u}/\tau_k$. Meanwhile, it is proved that the first derivative of $e_{k,u}^l\left(\rho_{k,u},f_{k,u}\right)$ is always positive in the given interval. Thus, the optimal CPU frequency of the terminal is $f_{k,u}^*=\left(1-\rho_{k,u}\right)d_{k,u}\varphi_{k,u}/\tau_k$.

After acquiring the optimal value $f_{k,u}^*$, the problem ($\mathcal{P}_1$) is simplified to ($\mathcal{P}_2$), which mainly involves the base station selection, power allocation and resource slicing

$$(\mathcal{P}_2): \min_{c_{k,u},p_{k,u},F_{k,u}^n} \sum_{k=1}^{K}\sum_{u=1}^{U}\left[e_{k,u}^o\left(c_{k,u},\rho_{k,u},p_{k,u},F_{k,u}^n\right)\right] \tag{16}$$

$$\text{s.t. C2, C4-C9}$$

Through observing the problem ($\mathcal{P}_2$), it is unable to solve directly due to the non-convex mixed nonlinear properties. First, deal with the coexistence of discrete and continuous variables to facilitate the solution, appropriately relax the feasible range of discrete variables, $c_{k,u}\in\{0,1,2,...,n,...,N-1\}$ are mapped to $\widetilde{c_{k,u}}\in[0,1]$. Since every user can only select one server for offloading at a time. Taking the energy consumption in MBS and SBS as the basis of selecting the server. The user association problem is given by

$$\left(\mathcal{P}_3\right): \min_{\widetilde{c_{k,u}},p_{k,u},F_{k,u}^n} \sum_{k=1}^{K}\sum_{u=1}^{U}\left(\widetilde{c_{k,u}}e_{k,u}^{SBS}\left(p_{k,u},F_{k,u}^n\right)+\left(1-\widetilde{c_{k,u}}\right)e_{k,u}^{MBS}\left(p_{k,u},F_{k,u}^n\right)\right) \tag{17}$$

$$\text{s.t. C2, C4-C8, C9}': \left(1-\widetilde{c_{k,u}}\right)t_{k,u}^{MBS}\le\alpha t_{k,u}^{SBS}.$$

where $e_{k,u}^{l}\left(\rho_{k,u}, f_{k,u}\right)$ is omitted, as it is a constant in problem ($\mathcal{P}_2$) and ($\mathcal{P}_3$). $e_{k,u}^{SBS}\left(p_{k,u}, F_{k,u}^{n}\right)$, $e_{k,u}^{MBS}\left(p_{k,u}, F_{k,u}^{n}\right)$ are the energy consumption of the current user connected to the relevant SBS and MBS, respectively.

According to the $\mathrm{C9}'$, when $e_{k,u}^{MBS}\left(p_{k,u}, F_{k,u}^{n}\right) \le e_{k,u}^{SBS}\left(p_{k,u}, F_{k,u}^{n}\right)$, the user association is given by $\widetilde{c_{k,u}} = \left[1 - \alpha t_{k,u}^{SBS} / t_{k,u}^{MBS}\right]^{+}$. On the contrary, when it is $e_{k,u}^{MBS}\left(p_{k,u}, F_{k,u}^{n}\right) \ge e_{k,u}^{SBS}\left(p_{k,u}, F_{k,u}^{n}\right)$, the optimal user association is denoted as $\widetilde{c_{k,u}} = \left[\left(\tau_k - t_{k,u}^{tr}\left(\rho_{k,u}, p_{k,u}\right)\right) / t_{k,u}^{SBS}\left(\rho_{k,u}, F_{k,u}^{n}\right), 1\right]^{+}$. Different base stations are selected for association, the interference existing in the system is inconsistent, and there is a certain coupling between the transmission power and the interference. By discussing two cases in which users associate with MBS and SBS respectively, the optimal solution of the relevant variables is finally obtained.

Case I: The slice user *u* is associated with the MBS.

Once the slice user is connected to the MBS, set $\widetilde{c_{k,u}} = 0$. Problem ($\mathcal{P}_3$) will be converted into

$$\min_{\widetilde{c_{k,u}}, p_{k,u}, F_{k,u}^{n}} \sum_{k=1}^{K}\sum_{u=1}^{U}\left(e_{k,u}^{MBS}\left(p_{k,u}, F_{k,u}^{n}\right)\right) \tag{18}$$

s.t. C2, C4, C6-C8

Regarding the total energy consumption $e_{k,u}^{MBS}\left(p_{k,u}, F_{k,u}^{n}\right) = e_{k,u}^{tr}\left(\rho_{k,u}, p_{k,u}\right) + e_{k,u}^{MBS}\left(\rho_{k,u}, F_{k,u}^{n}\right)$, and it represents the energy consumption in the transferring phase and the MBS processing phase. The specific description is

$$(\mathcal{P}_4): \min_{p_{k,u}, F_{k,u}^{n}} \sum_{k=1}^{K}\sum_{u=1}^{U}\left(\frac{p_{k,u}\rho_{k,u}d_{k,u}}{w_m \log_2\left(1+\left(M-1\right)p_{k,u}h_{k,u}^{n\ 2}/\sigma^2\right)} + \omega_n\left(F_{k,u}^{n}\right)^2 \rho_{k,u}d_{k,u}\psi_n \right) \tag{19}$$

s.t. C2, C4, C6-C8

Before discussing resources allocation, Lemma 1 is introduced for a given feasible offloading coefficient, which ensures that user can always obtain resource allocation strategies quickly when the MBS server is used.

Lemma 1. For fixed $\rho_{k,u}$, the objective problem ($\mathcal{P}_4$) is transformed into a convex function when the user is associated with MBS.

Proof. The first term of formula (19) is abstracted into the form $h\left(p_{k,u}\right) = \dfrac{\theta p_{k,u}}{\log_2\left(1+p_{k,u}\beta\right)}$ with respect of variable $p_{k,u}$, where $\theta = \rho_{k,u}d_{k,u}/w_m$ and $\beta = \left(M-1\right)h_{k,u}^{n\ 2}/\sigma^2$. The first derivative function of $h\left(x\right)$ is expressed as

$$\nabla_{p_{k,u}} h\left(p_{k,u}\right) = \frac{\theta\log_2\left(1+p_{k,u}\beta\right) - \dfrac{\theta p_{k,u}\beta}{\left(1+p_{k,u}\beta\right)\ln 2}}{\left[\log_2\left(1+p_{k,u}\beta\right)\right]^2} \tag{20}$$

Define $f\left(p_{k,u}\right) = \theta\log_2\left(1+p_{k,u}\beta\right) - \dfrac{\theta p_{k,u}\beta}{\left(1+p_{k,u}\beta\right)\ln 2}$ in $p_{k,u}$ with

$\nabla_{p_{k,u}} f\left(p_{k,u}\right)=\ln 2\log_2\left(1+p_{k,u}\beta\right)\geq 0$ of $p_{k,u}>0$, and hence $h\left(p_{k,u}\right)$ is monotonically increasing function. Therefore, the problem ($\mathcal{P}_4$) is monotonically increasing with respect to variable $p_{k,u}$. There is no interference term in the rate, and the power can be obtained directly, via Theorem 1.

Theorem 1: For fixing $F_{k,u}^n$, the optimal $p_{k,u}^*$ of (19) is given by

$$p_{k,u}^*=\begin{cases}(2^{\frac{R_k}{w_m}}-1)\eta_m, & if \quad R_k\geq \ell_m \\ (2^{\frac{\rho_{k,u}d_{k,u}}{w_m\left(\tau_k-\rho_{k,u}d_{k,u}\psi_n/F_{k,u}^n\right)}}-1)\eta_m, & if \quad R_k\leq \ell_m\end{cases}, n=0 \tag{21}$$

where $\ell_m=\dfrac{\rho_{k,u}d_{k,u}}{\tau_k-\rho_{k,u}d_{k,u}\psi_n/F_{k,u}^n}$, $\eta_m==\left(\sigma^2/\left(M-1\right)h_{k,u}^{n\ 2}\right)$.

Proof: Utilizing the constraint C6", it gets $r_{k,u}\left(p_{k,u}\right)\geq \ell_m$. If $R_k\geq \ell_m$, the power satisfies $p_{k,u}\geq(2^{\frac{R_k}{w_m}}-1)\eta_m$.Otherwise, if $R_k\leq \ell_m$, it is accepted $p_{k,u}\geq(2^{\frac{\rho_{k,u}d_{k,u}}{w_m\left(\tau_k-\rho_{k,u}d_{k,u}\psi_n/F_{k,u}^n\right)}}-1)\eta_m$. The available power range is further reduced to $p_{k,u}^{\min}\leq p_{k,u}\leq p_{k,u}^{\max}$. $e_{k,u}^{tr}\left(\rho_{k,u},p_{k,u}\right)$ is the monotonically increasing function of $p_{k,u}$, and the optimal $p_{k,u}^*$ will be obtained at the lowest boundary, where the first-derivation of $e_{k,u}^{tr}\left(\rho_{k,u},p_{k,u}\right)$ is equal to 0.

Case II: The slice user *u* is associated with SBS *n*.

When the user u is associated with a certain SBS, set $\widetilde{c_{k,u}}=1$ .For a given offloading coefficient, the problem ($\mathcal{P}_3$) is described as

$$\min_{\widetilde{c_{k,u}},p_{k,u},F_{k,u}^n}\sum_{k=1}^{K}\sum_{u=1}^{U}\left(e_{k,u}^{SBS}\left(p_{k,u},F_{k,u}^n\right)\right) \tag{22}$$

Similar to Case I, $e_{k,u}^{SBS}\left(p_{k,u},F_{k,u}^n\right)=e_{k,u}^{tr}\left(\rho_{k,u},p_{k,u}\right)+e_{k,u}^{SBS}\left(\rho_{k,u},F_{k,u}^n\right)$ is the offloading energy consumption when connected to the selected server. The formula (18) is expanded and rewritten as

$$(\mathcal{P}_5):\min_{p_{k,u},F_{k,u}^n}\sum_{k=1}^{K}\sum_{u=1}^{U}\left(\frac{p_{k,u}\rho_{k,u}d_{k,u}}{w_s\log_2\left(1+p_{k,u}h_{k,u}^{n\ 2}/\left(\sigma^2+Y^2\right)\right)}+\omega_n\left(F_{k,u}^n\right)^2\rho_{k,u}d_{k,u}\psi_n\right) \tag{23}$$

s.t. C2, C4, C6-C8

Due to the existence of interference terms in $r_{k,u}$, the problem ($\mathcal{P}_5$) is certified non-convex. For facilitating the solution, utilizing the D.C programming method to convert the rate function to the difference of two logarithmic functions, which is expressed as

$$\begin{aligned} r_{k,u}\left(p_{k,u}\right) &= w_s \sum_{n=1}^{N-1} I\left(c_{k,u}=n\right)\left[\log_2\left(1+p_{k,u}h_{k,u}^{n\ 2}/\left(\sigma^2+Y^2\right)\right)\right] \\ &= w_s \sum_{n=1}^{N-1} I\left(c_{k,u}=n\right)\left[z\left(p_{k,u}\right)-s\left(p_{k,u}\right)\right] \end{aligned} \tag{24}$$

In (24),

$$z\left(p_{k,u}\right)=\log_2\left(p_{k,u}h_{k,u}^{n\ 2}+\sigma^2+Y^2\right) \tag{25}$$

$$s\left(p_{k,u}\right)=\log_2\left(\sigma^2+Y^2\right) \tag{26}$$

To optimize the power distribution, the first-order Taylor expansion $\hat{s}\left(p_{k,u}\right)$ of $s\left(p_{k,u}\right)$ is approximated as

$$\hat{s}\left(p_{k,u}\right)=s\left(p_{k,u}^{(l)}\right)+<\nabla s\left(p_{k,u}^{(l)}\right),p_{k,u}-p_{k,u}^{(l)}> \tag{27}$$

where $p_{k,u}^{(l)}$ is the optimized value of $p_{k,u}$ obtaining after *l* iteration. The rate formula $r_{k,u}$ can be updated as

$$\widehat{r}_{k,u}^{(l)}\left(p_{k,u}\right)=w_s\sum_{n=1}^{N-1} I\left(c_{k,u}=n\right)\left[z\left(p_{k,u}\right)-\hat{s}^{(l)}\left(p_{k,u}\right)\right] \tag{28}$$

The optimization problem ($\mathcal{P}_5$) can be transformed into:

$$(\widetilde{\mathcal{P}_5}):\min_{p_{k,u},F_{k,u}^n}\sum_{k=1}^{K}\sum_{u=1}^{U}\left(\frac{p_{k,u}\rho_{k,u}d_{k,u}}{w_s\sum_{n=0}^{N-1} I\left(c_{k,u}=n\right)\left[z\left(p_{k,u}\right)-\hat{s}\left(p_{k,u}\right)\right]}+\omega_n\left(F_{k,u}^n\right)^2\rho_{k,u}d_{k,u}\psi_n\right) \tag{29}$$

$$\text{s.t C2, } \widetilde{\text{C6}}:\rho_{k,u}d_{k,u}-\widehat{r}_{k,u}\left(p_{k,u}\right)\left(\tau_k-\rho_{k,u}d_{k,u}\psi_n\Big/\sum_{n=0}^{N-1} I\left(c_{k,u}=n\right)F_{k,u}^n\right)\le 0$$

Initializing a $p_{k,u}^{(0)}$, the problem ($\widetilde{\mathcal{P}_5}$) can be solved iterative. $p_{k,u}^{(l)}$ is the power value updated after *l* iteration. Calculate the normalized error of $p_{k,u}^{(l)}$ at *l* iterations and $p_{k,u}^{(l)}$ at (*l*-1) iterations, which can be used as an admission criterion for whether to continue the iteration. Only $\eta_{k,u}=\left|p_{k,u}^{(l)}-p_{k,u}^{(l-1)}\right|\Big/p_{k,u}^{(l-1)}$ is less than a given minimum value, the optimal power is $p_{k,u}^{*}=p_{k,u}^{(l)}$. Conversely, it will continue to iterative update until the limit is reached. Under the premise of a given offloading decision, both the local computation and transmission energy consumption are constants. The optimal computing resource slices for user is obtained as

$$(\mathcal{P}_6):\min\sum_{k=1}^{K}\sum_{u=1}^{U}\left(\omega_n\left(F_{k,u}^n\right)^2\rho_{k,u}d_{k,u}\psi_n\right) \tag{30}$$

$$\text{s.t. C4, C6", C8.}$$

The Lagrangian function of (30) is described as

$$L\left(F_{k,u}^{n},\varsigma_{k,u},v\right)=\sum_{k=1}^{K}\sum_{u=1}^{U}\left(\omega_n d_{k,u}\psi_{k,u}\rho_{k,u}\left(F_{k,u}^{n}\right)^2\right)+\sum_{k=1}^{K}\sum_{u=1}^{U}\varsigma_{k,u}\left(\ t_{k,u}^{tr}\left(\rho_{k,u},p_{k,u}\right)+t_{k,u}^{s}\left(\rho_{k,u},F_{k,u}^{n}\right)-\tau_k\right)$$
$$+\sum_{k=1}^{K}\sum_{u=1}^{U}v\left(F_{k,u}^{n}-F_{n}^{\max}\right),n\in\mathcal{N} \tag{31}$$

where $\varsigma\geq 0,v\geq 0$ are the Lagrangian multipliers of C6",C8, then the KKT conditions are applied to get the optimal $F_{k,u}^{n}$.

2）Offloading Decision Sub-problem

Given the fixed $f_{u,k},c_{u,k},p_{k,u},F_{k,u}^{n}$, the offloading decision problem is expressed as

$$\left(\mathcal{P}_7\right):\min_{\rho_{k,u}}\sum_{k=1}^{K}\sum_{u=1}^{U}\left(\ \frac{\omega_{k,u}\varphi_{k,u}^{3}d_{k,u}^{3}\left(1-\rho_{k,u}\right)^3}{\tau_k^2}+p_{k,u}\frac{\rho_{k,u}d_{k,u}}{r_{k,u}}+\omega_n\left(F_{k,u}^{n}\right)^2\rho_{k,u}d_{k,u}\psi_n\ \right) \tag{32}$$
$$\text{s.t. C1, C6, C10: }\left(1-\rho_{k,u}\right)d_{k,u}\varphi_{k,u}-\tau_k f_{k,u}^{\max}\leq 0,\quad \forall u\in U,k\in\mathcal{K}$$

It is easy to find the problem ($\mathcal{P}_7$) and constraints are affine with respect to $\rho_{k,u}$ if the user associations are decided. The optimal solutions can be obtained in polynomial time.

3.3 Algorithm Implementation and Complexity

The realization process of the scheme can be divided into four parts. First, on the basis of alternating optimization, the objective problem is split into offloading decision and resource allocation. With the fixed offloading ratio, local CPU-frequency is adjusted using (15). Next, utilize (17) to find the association between slice user and the base station. Then, under the coverage of the MBS and SBS, the optimal power can be obtained according to (21) and (27), respectively, and the computing resources of server is sliced to provide services for the offloading task using (31). Finally, the offloading decision problem is optimized in the current iteration with the known resources. The above process repeated until convergence and the specific implementation is described in Algorithm 1. The execution complexity of the algorithm is $O\left(J\times U^{'}\times Z^3\right)$, where J, $U^{'}$, and Z are the number of iterations, users, and optimized variables.

Algorithm 1 Multi-slice joint task offloading and resource allocation

1: **Set** a suitable initial value for $\rho_{k,u},k\in\mathcal{K},u\in\mathcal{U}$
2: **Set** $j=1$, $\varepsilon_1=0.001$, $e_{k,u}=0$, $k\in\mathcal{K},u\in\mathcal{U}$
3: **while** $|e_{k,u}(j+1)-e_{k,u}(j)|>\varepsilon_1$ or $j\leq 20$ do
4: **For** slices *k*=1 to *K* do
5: **For** user *u*=1 to *U* do
6: Compute $f_{k,u}^{*}(j)$ from $f_{k,u}^{*}=\left(1-\rho_{k,u}\right)d_{k,u}\varphi_{k,u}/\tau_k$
7: At a given $\rho_{k,u}(j)$, compute $c_{k,u}(j)$ using (17)
8: **if** $c_{k,u}(j)=0$, do
9: Compute $p_{k,u}(j)$, and $F_{k,u}^{n}(j)$ at a given $\rho_{k,u}(j)$ by using (21) and (31)

else
  Initializing $\eta_{k,u}$=1, l=0, and a suitable power $p_{k,u}^{(0)}$
  While $\eta_{k,u}$>0.01, do
    $l = l + 1$
    Compute $\widehat{r_{k,u}}\left(p_{k,u}{}^{(l)}\right)$ using (28)
    Obtain the optimal $p_{k,u}{}^{(l)}$ using (29)
    Calculate $\eta_{k,u} = \left|p_{k,u}^{(l)} - p_{k,u}^{(l-1)}\right| / p_{k,u}^{(l-1)}$ at each iteration.
  end
  Compute $F_{k,u}^{n}(j)$ by using (31)
 end
 Calculate offloading decision $\rho_{k,u}(j)$ using (28).
end
end
Updating system total energy consumption by adopting (10)
Updating $j = j + 1$, go to 3 until convergence.
end

4. Simulation Implementation

To assess the proposed scheme, a heterogeneous network scenario is considered in this section, consisting of one MBS with 100 antennas and three single-antenna SBSs. The MBS is located in the center of the network and has a coverage of 100 meters, and three SBSs are randomly distributed within the range of the MBS with a coverage of 20 meters. The total bandwidths of MBS and the SBSs are 100MHz and 70MHz. The wireless bandwidths allocated to users are 10M and 5M, which are determined by linking to the MBS and SBSs. Further, the computing capabilities of edge servers connected to the MBS and the SBS are16GHz and 8GHz [26]. The computing capacities of mobile devices can be adjusted between the interval [0,2] GHz. The path loss is -128.1-37.6$\log_2(d)$, and the noise power is 172 dBm [18]. The user power factor $\omega_{k,u}$ is set to 10^{-27}. In addition, the maximum number of iterations j is set to 20, and the iteration termination value is set to 0.001. The mobile device need revolutionize 100cycles per bit, i.e., $\varphi_{k,u} = 100 cycles / bit$. In the same way, the server needs to convert 40cycles to process one bit, i.e., $\psi_n = 40 cycles / bit$. According to the 5G slices categories, the experiment defines four types of slices, which are Low Latency-Low rate Slices (LLS), Low Latency-High Rate Slices (LHS), High Latency-Low Rate Slices (HLS), and High Latency-High Rate Slices (HHS) [27]. Different types of slices own the different data sizes, maximum delay, and minimize rate requirements. Each type of slice can provide services for 8 users, and the maximum power of users is 23dbm. The partial parameters details of slices are summarized in **Table 1**.

Table 1. Simulation Parameters

Parameters	Value
Data size	{LLS: 1×10^6, LHS: 2×10^6, HLS: 1×10^6, HHS: 2×10^6 } (bits)
maximum delay	{LLS:0.1, LHS:0.1, HLS:0.3, HHS:0.3} (s)
minimum rate	{LLS:100 LHS:500, HLS:100, HHS:500} (kbps)

Fig. 2 draws the convergence performance of the proposed algorithm with different antenna numbers (M=100, 200, 300, 400). Simulation diagram reveals that the algorithm gradually converges after several iterations. It is clearly found that the number of antennas has a significant impact on energy consumption, but has little effect on the convergence speed. The increase in the number of antennas leads to lower energy cost and faster convergence. The reason is that the more antennas numbers, the higher the transmission rate and the lower the transmission delay, resulting in relatively low offloading energy consumption. In addition, more users can choose the MBS for processing, which avoids the interference among users and reduces total energy overhead. To verify the performance of the proposed algorithm, **Fig. 3** evaluates the system energy overhead by comparing proposed algorithm with several other baseline algorithms, which are introduced as follows

1) Full offloading algorithm [24]: All tasks of users are offloaded to the associated BS for processing using the corresponding server. During this process, the offloading ratios are set to1, and the local CPU frequencies are not optimized. Other optimization parameters, such as transmission power, server computing resources, are jointly optimized by Algorithm 1.
2) Local computing algorithm [24]: The slicing tasks are handled entirely locally, and the system energy consumptions depend only on the terminal energy consumption. The local CPU frequencies scheduling are determined by Algorithm 1.
3) Average allocation algorithm [24]: Only the computing resources provided by the MEC servers for offloading tasks are distributed equally, and other parameters such as offloading ratio, transmission power are jointly optimized by Algorithm 1.
4) BFGS algorithm [28]: It is one of the quasi-Newtonian algorithms, which uses quasi-Newton method to directly approximate the Hessian matrix, and then jointly optimize each variable.

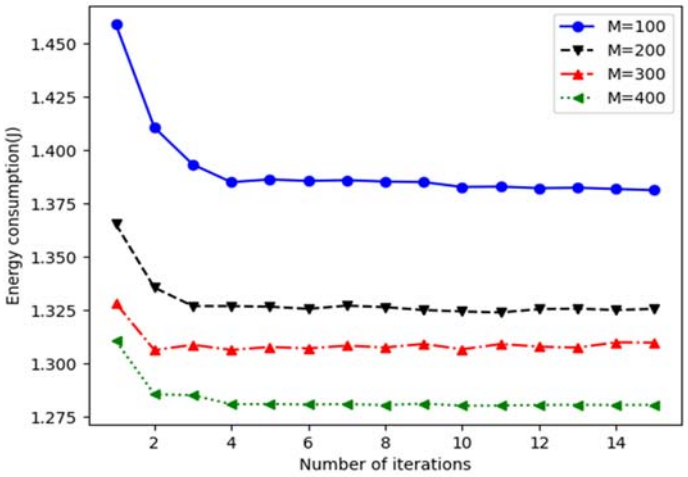

Fig. 2. Energy consumption versus different number of antennas.

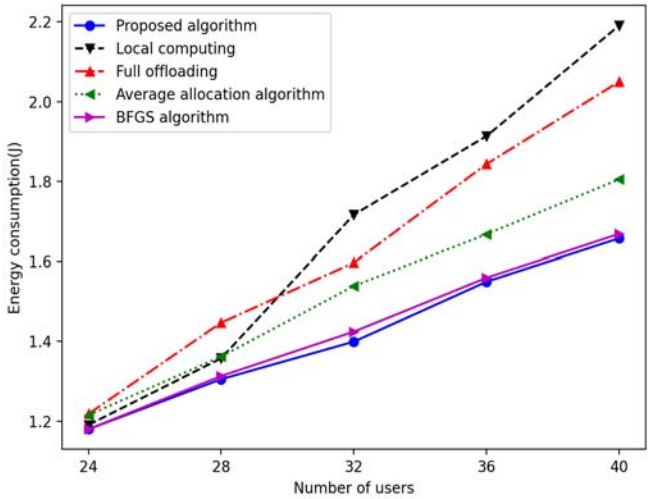

Fig. 3. Energy consumption versus different number of users with five algorithms.

As can be seen from the **Fig. 3**, energy consumption is positively correlated with the number of sliced users, and the increasing of five algorithms is different. This figure demonstrates that the least energy consumption using the proposed algorithm, followed by the BFGS. The average distribution, local offloading and full offloading algorithms consume relatively more energy. The proposed scheme utilizes partial offloading to execute tasks in parallel with the slicing resources from the perspective of terminals and servers, which can effective meet user's latency and rate requirements. At the same time, as the rate increases and delay decreases, the system energy consumption also decrease. Besides, as the further supplement, **Fig. 4** depicts the energy consumption of the above algorithms under different slices, respectively. When only considering a class of slices, the performance of the scheme is also comparatively better than other schemes.

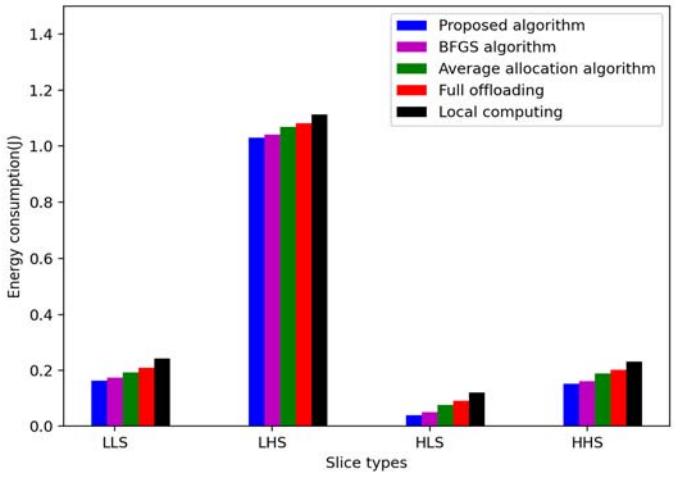

Fig. 4. Energy consumption versus different slices with five algorithms.

Fig. 5 plots the energy consumption of various slices versus the different number of users and antennas. As can be seen, the LHS consumes the most energy owing to its the highest requirements for delay and rate, and requires the server to provide more resources. HLS consumes the least energy and has no strict requirements on delay and rate. Likewise, LLS and HHS have only strict delay or rate requirements, so the energy consumption is moderate.

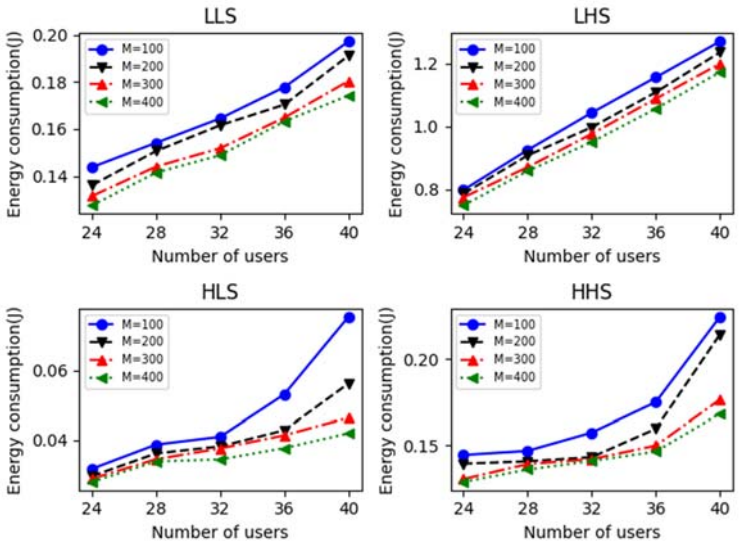

Fig. 5. Energy consumption of various slices versus different number of users.

Fig. 6 analyzes the overall energy expenditure under slice dissimilar data sizes and delay. First, remain the data sizes unchanged and expand the delay to {0.2, 0.2, 0.6, 0.6} s, which require fewer resources to complete the task within a wide period of time with the corresponding saving in energy. Intuitively, energy consumption is positively related to data sizes for each type of slices. Contrary to the above settings, expanding the data sizes to $\{2\times10^6,\ 4\times10^6,\ 2\times10^6,\ 4\times10^6\}$bits in the case of constant delay, and the system energy loss also drops significantly with more resource allocation for jointly processing in a shorter time. Meanwhile, **Fig. 7** studied the effects of maximum delay and minimum rate on overall energy consumption. Similar to the **Fig. 6**, doubles the delay while keeping the rate constant, it will require fewer resources to complete the task and produce less energy consumption. In contrast, if the delay does not change and the rate is doubled, the task processing will need more resources to maintain the new rate. Thus, the transmission power will increase accordingly, resulting in an increase in the energy consumption of the system.

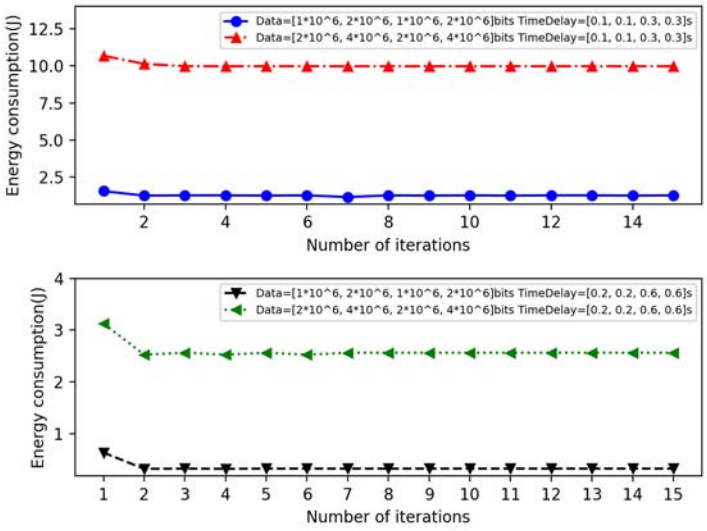

Fig. 6. Energy consumption versus different data sizes and delay.

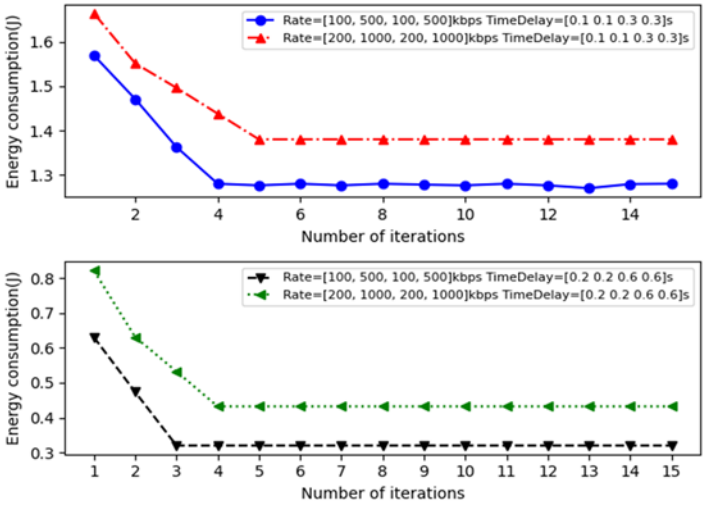

Fig. 7. Energy consumption versus different rate and delay requirements.

Fig. 8 further verifies the performance of the proposed algorithm and compares it with the existing methods, such as 1) JCCRAVM algorithm [29]: where only considered the resource allocation problem with the full offloading policy; 2) JTORAA algorithm [30]: The offloading policy is binary-based partial offloading, i.e., the whole task is as an offloading target, where a portion of tasks are executed locally, and the other tasks are executed by the MEC servers. At the same time, jointly optimized the user association, transmission power, and computing resources allocation of users and MEC severs. It is clear seen from the **Fig. 8** that the energy consumption of these three algorithms decreases with the number of iterations increasing, and the proposed algorithm requires fewer iterations to obtain the optimal solution. Compared with algorithms that considering the binary and full offloading strategies, the method with partial offloading can better save energy consumption. This is because the partial offloading decision can divide the task into two parts to process in parallel, and allocating resources for local and offloaded tasks independently, which reduces the task processing delay and the number of exchanges in the resource allocation phase. Furthermore, from the perspective of convergence,

it can be discovered that when reaching the lowest energy consumption, the proposed algorithms, JCCRAVM and JTORAA algorithm require 5, 10, 3 iterations, respectively. The proposed algorithm in this paper can complete the calculation more efficiently with less energy consumption, and converges more easily than the JCCRAVM algorithm, but not as quickly as the JTORAA algorithm.

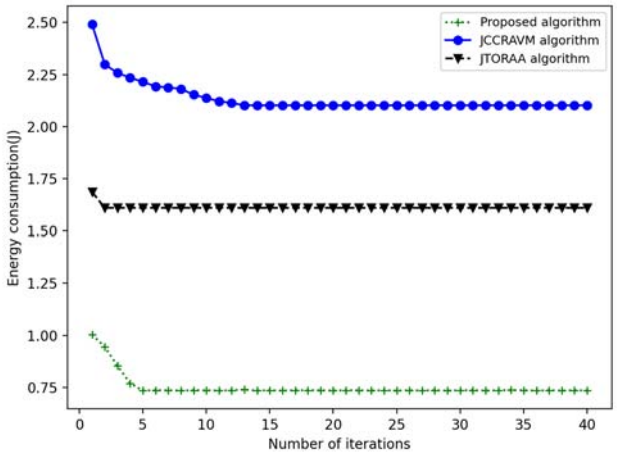

Fig. 8. The convergence of energy consumption under different algorithms.

Fig. 9 shows the impact of different numbers of SBSs on system energy consumption. When the number of SBSs is fixed, as the number of users increases, so does the number of tasks that users can offload. Therefore, under the limited system resources, the system energy consumption will gradually increase with the number of users. From another perspective, when the number of users is the same, the energy consumption will gradually decrease with the increase of the number of SBSs. The main reason is that the growth of SBSs will provide more computing and bandwidth resources, which in turn supports more offloading tasks and reduces energy consumption.

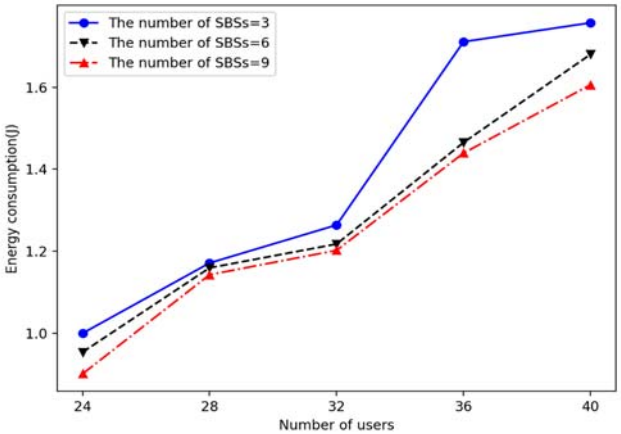

Fig. 9. Energy consumption with different number of SBSs and users.

5. Conclusion

In this paper, an energy efficient multi-slice task offloading and resource allocation scheme is developed, which guarantees the different QoS requirements of slices, and integrates massive MIMO technology and edge computing technology in heterogeneous networks. This scheme can work for multiple vertical industry coexistence scenarios with limited energy. The user association, power control, and resource slicing problem are jointly optimized to minimize the system energy consumption, and the optimal solution is obtained by using alternating optimization and KKT conditions. Through simulation and comparison with different offloading methods and different resource optimization algorithms, results prove the effectiveness of the proposed scheme. For slices with different configurations, the scheme can achieve targeted offloading and solve the coexistence problem of multiple services. Future work may focus on slicing resource allocation combined with cloud computing and edge computing.

References

[1] S. Jošilo and G. Dán, "Joint Wireless and Edge Computing Resource Management with Dynamic Network Slice Selection," *IEEE/ACM Transactions on Networking*, vol.30, no.4, pp.1865-1878, Aug. 2022. Article (CrossRef Link)

[2] H. Peng, E. Fitzgerald, W. Tärneberg and M. Kihl, "5G Radio Access Network Slicing in Massive MIMO Systems for Industrial Applications," in *Proc. of 2020 Seventh International Conference on Software Defined Systems (SDS)*, Paris, France, pp. 262-267, 2020. Article (CrossRef Link)

[3] A. Cárdenas and D. Fernández, "Network Slice Lifecycle Management Model for NFV-based 5G Virtual Mobile Network Operators," in *Proc. of 2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, Leganes, Spain, pp. 120-125, 2020. Article (CrossRef Link)

[4] F. Song, J. Li, C. Ma, Y. Zhang, L. Shi and D. N. K. Jayakody, "Dynamic Virtual Resource Allocation for 5G and Beyond Network Slicing," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 215-226, 2020. Article (CrossRef Link)

[5] Y. Ai, G. Qiu, C. Liu and Y. Sun, "Joint resource allocation and admission control in sliced fog radio access networks," *China Communications*, vol. 17, no. 8, pp. 14-30, Aug. 2020. Article (CrossRef Link)

[6] J. Lee and W. Na, "A Survey on Mobile Edge Computing Architectures for Deep Learning Models," in *Proc. of 2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 2346-2348, 2022. Article (CrossRef Link)

[7] X. Hu, L. Wang, K. -K. Wong, M. Tao, Y. Zhang and Z. Zheng, "Edge and Central Cloud Computing: A Perfect Pairing for High Energy Efficiency and Low-Latency," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1070-1083, Feb. 2020. Article (CrossRef Link)

[8] C. Ren, G. Zhang, X. Gu and Y. Li, "Computing Offloading in Vehicular Edge Computing Networks: Full or Partial Offloading?," in *Proc. of 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, China, pp. 693-698, 2022. Article (CrossRef Link)

[9] X. Tang, Z. Wen, J. Chen, Y. Li and W. Li, "Joint Optimization Task Offloading Strategy for Mobile Edge Computing," in *Proc. of 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, Chongqing, China, pp. 515-518, 2021. Article (CrossRef Link)

[10] N. O. Parchin, H. J. Basherlou, I. A. Yasir Al-Yasir, M. Sajedin, J. Rodriguez and R. A. Abd-Alhameed, "Multi-Mode Smartphone Antenna Array for 5G Massive MIMO Applications," in *Proc. of 2020 14th European Conference on Antennas and Propagation (EuCAP)*, Copenhagen, Denmark, pp. 1-4, 2020. Article (CrossRef Link)

[11] F. Guo, H. Zhang, H. Ji, X. Li and V. C. M. Leung, "An Efficient Computation Offloading Management Scheme in the Densely Deployed Small Cell Networks With Mobile Edge Computing," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2651-2664, Dec. 2018. Article (CrossRef Link)

[12] L. Nadeem, Y. Amin, J. Loo, M. A. Azam and K. K. Chai, "Efficient Resource Allocation Using Distributed Edge Computing in D2D Based 5G-HCN With Network Slicing," *IEEE Access*, vol. 9, pp. 134148-134162, 2021. Article (CrossRef Link)

[13] S. Chouhan, "Energy Optimal Partial Computation Offloading Framework for Mobile Devices in Multi-access Edge Computing," in *Proc. of 2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Split, Croatia, pp. 1-6, 2019. Article (CrossRef Link)

[14] Q. -V. Pham, L. B. Le, S. -H. Chung and W. -J. Hwang, "Mobile Edge Computing With Wireless Backhaul: Joint Task Offloading and Resource Allocation," *IEEE Access*, vol. 7, pp. 16444-16459, 2019. Article (CrossRef Link)

[15] R. Malik and M. Vu, "Energy-Efficient Computation Offloading in Delay-Constrained Massive MIMO Enabled Edge Network Using Data Partitioning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6977-6991, Oct. 2020. Article (CrossRef Link)

[16] K. Wang, Y. Zhou, J. Li, L. Shi, W. Chen and L. Hanzo, "Energy-Efficient Task Offloading in Massive MIMO-Aided Multi-Pair Fog-Computing Networks," *IEEE Transactions on Communications*, vol. 69, no. 4, pp. 2123-2137, April 2021. Article (CrossRef Link)

[17] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004. [Online]. Available: https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf

[18] T. T. Nguyen, L. B. Le and Q. Le-Trung, "Computation Offloading in MIMO Based Mobile Edge Computing Systems Under Perfect and Imperfect CSI Estimation," *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 2011-2025, 1 Nov.-Dec. 2021. Article (CrossRef Link)

[19] C. Ding, J. -B. Wang, H. Zhang, M. Lin and J. Wang, "Joint MU-MIMO Precoding and Resource Allocation for Mobile-Edge Computing," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1639-1654, March 2021. Article (CrossRef Link)

[20] S. Zarandi and H. Tabassum, "Delay Minimization in Sliced Multi-Cell Mobile Edge Computing (MEC) Systems," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1964-1968, June 2021. Article (CrossRef Link)

[21] Z. Tong, T. Zhang, Y. Zhu and R. Huang, "Communication and Computation Resource Allocation for End-to-End Slicing in Mobile Networks," in *Proc. of 2020 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1286-1291, 2020. Article (CrossRef Link)

[22] Y. K. Tun, M. Alsenwi, S. R. Pandey, C. W. Zaw and C. S. Hong, "Energy Efficient Multi-Tenant Resource Slicing in Virtualized Multi-Access Edge Computing," in *Proc. of 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, Matsue, Japan, pp. 1-4, 2019. Article (CrossRef Link)

[23] B. Xiang, J. Elias, F. Martignon and E. Di Nitto, "Joint Network Slicing and Mobile Edge Computing in 5G Networks," in *Proc. of ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, pp. 1-7, 2019. Article (CrossRef Link)

[24] Y. K. Tun, D. H. Kim, M. Alsenwi, N. H. Tran, Z. Han and C. S. Hong, "Energy Efficient Communication and Computation Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond," *IEEE Access*, vol. 8, pp. 136024-136035, 2020. Article (CrossRef Link)

[25] Y. Ren, A. Guo, C. Song and Y. Xing, "Dynamic Resource Allocation Scheme and Deep Deterministic Policy Gradient-Based Mobile Edge Computing Slices System," *IEEE Access*, vol. 9, pp. 86062-86073, 2021. Article (CrossRef Link)

[26] Y. He, J. Ren, G. Yu and Y. Cai, "D2D Communications Meet Mobile Edge Computing for Enhanced Computation Capacity in Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1750-1763, March 2019. Article (CrossRef Link)

[27] G. Sun, K. Xiong, G.O. Boateng, G. Liu, and W. Jiang, "Resource slicing and customization in RAN with dueling deep Q-network," *J.Netw. Comput. Appl*., vol.157, 102573, May 2020. Article (CrossRef Link)

[28] L. Li and J. Hu, "An Efficient Linear Detection Scheme Based on L-BFGS Method for Massive MIMO Systems," *IEEE Communications Letters*, vol. 26, no. 1, pp. 138-142, Jan. 2022. Article (CrossRef Link)
[29] Y. Hao, Q. Ni, H. Li and S. Hou, "Energy-Efficient Multi-User Mobile-Edge Computation Offloading in Massive MIMO Enabled HetNets," in *Proc. of ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, pp. 1-6, 2019. Article (CrossRef Link)
[30] G. Zheng, C. Xu, H. Long and X. Zhao, "MEC in NOMA-HetNets: A Joint Task Offloading and Resource Allocation Approach," in *Proc. of 2021 IEEE Wireless Communications and Networking Conference (WCNC)*, Nanjing, China, pp. 1-6, 2021. Article (CrossRef Link)

YIN REN received the B.E. degree in communication engineering from the Northeast Forestry University of Harbin, China, in 2015, and the M.S. degree in Electronics and Communication Engineering from the Harbin Engineering University, Harbin, China, in 2017. She is currently pursuing the Ph.D. degree in Information and Communication Engineering with the University of Tongji, Shanghai, China. Her research interests include mobile communication, network slicing, mobile edge computing, and machine learning.

AIHUANG GUO received the B.E. degree in applied physics from the China University of Mining and Technology, the master's degree in circuit signal and system from the Coal Science Research Institute, and the Ph.D. degree in electronic science and technology from Xi'an Jiaotong University. He is currently a Professor with the Department of Information and Communication Engineering, Tongji University, Shanghai, China. His main research directions are broadband communication and signal processing.

CHUNLIN SONG received the Ph.D. degree in computer science and technology from the Northeast University. He is currently an associate Professor with the Department of Information and Communication Engineering, Tongji University, Shanghai, China. His main research directions are mobile communication, wireless communication, embedded system, and Internet of Vehicles.