SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2024, Vol. 67, Iss. 2, 122302:1–122302:18
https://doi.org/10.1007/s11432-023-3855-5

Energy consumption optimization for edge computing-supported cellular networks based on optimal transport theory

Xiangyu LV, Xiaohu GE*, Yi ZHONG, Qiang LI & Yong XIAO

School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

Received 28 February 2023/Revised 22 June 2023/Accepted 18 August 2023/Published online 25 January 2024

Abstract With advancements in mobile communication technologies, there have been multiple user device (UD) connections to cellular networks. Because of changes in the spatial distribution of UDs and application requirements, the traditional offloading mechanism based on the nearest distance will result in heavy loads for some base stations (BSs). Because there is a high-order relationship between the energy consumption of the BS processing tasks and the number of computing tasks, the traditional offloading mechanism will result in high energy consumption, and UD's offloading decision must be dynamically adjusted. The offloading decision of UDs is optimized based on detailed information about various parameters associated with a network from the standpoint of distribution. This information includes details regarding the spatial distribution of UDs, application requirements, and the offloading period of computing tasks. Based on optimal transport theory, an energy consumption optimization algorithm is suggested to lower the total amount of energy consumed by the offloading process of UDs' computing tasks by reasonably planning the offloading BSs of the UDs in the networks. The simulation results show that the proposed offloading mechanism based on energy consumption optimization reduces the total energy consumption of the offloading process by 28.09% when compared to the traditional offloading mechanism, and the traffic managed by each BS is balanced.

Keywords optimal transport theory, edge computing, energy consumption reduction, dynamic voltage frequency scaling

1 Introduction

With the rapid deployment of fifth-generation mobile communication technology (5G) networks and an increase in demand for smart services, many resource-intensive applications have been introduced [1], including augmented reality (AR)/virtual reality (VR) [2], interactive online games [3], real-time ultrahigh-definition (UHD) video transmission [4], panorama navigation [5], smart wearable monitoring systems [6, 7]. Many of these new applications involve latency-sensitive and computationally intensive computing tasks that exceed the capability of the user device (UD). In other words, accomplishing all necessary computing tasks within the latency constraint can be a considerable challenge for UDs due to limited computing resources and battery capacity. Offloading of the computing tasks can effectively address the aforementioned issues by migrating computationally intensive tasks from UDs to cloud servers for processing [8]. In a traditional cloud data center (CDC), a UD offloads computing tasks to remote clouds for processing in wireless networks [9, 10]. However, with the dramatic increase in the number of UDs, offloading massive amounts of data to the cloud far from the UD may cause congestion of backbone networks and increase transmission latency [11]. To compensate for the drawbacks of CDCs, mobile edge computing (MEC) is proposed as an emerging computing paradigm. In comparison to CDC, MEC employs a distributed deployment strategy that frequently consists of a large number of edge servers deployed close to the base station (BS) [12]. This deployment strategy can help to reduce transmission latency and energy consumption.

* Corresponding author (email: xhge@mail.hust.edu.cn)

The majority of the existing research has concentrated on lowering the energy consumption of UDs [4, 13–16]. Li et al. [4] achieved their goal of minimizing the total energy consumption of UDs by jointly optimizing the transmission power and computational resource allocation in a single-edge server scenario that also considered the circuit power of the UD. Li et al. [13] decreased the total energy consumption of UDs under the delay constraint by jointly optimizing the transmission power of UDs and computation offloading strategy in the scenario with multiple edge servers. To tackle the problem of mutual interference among multiple communication zones, Sardellitti et al. [14] lowered the total energy consumption of UDs by jointly optimizing the pre-coding matrix of UDs and the allocation of computational resources.

Sun et al. [15] investigated the trade-off between local computing and edge offloading. Using iterative and gradient descent methods, the sum of UDs' computing utility can be maximized efficiently. Huang et al. [16] investigated the impact of various parameters in the traditional coupled access mode and the decoupled access mode on uplink transmission energy efficiency in ultra-dense heterogeneous cellular networks. Furthermore, You et al. [17] investigated the downlink energy efficiency maximized when the transmitter only has statistical channel state information in multi-cell massive multiple-input multiple-output systems.

However, the preceding studies only considered lowering the energy consumption of the UD. With the implementation of the carbon emissions and carbon neutral goal, as well as the current situation of power shortage in China, energy savings at the edge are also critical. Therefore, it is necessary to comprehensively consider the total system energy consumption, which includes both the UD and edge sides in the cellular networks. Bi et al. [18] devised an optimization problem to reduce the total energy consumed by UDs and edge servers by jointly optimizing the offloading ratio of computing tasks, UDs CPU speed, allocated bandwidth of available channels, and transmission power of each UD in each time slot. Dai et al. [19] investigated the problem of minimizing total system energy consumption by jointly optimizing transmission power and computational resource allocation in a heterogeneous network scenario. Merluzzi et al. [20] minimized the total system energy consumption with guaranteed quality of service in a heterogeneous networks scenario using Lyapunov stochastic optimization theory. Xu et al. [21] proposed an energy-aware computational offloading scheme that discovered the optimal solution based on simple weighting and multi-criteria decision labeling. Lin et al. [22] investigated the computational offloading problem in 5G networks, considering backhaul links with time constraints to minimize the overall energy consumption using the artificial fish swarm algorithm.

Given the large number of UDs connected to cellular networks, the uneven distribution of UDs, and diverse application requirements, the traditional individually optimizing method will result in high computational complexity and communication overhead in the studies described above. Hence, the offloading decision of UDs from a distribution standpoint may be beneficial in resolving the aforementioned issues. Because the offloading decision is essentially a transportation strategy for computing tasks between the UD and the BS, and transportation cost is the optimization objective, optimal transport theory is used to study the edge offloading decision optimization. Optimal transport theory has been extensively used in economics [23], transportation [24], artificial intelligence [25, 26], and wireless networks [27–30]. Considering the impact of the battery capacity of unmanned aerial vehicles (UAVs) on flight time, Wang et al. [27] minimized the total energy by jointly optimizing the UAVs' trajectory and service areas. Mozaffari et al. [28] investigated how to reduce the average hovering time of the UAVs by reasonably allocating bandwidth and service areas among UAVs while meeting the load requirements of ground UDs. Furthermore, Mozaffari et al. [29] studied the problem of minimizing the average network delay by reasonably allocating service areas under the joint service of UAVs-supported wireless cellular networks and ground BSs. Zhang et al. [30] investigated how to reasonably divide the BSs service areas in the multiple-edge server scenario to minimize the average delay and energy consumption of UDs when offloading computing tasks to the BS.

Although previous research has used optimal transport theory to solve various resource optimization problems in computing tasks edge offloading, many potential problems remain unsolved. Firstly, most existing works focused solely on the transmission energy consumption of computing tasks, ignoring the processing energy consumption and backhaul energy consumption of computing tasks on the BS side. Secondly, existing works frequently assumed the same requirements of the UD in the service area, i.e., the same data size of the generated computing tasks, which did not correspond to the current trend of diversification of the UD's application requirements. To address the aforementioned challenges, we investigate a total energy consumption reduction solution for the UDs' computing tasks offloading process,

which includes the energy consumption of the computing tasks upload, processing, and results return stages. It should be noted that in the fully offloading cellular networks scenario, the offloading decision of the UD with discrete spatial distribution is expressed as the correspondence between the UD and the offloading BS. In this paper, we consider a scenario in which the UD with continuous spatial distribution offloads all computing tasks to the MEC server on the BS side, with the offloading decision represented by the correspondence between the location of the UD and the offloading BS. When multiple locations of offloading to the same BS are linked, it acts as the service area of the BS. In this paper, the offloading decision of the UD and the division of the service area of the BS are comparable in the following description. Therefore, the offloading decision of UD is optimally equivalent to the optimal division of the service area of the BS. The task model of periodic offloading is considered in multi-BS scenarios with the spatial distribution of UDs and diverse application requirements. The total system energy consumption is reduced by carefully planning the service area of each BS. The following are the main contributions of this paper are given as follows:

• Considering the offloading period of computing tasks as well as the spatial distribution of UDs, an energy consumption model that includes computing tasks uploading, edge computing, and result returning is developed.

• Using the properties of Wasserstein distance in optimal transport theory, the optimization problem of energy consumption in the offloading process is investigated. Furthermore, the existence and uniqueness theorems of optimal service areas of BSs are established. Also proposed is an energy consumption optimization algorithm based on optimal transport theory.

• The simulation results show that the proposed offloading mechanism based on energy consumption optimization reduces the total energy consumption of the offloading process by 28.09% when compared to the offloading mechanism based on the nearest distance.

The remainder of the paper is structured as follows. Section 2 describes the system model and problem formulation. Section 3 analyzes the energy minimization problem using optimal transport theory and proposes an iterative algorithm based on the area division formula to solve the optimal division of BSs service areas. Section 4 presents the simulation results of the proposed algorithm. Section 5 discusses the conclusion.

2 System model and problem description

2.1 System model

We consider the system model illustrated in Figure 1. Let the set of UDs in the considered area labeled as $\mathcal{D}$ be $\mathcal{U} = \{1, 2, \ldots, U\}$ where U is the number of UDs that are independently distributed in the area $\mathcal{D}$. Each UD is equipped with one antenna. The set of BSs is denoted as $\mathcal{M} = \{1, \ldots, i, \ldots, M\}$, and M is the number of BSs. The coordinate of BS_i is denoted as (X_i, Y_i) , and each BS is equipped with N antennas [31]. The MEC server is deployed near the BS to provide data processing services to nearby UDs. The tasks processing capability of the BS_i is characterized by the CPU frequency f_i of the MEC server near the BS_i, and the tasks processing capability of all BSs in the area $\mathcal{D}$ are denoted by $\mathcal{F} = \{f_i\}_{i\in\mathcal{M}}$. The spatial distribution of UDs in the area $\mathcal{D}$ is not necessarily uniform. Generally, the spatial distribution of UDs is modeled as a continuous distribution and a two-dimensional distribution $f(x, y)$ is used to describe the spatial distribution of UDs in the area $\mathcal{D}$, which also satisfies $\iint_{\mathcal{D}} f(x, y)\mathrm{d}x\mathrm{d}y = 1$. In other words, the area $\mathcal{D}_r(x, y)$ represents the circle with (x, y) as the center and $r > 0$ as the radius. $f(x, y)$ represents the ratio of the number of UDs in the area $\mathcal{D}_r(x, y)$ to the total number of UDs in the area $\mathcal{D}$. We consider a typical location in the area $\mathcal{D}$ denoted by (x_0, y_0), and u_0 represents a group of UDs in the area $\mathcal{D}_r(x_0, y_0)$. The number of UDs in the group u_0 is expressed as

$$u(x_0, y_0) = Uf(x_0, y_0). \tag{1}$$

The area $\mathcal{D}$ can be divided into M sub-areas based on the areas served by different BSs, and $\mathcal{D}_i$ represents the sub-area served by the BS_i, then $\mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_M = \mathcal{D}$. The number of UDs served by the BS_i can be expressed as

$$U_i = U \iint_{\mathcal{D}_i} f(x, y)\mathrm{d}x\mathrm{d}y, \tag{2}$$

and we have $\sum_{i=1}^{M} U_i = U$.

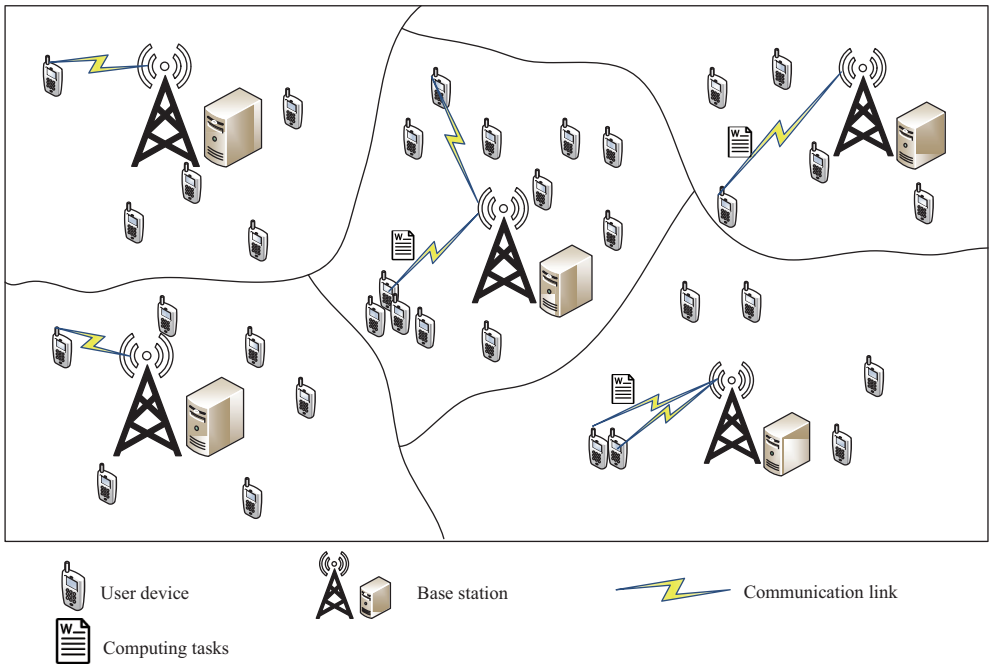

Figure 1 (Color online) System model.

2.2 Tasks model

In this paper, we assume that all computing tasks are offloaded for each UD. The computing tasks are offloaded periodically, and the computing tasks offloading period of the UD is T, which is divided into computing tasks upload time and processing time. For a more intuitive description, a time control factor τ is introduced to represent the ratio of the computing tasks upload time to the offloading period, where the computing tasks upload time and processing time at the BS side are expressed as $t_{\text{up}} = \tau T$ and $t_{\text{com}} = (1-\tau)T$, respectively. To avoid the interference between different channels, the communication link between the UD and the BS uses orthogonal frequency division multiple access (OFDMA) [32], and the bandwidth of each channel is W. The offloading process is shown in Figure 2. In computing tasks upload stage, i.e., $[nT, nT+\tau T]$, the UD uploads the computing tasks generated in $[(n-1)T, nT]$ to the assigned BS without interference. In computing tasks processing stage, i.e., $[nT+\tau T, (n+1)T]$, the MEC server on the BS side processes the offloaded tasks. And in the same time, the BS transmits the processing results of the computing tasks uploaded in $[(n-1)T, (n-1)T+\tau T]$ back to the UD without interference, which is called the processing results return stage, where $n \in Z^+$. Since the allocation mechanism divides BS service areas according to the amount of computing task data generated by the UDs during T, it is highly scalable, i.e., the time control factor τ and the offloading period T can be dynamically adjusted without exceeding the computing tasks latency requirement to meet the power limit of the BS. Specifically, when the UD offloads plenty of computing tasks, the time control factor τ can be reduced so that the BS has a longer time to process the computing tasks, which in turn reduces the energy consumption of the BS. In this paper, we assume that the BS's power can meet the requirements of the UD's computing tasks offloading.

We consider that the UD generates different types of computing tasks, and the amount of computing task data generated by different the UD may be different. For the convenience of processing, the average amount of computing task data generated by the UDs at the same location is used to represent the amount of computing task data generated by a UD at that location, i.e., $\theta(x_0, y_0)$ is used to represent the amount of computing task data generated by a UD in the group of u_0 during $[(n-1)T, nT]$, where $\theta(x_0, y_0) \in [\theta_{\min}, \theta_{\max}]$, and $\theta_{\min}$ and $\theta_{\max}$ are the minimum and maximum amount of computing task data generated by the UD, respectively. The total data amount of computing tasks generated by all UDs in the group of u_0 during $[(n-1)T, nT]$ is

$$\phi(x_0, y_0) = \theta(x_0, y_0) U f(x_0, y_0). \tag{3}$$

Since the BS_i serves the UDs in the area $\mathcal{D}_i$, the amount of computing task data π_i arriving at the

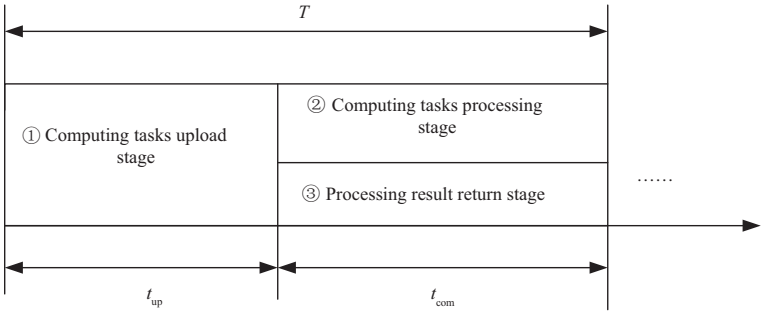

Figure 2 Computing tasks offloading process.

BS_i during $[nT, nT+\tau T]$ is

$$\pi_i = \iint_{\mathcal{D}_i} \phi(x,y)\mathrm{d}x\mathrm{d}y. \tag{4}$$

The amount of computing task data that are offloaded for all BSs in the area $\mathcal{D}$ during the period of $[nT, nT+\tau T]$ can be expressed as $\boldsymbol{\pi} = \{\pi_i\}_{i\in\mathcal{M}}$.

2.3 Channel model

In this paper, since the total energy consumption is minimized through the optimal division of the BSs service areas in the area $\mathcal{D}$, the computing tasks of the UDs at the same location will be offloaded as a whole. For the convenience of processing, the average channel gain from each UD to the BS at the same location is used to represent the channel gain from the UD to the BS at the location. Specifically, $\boldsymbol{h}_i^j(x_0,y_0)$ represents the channel gain of the jth UD in the group u_0 to the BS_i, and $\boldsymbol{h}_i(x_0,y_0)$ represents the channel gain between a UD at (x_0,y_0) and the BS_i, then we have

$$\|\boldsymbol{h}_i(x_0,y_0)\|_2^2 = \frac{1}{u(x_0,y_0)} \sum_{j=1}^{u(x_0,y_0)} \left\|\boldsymbol{h}_i^j(x_0,y_0)\right\|_2^2. \tag{5}$$

Considering that there is a Rayleigh channel between the UD and the BS [33], the receiver adopts the equal gain combining (EGC) to process the received signals, so the channel gain between the jth UD in the group u_0 and the BS_i is [34]

$$\left\|\boldsymbol{h}_i^j(x_0,y_0)\right\|_2^2 = \boldsymbol{\varphi}_{i0}^j \boldsymbol{\beta}_{i0}^j, \tag{6}$$

where $\boldsymbol{\varphi}_{i0}^j = [\varphi_{i0,1}^j, \ldots, \varphi_{i0,n}^j, \ldots, \varphi_{i0,N}^j] \in \mathbb{R}^{1\times N}$ is the small-scale fading coefficient matrix, and $\varphi_{i0,n}^j$ represents the small-scale fading coefficient from the jth UD in the group u_0 to the nth antenna of the BS_i. $\varphi_{i0,n}^j$ follows the Rayleigh distribution with parameter ϑ [35], i.e., the probability density function of $\varphi_{i0,n}^j$ is expressed as

$$f\left(\varphi_{i0,n}^j\right) = \frac{\varphi_{i0,n}^j}{\vartheta^2} \mathrm{e}^{-\frac{(\varphi_{i0,n}^j)^2}{2\vartheta^2}}, \quad \varphi_{i0,n}^j > 0. \tag{7}$$

$\boldsymbol{\beta}_{i0} = [(d_{i0,1}^j)^{-\alpha}, \ldots, (d_{i0,n}^j)^{-\alpha}, \ldots, (d_{i0,N}^j)^{-\alpha}]^{\mathrm{T}} \in \mathbb{R}^{N\times 1}$ is the large-scale fading coefficient matrix, where α is the path loss index, and $d_{i0,n}^j$ is the Euclidean distance from the jth UD in the group u_0 to the nth antenna of the BS_i. Since the distance between antennas is negligible compared to the distance between the UD and the BS, the distance between the UD and all antennas of the same BS is the same, i.e., $(d_{i0}^j)^{-\alpha} = (d_{i0,1}^j)^{-\alpha} = \cdots = (d_{i0,N}^j)^{-\alpha}$. Therefore, the channel gain between the jth UD in the group u_0 and the BS_i can be written as

$$\left\|\boldsymbol{h}_i^j(x_0,y_0)\right\|_2^2 = \boldsymbol{\varphi}_{i0}^j \boldsymbol{\beta}_{i0}^j = \left[\varphi_{i0,1}^j, \ldots, \varphi_{i0,n}^j, \ldots, \varphi_{i0,N}^j\right] \begin{bmatrix} (d_{i0}^j)^{-\alpha} \\ \vdots \\ (d_{i0}^j)^{-\alpha} \end{bmatrix}. \tag{8}$$

2.4 Problem description

In the offloading process of computing tasks from the UD to the BS, the total energy consumption of the system can be expressed as

$$E_{\text{total}} = E_{\text{total,up}} + E_{\text{total,com}} + E_{\text{total,do}}, \tag{9}$$

where $E_{\text{total,up}}$ represents the energy consumption of all UDs in the area $\mathcal{D}$ during the upload stage of computing tasks, $E_{\text{total,com}}$ represents the energy consumption of all BSs in the area $\mathcal{D}$ to process computing tasks, and $E_{\text{total,do}}$ represents the energy consumption that the processing results of computing tasks in the area $\mathcal{D}$ returned from the BS to the UD.

In the computing tasks upload stage, since the computing tasks need to be transmitted from the UD to the BS within t_{up}, the transmission rate of the jth UD in the group u_0 is

$$R_{\text{up}}^{j}\left(x_0, y_0\right) = \frac{\theta\left(x_0, y_0\right)}{\tau T}. \tag{10}$$

According to (10) and Shannon's theorem, the average upload power of the jth UD in the group u_0 offloading computing tasks to BS_i is

$$P_{i,\text{up}}^{j}\left(x_0, y_0\right) = \frac{\left(2^{\frac{\theta(x_0,y_0)}{W\tau T}} - 1\right)N_0}{\left\|\boldsymbol{h}_i\left(x_0, y_0\right)\right\|_2^2}, \tag{11}$$

where N_0 is the noise power.

According to (11), the average energy consumption of the jth UD in the group u_0 during the upload stage of computing tasks offloading process is

$$E_{i,\text{up}}^{j}(x_0, y_0) = P_{i,\text{up}}^{j}\left(x_0, y_0\right)\tau T. \tag{12}$$

The total energy consumption of all UDs in the group u_0 during the process of uploading computing tasks to the BS_i is

$$E_{i,\text{up}}\left(x_0, y_0\right) = E_{i,\text{up}}^{j}\left(x_0, y_0\right) U f\left(x_0, y_0\right). \tag{13}$$

Since the BS_i receives computing tasks offloaded by the UDs in the area $\mathcal{D}_i$, the total energy consumption of all UDs in the area $\mathcal{D}_i$ uploading computing tasks to the BS_i is

$$E_{\mathcal{D}_i,\text{up}} = \iint_{\mathcal{D}_i} E_{i,\text{up}}(x, y)\mathrm{d}x\mathrm{d}y. \tag{14}$$

According to (14), the total upload energy consumption of all UDs in the area $\mathcal{D}$ transmitting computing tasks to the BS is

$$E_{\text{total,up}} = \sum_{i=1}^{M} E_{\mathcal{D}_i,\text{up}}. \tag{15}$$

The computing tasks are processed by the MEC server on the BS side. The number of CPU cycles required to process single-bit data is denoted as ω. According to (4), when the BS_i processes computing tasks offloaded by the UDs in the area $\mathcal{D}_i$, the CPU frequency is

$$f_i = \frac{\pi_i \omega}{(1-\tau)\, T}. \tag{16}$$

In the computing tasks processing stage, the total energy consumption of the BS_i is [36, 37]

$$E_{\mathcal{D}_i,\text{com}} = \xi f_i^3\,(1-\tau)\, T, \tag{17}$$

where ξ is the effective capacitance coefficient, which depends on the chip architecture.

According to (17), the total energy consumption of all BSs in the area $\mathcal{D}$ during the computing tasks processing stage is

$$E_{\text{total,com}} = \sum_{i=1}^{M} E_{\mathcal{D}_i,\text{com}}. \tag{18}$$

The computing tasks processing results of the UD will be sent back to the UD when the new computing tasks are processed in the next period. The computing tasks uplink transmission and the processing results downlink transmission occupy the same channel. The ratio of the returned processing results data volume to the uploaded computing task data volume is denoted as δ, and the transmission rate of the processing results from the BS_i to the jth UD in the group u_0 is

$$R_{\mathrm{do}}^{j}\left(x_0, y_0\right) = \frac{\delta\theta\left(x_0, y_0\right)}{(1-\tau)T}. \tag{19}$$

According to (19) and Shannon's theorem, the average transmission power from the BS_i to the jth UD in the group u_0 is

$$P_{i,\mathrm{do}}^{j}\left(x_0, y_0\right) = \frac{\left(2^{\frac{\delta\theta(x_0,y_0)}{W(1-\tau)T}} - 1\right)N_0}{\left\|\boldsymbol{h}_i\left(x_0, y_0\right)\right\|_2^2}. \tag{20}$$

According to (20), the average energy consumption during the process of the BS_i returning results to the jth UD in the group u_0 is

$$E_{i,\mathrm{do}}^{j}\left(x_0, y_0\right) = P_{i,\mathrm{do}}^{j}\left(x_0, y_0\right)\left(1-\tau\right)T. \tag{21}$$

According to (21), when the processing results are returned from BS_i to all UDs in the group u_0, the total energy consumption is

$$E_{i,\mathrm{do}}\left(x_0, y_0\right) = E_{i,\mathrm{do}}^{j}\left(x_0, y_0\right) U f\left(x_0, y_0\right). \tag{22}$$

The processing results need to be sent back from the BS_i to all UDs in the area $\mathcal{D}_i$. According to (22), the total energy consumption of the BS_i for the processing results transmission is

$$E_{\mathcal{D}_i,\mathrm{do}} = \iint_{\mathcal{D}_i} E_{i,\mathrm{do}}(x,y)\mathrm{d}x\mathrm{d}y. \tag{23}$$

According to (23), the total energy consumption of all BSs in the area $\mathcal{D}$ for the processing results transmission is

$$E_{\mathrm{total,do}} = \sum_{i=1}^{M} E_{\mathcal{D}_i,\mathrm{do}}. \tag{24}$$

According to (9), (15), (18), and (24), the total energy consumption minimization problem of the system during the offloading process of the UDs in the area $\mathcal{D}$ can be expressed as

$$\begin{aligned}
\text{(P1)}\ \min_{\mathcal{D}_i}\ & \sum_{i=1}^{M}\left(\iint_{\mathcal{D}_i}\left(\frac{(2^{\frac{\theta(x,y)}{W\tau T}}-1)N_0\tau T}{\|\boldsymbol{h}_i(x,y)\|_2^2} + \frac{(2^{\frac{\delta\theta(x,y)}{W(1-\tau)T}}-1)N_0(1-\tau)T}{\|\boldsymbol{h}_i(x,y)\|_2^2}\right) U f(x,y)\mathrm{d}x\mathrm{d}y\right. \\
& \left. + \xi\frac{U^3\omega^3\left(\iint_{\mathcal{D}_i}\theta(x,y)f(x,y)\mathrm{d}x\mathrm{d}y\right)^3}{\left((1-\tau)T\right)^2}\right), \qquad (25)\\
\text{s.t.}\ & \mathcal{D}_p \cap \mathcal{D}_q = \emptyset, \quad \forall p \neq q \in \mathcal{M}, \qquad (25a)\\
& \bigcup_{i\in\mathcal{M}} \mathcal{D}_i = \mathcal{D}. \qquad (25b)
\end{aligned}$$

Constraint (25a) ensures that the divided areas do not overlap, and ensures that each UD selects a BS for offloading computing tasks. Constraint (25b) ensures that UDs in the area can be fully served. It is difficult to solve the above problem. Firstly, the optimization variable $\mathcal{D}_i, \forall i \in \mathcal{M}$ are interdependent continuous variable. Secondly, in order to more accurately fit the spatial distribution of UDs, $f(x, y)$ is considered to be a generic function of x and y. The above factors increase the complexity of the given double Integral. Optimal transport theory is adopted to solve the above problem, because it is suitable for the problem of mapping cost between two distributions.

3 Energy optimization modeling

3.1 Optimal transport theory

The optimal transport problem was first proposed by the French mathematician Gaspard Monge [38] in 1781. The main research is how to move the sand from the sand pile to the bunker with the minimum cost when the volume of the sand pile and the bunker are the same. From a mathematical point of view, the optimal transport problem can be abstracted as the minimum cost required to transform one probability distribution into another probability distribution. The above two probability distributions can be discrete or continuous. The optimal transport problem transformed from a continuous distribution to a discrete distribution is called a semi-discrete optimal transport problem.

The Monge problem can be expressed in mathematical language as follows. Given two separable and complete metric spaces $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$ on the polish space, $\mathbb{R}^n$ represents the n-dimensional space. Considering the probability distribution $f_1 \in \mathcal{P}(X)$ and $f_2 \in \mathcal{P}(Y)$, $\mathcal{P}(X)$ represents the space composed of all probability measures on X, and $\mathcal{P}(Y)$ represents the space composed of all probability measures on Y. For any measurable set $A \subset Y$, it can find a transport mapping F from X to Y that minimizes the total cost of the transformation while transforming the random variable $\boldsymbol{x}$ into the random variable $\boldsymbol{y}$. The mathematical expression is

$$\min_{F} \int_X c\left(\boldsymbol{x}, F\left(\boldsymbol{x}\right)\right) f_1\left(\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x}, \tag{26}$$

$$\text{s.t.} \int_A f_2\left(\boldsymbol{y}\right) \mathrm{d}\boldsymbol{y} = \int_{F^{-1}(A)} f_1\left(\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x}, \ \forall A \subset Y, \tag{26a}$$

where $c\left(\boldsymbol{x}, F\left(\boldsymbol{x}\right)\right)$ represents the unit cost of transforming from the source $\boldsymbol{x}$ to the corresponding destination $\boldsymbol{y} = F\left(\boldsymbol{x}\right)$, $F^{-1}(A) = \{\boldsymbol{x} | \boldsymbol{x} \in X, F(\boldsymbol{x}) \in A\}$. Constraints (26a) are usually abbreviated as $F_{\#} f_1 = f_2$, and $F_{\#}$ is called forward measure [39].

In order to better describe the size of the transformation between two probability distributions, the Wasserstein distance is introduced mathematically to measure, and the p-order Wasserstein distance of the Monge problem is defined as

$$M_p^p\left(f_1, f_2\right) = \min_{F} \left(\int_X \left| \boldsymbol{x} - F\left(\boldsymbol{x}\right) \right|^p f_1\left(\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} \right). \tag{27}$$

Solving the Monge problem is challenging. Firstly, the Monge problem is highly nonlinear [40]. Secondly, the Monge problem requires that each point in the source distribution only maps to one location in the destination distribution, so the optimal mapping does not necessarily exist. Based on the research of the Soviet mathematician Kantorovich, the mapping scheme is relaxed into a transport scheme so that each point in the source distribution can be transformed into multiple points in the destination distribution. The relaxed Monge problem is called the Kantorovich problem. In the following, we describe the Kantorovich problem with mathematical language. Given two separable and complete metric spaces $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$ on the polish space, $\mathbb{R}^n$ represents the n-dimensional space. Considering the probability distribution $f_1 \in \mathcal{P}(X)$ and $f_2 \in \mathcal{P}(Y)$, $\mathcal{P}(X)$ represents the space composed of all probability measures on X, and $\mathcal{P}(Y)$ represents the space composed of all probability measures on Y. All transport schemes between the two probability distributions, that is, the joint probability distribution, can be uniformly expressed as $\pi\left(\boldsymbol{x}, \boldsymbol{y}\right)$. The optimal transport scheme $\pi^*\left(\boldsymbol{x}, \boldsymbol{y}\right)$ is found to minimize the total cost of transforming from the source distribution f_1 to the destination distribution f_2. The mathematical expression is

$$\min_{\pi} \int_{X \times Y} c\left(\boldsymbol{x}, \boldsymbol{y}\right) \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}, \tag{28}$$

$$\text{s.t.} \int_Y \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} = f_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \ \int_X \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} = f_2(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}, \tag{28a}$$

where $c\left(\boldsymbol{x}, \boldsymbol{y}\right)$ represents the unit cost of transforming from the source $\boldsymbol{x}$ to the corresponding destination $\boldsymbol{y}$. Note that the transport scheme is always non-empty. The p-order Wasserstein distance that defines the Kantorovich problem is

$$W_p^p\left(f_1, f_2\right) = \min_{\pi} \left(\int_{X \times Y} \left| \boldsymbol{x} - \boldsymbol{y} \right|^p \pi\left(\boldsymbol{x}, \boldsymbol{y}\right) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \right). \tag{29}$$

It is worth noting that when the source distribution function and the cost function are continuous, the Monge problem and the Kantorovich problem are equivalent [41]

$$M_p^p\left(f_1, f_2\right) = W_p^p\left(f_1, f_2\right). \tag{30}$$

3.2 Energy consumption optimal solution

In this paper, the total energy consumption of the system is minimized by optimizing the service areas between BSs. The spatial distribution of the UDs in the area $\mathcal{D}$ is modeled as a continuous distribution, and the spatial distribution of the BSs is discrete. The process by which the UD offloads computing tasks to the BS can be modeled as a semi-discrete optimal transport process. The total energy consumption of the computing tasks offloading process can be regarded as the total cost during transport. The spatial distribution function of the UDs is $f(x, y)$, and the spatial distribution function of the BSs can be expressed as

$$\Theta = \sum_{i=1}^{M} \kappa_i \delta_i, \tag{31}$$

where κ_i represents the normalized computing tasks processing capability of the BS_i, that is, the ratio of computing task data volume processed by the BS_i to the total computing task data volume in the area $\mathcal{D}$, and $\sum_{i=1}^{M} \kappa_i = 1$, δ_i is a Dirac function.

We define $\mathcal{K} \triangleq \{(\kappa_1, \kappa_2, \ldots, \kappa_M) | \sum_{i=1}^{M} \kappa_i = 1, \kappa_i \geqslant 0, \forall i \in \mathcal{M}\}$ and the function $\Lambda : \mathcal{K} \to \mathbb{R}$, then the Wasserstein distance between different values of $(\kappa_i)_{i\in\mathcal{M}}$ and p-order can be expressed as

$$\Lambda\left(\kappa_1, \kappa_2, \ldots, \kappa_M\right) = W_p^p\left(f_1, \sum_{i=1}^{M} \kappa_i \delta_i\right). \tag{32}$$

Lemma 1. When $p \geqslant 1$, $\Lambda(\kappa_1, \kappa_2, \ldots, \kappa_M) = W_p^p(f_1, \sum_{i=1}^{M} \kappa_i \delta_i)$ is a continuous convex function [42].

Next, we prove that the optimal BSs service areas exist for problem (P1) and the optimal solution is unique.

The transformation of problem (P1) is given as follows:

$$\begin{aligned}
(\mathrm{P2}) \quad & \min_{\mathcal{D}_i} \sum_{i=1}^{M} \left(\iint_{\mathcal{D}_i} L_i(x, y) O(x, y) \mathrm{d}x\mathrm{d}y + \xi \frac{U^3 \omega^3 {s_i}^3}{\left((1-\tau)T\right)^2} \right) && (33)\\
& \text{s.t. } \mathcal{D}_p \cap \mathcal{D}_q = \emptyset, \ \forall p \neq q \in \mathcal{M}, && (33a)\\
& \quad \bigcup_{i\in\mathcal{M}} \mathcal{D}_i = \mathcal{D}, && (33b)\\
& \quad L_i(x, y) = \frac{(2^{\frac{\theta(x,y)}{W\tau T}} - 1) N_0 \tau T U}{\|\boldsymbol{h}_i(x, y)\|_2^2 \, \theta(x, y)} + \frac{(2^{\frac{\delta\theta(x,y)}{W(1-\tau)T}} - 1) N_0 (1-\tau) T U}{\|\boldsymbol{h}_i(x, y)\|_2^2 \, \theta(x, y)}, && (33c)\\
& \quad O(x, y) = \theta(x, y) f(x, y), && (33d)\\
& \quad s_i = \iint_{\mathcal{D}_i} O(x, y) \mathrm{d}x\mathrm{d}y. && (33e)
\end{aligned}$$

Lemma 2. Considering two probability measures f and Θ on the polish space, f is a continuous probability measure, and $\Theta = \sum_{i=1}^{M} \kappa_i \delta_i$ is an discrete probability measure, where δ_i is a Dirac function and κ_i is the probability. For Wasserstein distance with transport cost $p \geqslant 1$, there is an optimal mapping scheme from $f \to \Theta$, and this scheme is unique [43–45].

Next, the correspondence between the optimal BSs service areas and the normalized computing tasks processing capacity of the BS is given according to Lemma 2.

Proposition 1. When the spatial distribution of UDs follows the continuous distribution $f(x, y)$ and the spatial distribution of BSs follows the discrete distribution Θ, the computing tasks offloading process from the UD to the BS is equivalent to establishing a mapping relationship between the UD and the BS at different locations. For the Wasserstein distance with the transport cost $p \geqslant 1$, the area division $(\mathcal{D}_i)_{i\in\mathcal{M}}$ is in one-to-one correspondence with the normalized computing tasks processing capacity $(\kappa_i)_{i\in\mathcal{M}}$ of the BSs.

Proof. The mapping from the UD to the BS represents the destination where the UD offloads computing tasks. Therefore, from the perspective of the entire area, the mapping represents area division, that is, each mapping F corresponds with an area $(\mathcal{D}_i)_{i\in\mathcal{M}}$. Let $\boldsymbol{v}=(x,y)$ and the mapping from the UD to the BS can be represented by a transport graph as follows:

$$F(\boldsymbol{v})=\sum_{i=1}^{M} i\mathbf{1}_i(\boldsymbol{v}), \tag{34}$$

$$\iint_{\mathcal{D}_i} f(x,y)\mathrm{d}x\mathrm{d}y=\kappa_i, \tag{35}$$

where $\mathbf{1}_i(\cdot)$ is an indicator function. When $\boldsymbol{v}\in\mathcal{D}_i$, the value of $\mathbf{1}_i(\boldsymbol{v})$ is 1, $F(\boldsymbol{v})=i$, indicating that the UDs located at (x,y) offload computing tasks to the BS_i. When $\boldsymbol{v}\notin\mathcal{D}_i$, the value of $\mathbf{1}_i(\boldsymbol{v})$ is 0, indicating that the UDs at (x,y) do not offload the computing tasks to BS_i.

When the partitions $(\mathcal{D}_i)_{i\in\mathcal{M}}$ of the area $\mathcal{D}$ satisfy the mapping F described in (34), the integral over the partitioned area satisfy (35). According to Lemma 2, when the transport cost is the Wasserstein distance of $p\geqslant 1$, there is an optimal and unique mapping from $f\to\Theta$, and each mapping corresponds to an optimal area division. Therefore, for the Wasserstein distance with the transport cost $p\geqslant 1$, the optimal area division $(\mathcal{D}_i)_{i\in\mathcal{M}}$ is in one-to-one correspondence with the normalized computing tasks processing capacity $(\kappa_i)_{i\in\mathcal{M}}$ of the BSs. This proves Proposition 1.

Next, the existence and uniqueness theorems for the optimal division of the BSs service areas will be given based on Lemma 1 and Proposition 1.

Theorem 1. For problem (P2), when $\alpha\geqslant 1$, in the offloading process of computing tasks from UDs with continuous spatial distribution $f(x,y)$ to BSs with discrete spatial distribution Θ, there is an optimal division of BSs service areas and a unique scheme.

Proof. See Appendix A.

Lemma 3. If $f(x,y)$ is continuous over a bounded closed area $\mathcal{D}$ and the function $g(x,y)$ is integrable over $\mathcal{D}$ with invariant sign, then there exists a point $(\xi,\eta)\in\mathcal{D}$ such that the following holds [46]:

$$\iint_{\mathcal{D}} f(x,y)g(x,y)\mathrm{d}x\mathrm{d}y=f(\xi,\eta)\iint_{\mathcal{D}} g(x,y)\mathrm{d}x\mathrm{d}y. \tag{36}$$

Next, according to Theorem 1 and Lemma 3, the optimal division of BSs service areas theorem corresponding to the minimization of the total system energy consumption will be given.

Theorem 2. For the problem (P2), in the area $\mathcal{D}$, the total energy consumption of the system minimized in the offloading process of computing tasks from UDs with continuous spatial distribution $f(x,y)$ to BSs with discrete spatial distribution $\Theta=\sum_{i=1}^{M}\kappa_i\delta_i$ is

$$E_{\mathrm{OT}}=\sum_{i=1}^{M}\left(\iint_{\mathcal{D}_i^*} L_i(x,y)O(x,y)\mathrm{d}x\mathrm{d}y+\xi\frac{U^3\omega^3{s_i^*}^3}{\left((1-\tau)T\right)^2}\right), \tag{37}$$

where $\mathcal{D}_i^*$ is the area served by BS_i after the optimal division, given by

$$\mathcal{D}_i^*=\left\{(x,y)|L_i(x,y)+\xi\frac{3U^3\omega^3{s_i}^2}{\left((1-\tau)T\right)^2}<L_j(x,y)+\xi\frac{3U^3\omega^3{s_j}^2}{\left((1-\tau)T\right)^2},\forall i\neq j\in\mathcal{M}\right\}. \tag{38}$$

Proof. See Appendix B.

It can be seen from (38) that for $\forall i\in\mathcal{M}$, there is interdependence between s_i and $\mathcal{D}_i$, so there is no clear form to solve (38). Therefore, according to [42], we propose an iterative algorithm that can converge to the optimal solution within a limited number of iterations. By solving (38), the optimal BSs service areas division and the minimum value of total system energy consumption can be found. The energy consumption optimization algorithm of cellular networks based on optimal transport theory is shown in Algorithm 1.

The core idea of Algorithm 1 is the cautious self-interest of the UD, that is, focusing on minimizing the overall system cost. For the computing task offloading of the UD studied in this paper, we focus on minimizing the system total energy consumption. The first step is initialization, and the UD in different

locations randomly selects a BS to offload. The UD selects an offloading mechanism based on the nearest distance to initialize in Algorithm 1. Due to the influence of the uneven spatial distribution of UDs and the uneven spatial distribution of computing tasks, the load of some BSs is too large. Since the high order relationship between the energy consumption of processing computing tasks and the number of computing tasks, those BSs consume too much energy. Therefore, according to (38), the UD will choose the BS that reduces the cost to offload. However, this may lead to a result that the lighter load BS becomes the heavier load BS due to the offloading of tasks by a large number of UDs, and the original heavier load BS becomes the lighter load BS due to the departure of a large number of UDs. In order to avoid the UD choosing back and forth between BSs due to self-interest, the caution parameter $\Phi(x,y)$ is introduced, which is the caution of the UD to change in the next selection. In the third step, for each BS in the area $\mathcal{D}$, when the UD is within the service range of the BS, the UD's willingness to leave the BS are $\Phi_i^{(z+1)}(x,y)=(1-\frac{1}{z})\Phi_i^{(z)}(x,y)$. On the contrary, when the UD is not within the service range of the BS, the UD's willingness that does not uninstall it to the BS is $\Phi_i^{(z+1)}(x,y)=1-(1-\frac{1}{z})(1-\Phi_i^{(z)}(x,y))$. In the fourth step, according to the caution parameter, the amount of offloaded task is solved one by one for the BS in the area $\mathcal{D}$. In the sixth step, the service area of each BS is updated by using (38).

Algorithm 1 The energy consumption optimization algorithm of cellular networks based on optimal transport theory

Input: $f(x,y)$, U, T, τ, $\theta(x,y)$, ω, (X_i,Y_i), $\forall i\in\mathcal{M}$, α, ϑ, W, ξ, Z.
Output: $\mathcal{D}_i^*,\forall i\in\mathcal{M}$, E_{OT}.
1: $z=1$, generate an initial cell partitions $\mathcal{D}_i^{(z)}$, and set $\Phi_i^{(z)}(x,y)=0,\forall i\in\mathcal{M}$.
2: **while** $z<Z$ **do**
3: Compute $\Phi_i^{(z+1)}(x,y)=\begin{cases}\left(1-\frac{1}{z}\right)\Phi_i^{(z)}(x,y), & \text{if } (x,y)\in\mathcal{D}_i^{(z)},\\ 1-\left(1-\frac{1}{z}\right)(1-\Phi_i^{(z)}(x,y)), & \text{o.w.;}\end{cases}$
4: Compute $s_i=\int_{\mathcal{D}}(1-\Phi_i^{(z+1)}(x,y))O(x,y)\mathrm{d}x\mathrm{d}y$, $\forall i\in\mathcal{M}$;
5: $z\to z+1$;
6: Update cell partitions using (38);
7: **end while**
8: $\mathcal{D}_i^*=\mathcal{D}_i^{(z)}$;
9: Compute E_{OT} using (37) based on $\mathcal{D}_i^*$, $\forall i\in\mathcal{M}$.

4 Simulation results

In this section, we simulate and analyze the energy consumption optimization problem of cellular networks according to optimal transport theory. Furthermore, we compare the offloading mechanism based on energy consumption optimization proposed in this paper with the offloading mechanism based on the nearest distance.

For the simulations, we consider a rectangular area of $L_x\times L_y$, where L_x and L_y are the length and width of the rectangular area, respectively. Five BSs are deployed at random in the considered area. The UDs in the area under consideration can have any two-dimensional continuous spatial distribution. Because the spatial distribution of the UDs is uneven, we chose a two-dimensional truncated Gaussian distribution with hotspot characteristics, which is expressed as follows:

$$f(x,y)=\frac{1}{\varpi}\exp\left[-\left(\frac{x-\mu_x}{\sqrt{2}\sigma_x}\right)^2\right]\exp\left[-\left(\frac{y-\mu_y}{\sqrt{2}\sigma_y}\right)^2\right], \tag{39}$$

where $\varpi=2\pi\sigma_x\sigma_y\mathrm{erf}(\frac{L_x-\mu_x}{\sqrt{2}\sigma_x})\mathrm{erf}(\frac{L_y-\mu_y}{\sqrt{2}\sigma_y})$. μ_x and μ_y represents the mean value of the x-coordinate and y-coordinate respectively, and σ_x and σ_y represent the standard deviation of the x-coordinate and y-coordinates, respectively, and $\mathrm{erf}(s)=\frac{2}{\sqrt{\pi}}\int_0^s \mathrm{e}^{-t^2}\mathrm{d}t$. According to the properties of the two-dimensional truncated Gaussian distribution, the hot spot coordinates are (μ_x,μ_y), and σ_x and σ_y are the reciprocals of the UDs density around the hotspot in the x-coordinate and y-coordinates, respectively. We define the density of UDs in the x-coordinate and y-coordinate surrounding the hotspot as $\rho_x=\frac{1}{\sigma_x}$ and $\rho_y=\frac{1}{\sigma_y}$ respectively, and set $\rho_x=\rho_y$ in the subsequent simulation. Unless stated otherwise, other simulation parameters are shown in Table 1 [18, 35, 47–49].

Table 1 Simulation parameters

Parameter	Description	Value
L_x, L_y	Rectangular area	1000 m, 1000 m
μ_x, μ_y	Mean of a 2-D truncated Gaussian distribution	430 m, 450 m
σ_x, σ_y	Standard deviation of a 2-D truncated Gaussian distribution	300 m, 300 m
$\{(X_i, Y_i)\}_{i\in\mathcal{M}}$	BS location coordinates	(200, 200) (200, 800) (500, 500) (800, 200) (800, 800)
$\theta_{\min}$	The lower limit of the computing task data volume	10 kbit
$\theta_{\max}$	The upper limit of the computing task data volume	20 kbit
N_0 [18]	Noise power density	10^{-11} W
τ	Time control factor	0.3
W	Bandwidth	10 kHz
ϑ [35]	Rayleigh distribution standard deviation	0.5
ξ [47]	Effective capacitance constant	10^{-26}
ω [48]	Number of CPU cycles to process a single bit	800 cycle/bit
T [49]	Offloading period	20 ms
U	Number of UDs	8000

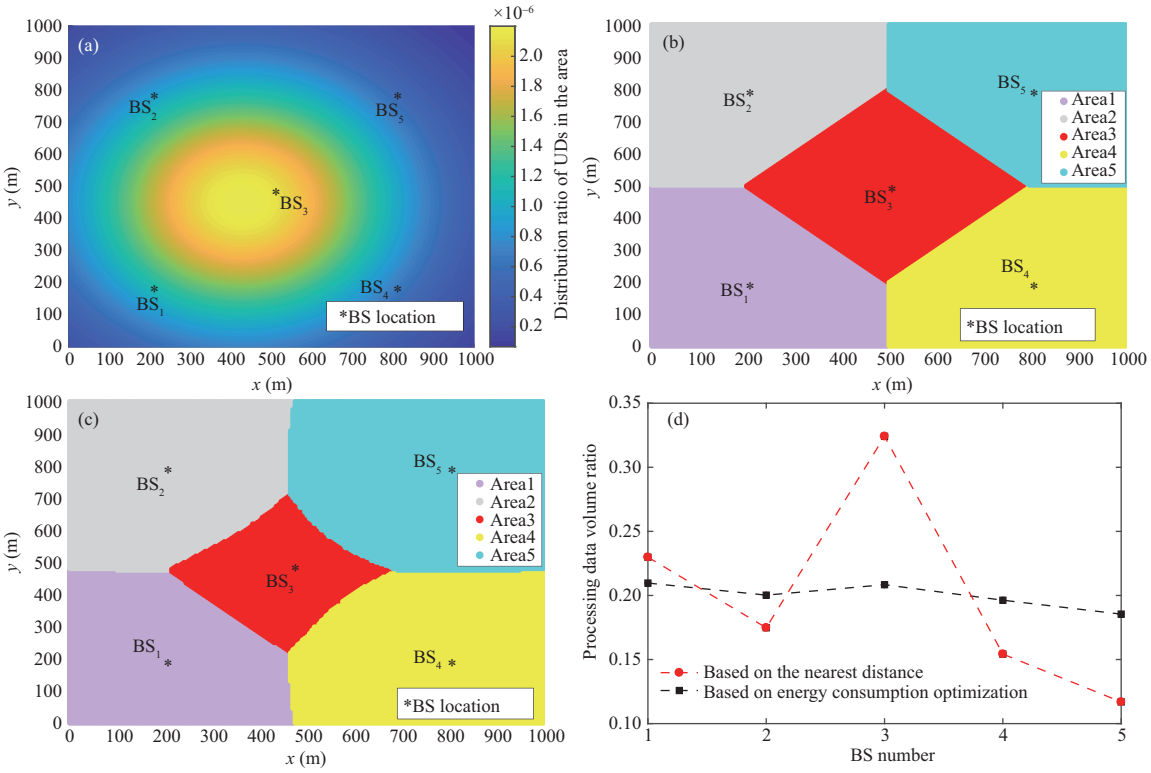

Figure 3 (a) Spatial distribution of the UDs; (b) area division of the offloading mechanism based on the nearest distance; (c) area division of the offloading mechanism based on energy consumption optimization; (d) processing data volume ratio as a function of the BS number.

Figure 3 depicts the spatial distribution of the UDs, as well as the area division and BS load, based on various offloading mechanisms. Figure 3(a) illustrates the spatial distribution of the UDs with hotspot characteristics. The density of the UDs is highest near the hotspot (430, 450) and gradually decreases in the radial direction outward. Figure 3(d) illustrates the ratio of computing task data processed by different BSs to the total amount of computing task data generated by the UDs in the area $\mathcal{D}$ under different offloading mechanisms. When the nearest distance offloading mechanism is used, because of the uneven spatial distribution of the UDs and the diversity of computing tasks generated by the UDs, the amount of computing task data processed by the BS_1 and BS_3 near the hot spot is significantly higher than that processed by the BS_5 far away from the hot spot. The BS_1 and BS_3 process 1.96 and 2.77 times as much computing task data as the BS_5, respectively. When the offloading mechanism based

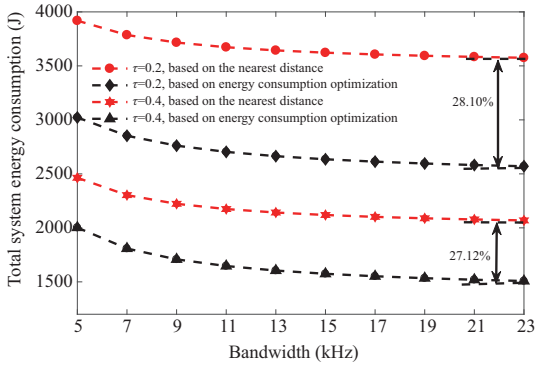

Figure 4 (Color online) Total system energy consumption as a function of bandwidth.

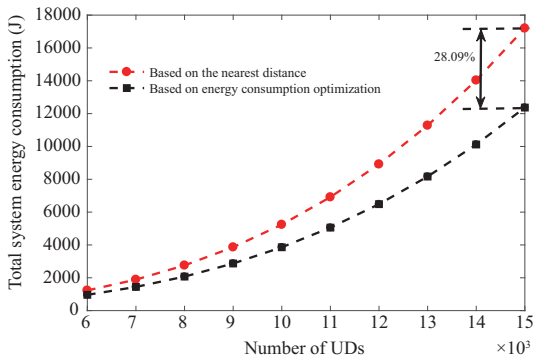

Figure 5 (Color online) Total system energy consumption as a function of the number of UDs.

on energy consumption optimization is used, the amount of computing task data processed by each BS is more balanced. Figures 3(b) and (c) depict the area division obtained by the offloading mechanism based on the nearest distance and the area division obtained by the offloading mechanism based on energy consumption optimization, respectively. According to Figures 3(b) and (d), the UDs in the service area of BS_3 are dense, increasing total energy consumption in the entire area to process the computing tasks generated by the UDs rise. According to Figures 3(c) and (d), some UDs assigned to BS_1 and BS_3 in the nearest distance offloading mechanism are assigned to BS_2, BS_4, and BS_5 in the offloading mechanism based on energy consumption optimization. At this time, the UDs in the area must expend more energy to transmit computing tasks to the BS located further away. However, because of the reduction of the total energy consumption of the BS side computing task processing, the total energy consumption of the system is reduced.

Figure 4 depicts the relationship between the total energy consumption of the system and the transmission bandwidth of the UD in the case of various offloading mechanisms and time control factors. As shown in Figure 4, when the time control factor is fixed, the total energy consumption of the system decreases as the transmission bandwidth of the UD increases under all schemes. When the transmission bandwidth remains constant and the time control factor remains constant, the total energy consumption of the system obtained by the offloading mechanism based on energy consumption optimization is less than that obtained by the offloading mechanism based on the nearest distance. When the channel transmission bandwidth is set to 23 kHz and the time control factor is set to $\tau = 0.2$, the total energy consumption of the system is reduced by 28.10%. When the transmission channel bandwidth is set to 23 kHz and the time control factor is set to $\tau = 0.4$, the total energy consumption of the system is reduced by 27.12%. When the transmission channel bandwidth is fixed, the smaller the time control factor, the higher the reduction in the total energy consumption of the system.

Figure 5 depicts the relationship between the total energy consumption of the system and the number of UDs under various offloading mechanisms. As shown in Figure 5, the total energy consumption of the system increases as the number of UDs across all schemes. When the number of UDs is fixed in the entire area, the total energy consumption of the system obtained by the offloading mechanism based on energy consumption optimization is less than that obtained by the offloading mechanism based on the nearest distance. When the number of UDs is 15000 in the considered area, the total energy consumption of the system is decreased by 28.09%.

Figure 6 depicts the relationship between the total energy consumption of the system and the upper limit of computing task data volume generated by UDs under various offloading mechanisms. Because the amount of computing task data generated by UDs is randomly selected from $(\theta_{\min}, \theta_{\max})$, the total amount of computing task data generated by UDs increases as the upper limit $\theta_{\max}$ is increased. The total energy consumption of the system increases as the upper limit of computing task data volume increases under all schemes. When the upper limit of the computing task data volume is fixed, the total energy consumption of the system obtained by the offloading mechanism based on energy consumption optimization is less than that obtained by the offloading mechanism based on the nearest distance. When the upper limit of the computing task data volume $\theta_{\max}$ is set to 60 kbit, the total energy consumption of the system is reduced by 28.53%.

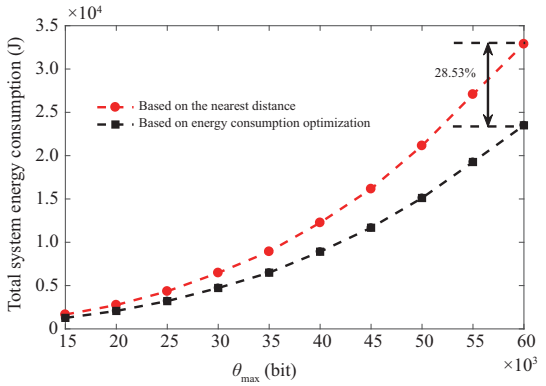

Figure 6 (Color online) Total system energy consumption as a function of the upper limit of computing task data volume.

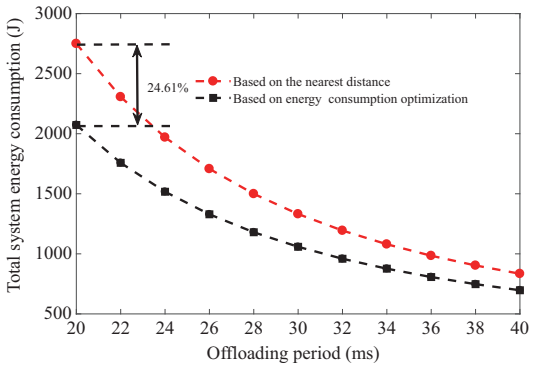

Figure 7 (Color online) Total system energy consumption as a function of the offloading period of the UDs.

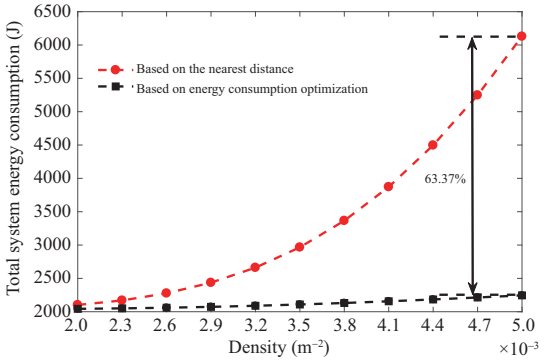

Figure 8 (Color online) Total system energy consumption as a function of the density of UDs at hotspot.

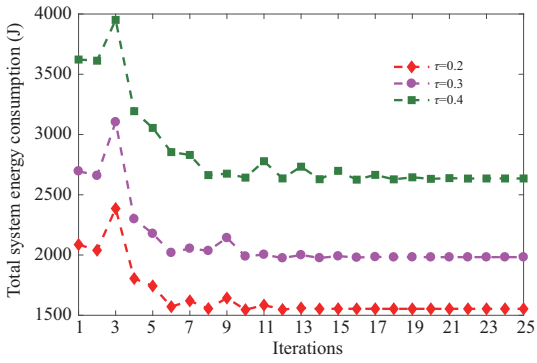

Figure 9 (Color online) Total system energy consumption as a function of the number of iterations.

Figure 7 depicts the relationship between the total energy consumption of the system and the offloading period of computing tasks of a UD under various offloading mechanisms. According to Figure 7, the total energy consumption of the system decreases as the offloading increases under all schemes. When this period is fixed, the total energy consumption of the system obtained using the offloading mechanism based on energy consumption optimization is less than that obtained by the offloading mechanism based on the nearest distance. When the offloading period is 20 ms, the total energy consumption of the system decreases by 24.61%.

Figure 8 depicts the relationship between the total energy consumption of the system and the density of UDs hotspots under various UDs offloading mechanisms. Unlike in Figure 5, where the spatial distribution of UDs is fixed and the number of UDs increases, Figure 8 investigates the system performance change when the number of UDs is fixed and the spatial distribution of UDs changes. The total energy consumption of the system increases with the density of UDs hotspots under all schemes, and the total energy consumption of the system increases faster when the nearest distance offloading mechanism is used. When the density of UDs at the hotspot is fixed, the total energy consumption of the system obtained by the offloading mechanism based on energy consumption optimization is less than that obtained by the offloading mechanism based on the nearest distance. When the density of UDs at the hotspot is 0.005, the total energy consumption of the system is reduced by 63.37%.

Figure 9 depicts the relationship between the total energy consumption of the system and the number of iterations when the time control factor is changed under the offloading mechanism based on energy consumption optimization. Figure 9 shows that when the number of iterations is fixed, the total energy consumption of the system increases as the time control factor increases. When the time control factor is fixed, the total energy consumption of the system tends to decrease as the number of iterations increases. Furthermore, once a certain number of iterations is reached, it will remain stable as the number of iterations increases. When the number of iterations is less than 15, the total energy consumption of the

system decreases as the number of iterations increases, and the difference is more noticeable. At the moment, the transmission energy consumption of the UD and the processing energy consumption of the MEC server on the BS side are competing to achieve the lowest sum in the system. When the number of iterations exceeds 15, the total energy consumption of the system remains constant as the number of iterations increases. At this point, the total energy consumption of the system is at its lowest, and the association between the UD and the BS is at its best.

5 Conclusion

In this paper, the problem of multi-BS cooperatively providing computing tasks offloading service for the UD is investigated. Considering the spatial distribution of UDs, diversity application requirements, and offloading period, the total energy consumption of the system in the process of computing tasks offloading is optimized by reasonably planning the service areas between BSs. Given that the offloading of computing tasks involves the transformation of one distribution into another, optimal transport theory is introduced. Using the properties of Wasserstein distance in optimal transport theory, the optimization problem of energy consumption in the offloading process is examined, and an energy consumption optimization algorithm for cellular networks is proposed. The simulation results show that, when compared to the offloading mechanism based on the nearest distance, the offloading mechanism based on energy consumption optimization proposed in this paper reduces the total energy consumption of the system. When the number of UDs is 15000 in the considered area, the total energy consumption of the system is reduced by 28.09%.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. U2001210).

References

1 Wang F, Xu J, Wang X, et al. Joint offloading and computing optimization in wireless powered mobile-edge computing systems. IEEE Trans Wireless Commun, 2017, 17: 1784–1797
2 Xu J Y, Wei Z C, Lv Z W, et al. Throughput maximization of offloading tasks in multi-access edge computing networks for high-speed railways. IEEE Trans Veh Technol, 2021, 70: 9525–9539
3 Mao Y Y, Zhang J, Letaief K B. Dynamic computation offloading for mobile-edge computing with energy harvesting devices. IEEE J Sel Areas Commun, 2016, 34: 3590–3605
4 Li S, Sun W, Sun Y, et al. Energy-efficient task offloading using dynamic voltage scaling in mobile edge computing. IEEE Trans Netw Sci Eng, 2020, 8: 588–598
5 Huang X M, Ye D D, Yu R, et al. Securing parked vehicle assisted fog computing with blockchain and optimal smart contract design. IEEE CAA J Autom Sin, 2020, 7: 426–441
6 Fang L W, Wang Z W, Chen Z Q, et al. 3D shape reconstruction of lumbar vertebra from two X-ray images and a CT model. IEEE CAA J Autom Sin, 2019, 7: 1124–1133
7 Wang B, Li M C, Jin X, et al. A reliable IoT edge computing trust management mechanism for smart cities. IEEE Access, 2020, 8: 46373–46399
8 Dinh H T, Lee C, Niyato D, et al. A survey of mobile cloud computing: architecture, applications, and approaches. Wireless Commun, 2013, 13: 1587–1611
9 Guo S T, Liu J D, Yang Y Y, et al. Energy-efficient dynamic computation offloading and cooperative task scheduling in mobile cloud computing. IEEE Trans Mobile Comput, 2018, 18: 319–333
10 Kuang Z K, Shi Y W, Guo S T, et al. Multi-user offloading game strategy in OFDMA mobile cloud computing system. IEEE Trans Veh Technol, 2019, 68: 12190–12201
11 Chen L X, Zhou S, Xu J. Computation peer offloading for energy-constrained mobile edge computing in small-cell networks. IEEE ACM Trans Networking, 2018, 26: 1619–1632
12 He L, Li F C, Xu H K, et al. Blockchain-based vehicular edge computing networks: the communication perspective. Sci China Inf Sci, 2023, 66: 172301
13 Li S L, Tao Y Z, Qin X Q, et al. Energy-aware mobile edge computation offloading for IoT over heterogenous networks. IEEE Access, 2019, 7: 13092–13105
14 Sardellitti S, Scutari G, Barbarossa S. Joint optimization of radio and computational resources for multicell mobile-edge computing. IEEE Trans Signal Inf Process Netw, 2015, 1: 89–103
15 Sun H, Zhou F, Hu R Q. Joint offloading and computation energy efficiency maximization in a mobile edge computing system. IEEE Trans Veh Technol, 2019, 68: 3052–3056
16 Huang T, Zheng F C, Lai L. On the local delay and energy efficiency under decoupled uplink and downlink in HetNets. Sci China Inf Sci, 2022, 65: 132304
17 You L, Huang Y F, Zhong W, et al. Robust online energy efficiency optimization for distributed multi-cell massive MIMO networks. Sci China Inf Sci, 2023, 66: 132302
18 Bi J, Yuan H T, Duanmu S F, et al. Energy-optimized partial computation offloading in mobile-edge computing with genetic simulated-annealing-based particle swarm optimization. IEEE Int Things J, 2020, 8: 3774–3785
19 Dai Y Y, Xu D, Maharjan S, et al. Joint computation offloading and user association in multi-task mobile edge computing. IEEE Trans Veh Technol, 2018, 67: 12313–12325
20 Merluzzi M, Pietro N, Di Lorenzo P, et al. Discontinuous computation offloading for energy-efficient mobile edge computing. IEEE Trans Green Commun Netw, 2021, 6: 1242–1257
21 Xu X L, Li Y C, Huang T, et al. An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks. J Network Comput Appl, 2019, 133: 75–85

22 Lin X, Wang Y Z, Xie Q, et al. Task scheduling with dynamic voltage and frequency scaling for energy minimization in the mobile cloud computing environment. IEEE Trans Serv Comput, 2014, 8: 175–186
23 Mashkin A L, Telushkina E K, Ulitskaya N M, et al. Digital technologies of public administration in transport. In: Proceedings of the Intelligent Technologies and Electronic Devices in Vehicle and Road Transport Complex (TIRVED), 2021. 1–6
24 Li S Q, Lang M X, Li S Y, et al. Optimization of high-speed railway line planning with passenger and freight transport coordination. IEEE Access, 2022, 10: 110217
25 Oh G, Sim B, Chung H J, et al. Unpaired deep learning for accelerated MRI using optimal transport driven CycleGAN. IEEE Trans Comput Imag, 2020, 6: 1285–1296
26 Akbari A, Awais M, Fatemifar S, et al. Deep order-preserving learning with adaptive optimal transport distance. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 313–328
27 Wang D, Tian J, Zhang H X, et al. Task offloading and trajectory scheduling for UAV-enabled MEC networks: an optimal transport theory perspective. IEEE Wireless Commun Lett, 2021, 11: 150–154
28 Mozaffari M, Saad W, Bennis M, et al. Wireless communication using unmanned aerial vehicles (UAVs): optimal transport theory for hover time optimization. IEEE Trans Wireless Commun, 2017, 16: 8052–8066
29 Mozaffari M, Saad W, Bennis M, et al. Optimal transport theory for cell association in UAV-enabled cellular networks. IEEE Commun Lett, 2017, 21: 2053–2056
30 Zhang Q, Jiang Y, Ge X, et al. Resource allocation based on optimal transport theory in IoT edge computing. Chinese J Int Things, 2021, 5: 60–70
31 Aredo S C, Negash Y, Marye Y W, et al. Hardware efficient massive MIMO systems with optimal antenna selection. Sensors, 2022, 22: 1743
32 Cui G, Xu Y, Zhang H, et al. Secure data offloading strategy for multi-UAV wireless networks based on minimum energy consumption. J Commun, 2021, 42: 51–62
33 Zhong Y, Mao G Q, Ge X H, et al. Spatio-temporal modeling for massive and sporadic access. IEEE J Sel Areas Commun, 2020, 39: 638–651
34 Ti N T, Le L B, Le-Trung Q. Computation offloading in MIMO based mobile edge computing systems under perfect and imperfect CSI estimation. IEEE Trans Serv Comput, 2019, 14: 2011–2025
35 Gómez-Déniz E, Gómez-Déniz L. A generalisation of the Rayleigh distribution with applications in wireless fading channels. Wireless Commun, 2013, 13: 85–94
36 Bi J, Yuan H T, Zhang K Y, et al. Energy-minimized partial computation offloading for delay-sensitive applications in heterogeneous edge networks. IEEE Trans Emerg Top Comput, 2022, 10: 1941–1954
37 Sun Y Y, Song C H, Yu S M, et al. Energy-efficient task offloading based on differential evolution in edge computing system with energy harvesting. IEEE Access, 2021, 9: 16383–16391
38 Monge G. Mémoire sur la théorie des déblais et des remblais. Mem Math Phys Acad Royale Sci, 1781, 666–704
39 Shen X. A Fast Algorithm for Monge-Kantorovich Problem Based on Partial Differential Equations/Probability Theory. Shanghai: East China Normal University, 2012
40 Villani C. Topics in Optimal Transportation. Providence: American Mathematical Society, 2003
41 Ambrosio L, Bressan A, Helbing D, et al. A user's guide to optimal transport. In: Modelling and Optimisation of Flows on Networks. Berlin: Springer, 2013. 1–155
42 Crippa G, Jimenez C, Pratelli A. Optimum and equilibrium in a transport problem with queue penalization effect. Adv Calc Var, 2009, 2: 207–246
43 Rüschendorf L. Optimal solutions of multivariate coupling problems. Appl Math (Warsaw), 1995, 23: 325–338
44 Caffarelli L, Feldman M, McCann R. Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. J Amer Math Soc, 2002, 15: 1–26
45 Ambrosio L, Pratelli A, Caffarelli L A, et al. Existence and stability results in the L^1 theory of optimal transportation. In: Optimal Transportation and Applications. Berlin: Springer, 2001. 123–160
46 Yin F, Wang P. The extension of double integral mean value theorem. J Xinzhou Teach Univ, 2011, 27: 15–16
47 Nguyen P X, Tran D H, Onireti O, et al. Backscatter-assisted data offloading in OFDMA-based wireless-powered mobile edge computing for IoT networks. IEEE Int Things J, 2021, 8: 9233–9243
48 Su C X, Ye F, Liu T T, et al. Computation offloading in hierarchical multi-access edge computing based on contract theory and Bayesian matching game. IEEE Trans Veh Technol, 2020, 69: 13686–13701
49 Malik R, Vu M. On-request wireless charging and partial computation offloading in multi-access edge computing systems. IEEE Trans Wireless Commun, 2021, 20: 6665–6679

Appendix A Proof of Theorem 1

(a) Existence. A unit simplex is defined as follows:

$$\mathcal{K}=\left\{\boldsymbol{\kappa}=(\kappa_1,\kappa_2,\ldots,\kappa_M)\in\mathbb{R}^M;\sum_{i=1}^{M}\kappa_i=1,\kappa_i\geqslant 0,\forall i\in\mathcal{M}\right\},\tag{A1}$$

where $\kappa_i=\frac{\iint_{\mathcal{D}_i}O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}}O(x,y)\mathrm{d}x\mathrm{d}y}$.

The BSs service areas division is performed when the spatial distribution of UDs and the spatial distribution of the computing tasks generated by UDs are determined. Therefore, when $f(x,y)$ and $\theta(x,y)$ are given, $\iint_{\mathcal{D}}O(x,y)\mathrm{d}x\mathrm{d}y$ is a constant value. For any given $\boldsymbol{\kappa}$, $s_i=\kappa_i\iint_{\mathcal{D}}O(x,y)\mathrm{d}x\mathrm{d}y,\forall i\in\mathcal{M}$ is determined. That is, no matter how κ_i, $\forall i\in\mathcal{M}$ changes, s_i does the same change.

According to the Wasserstein distance definition of the Kantorovich problem, the cost function represented by (33c) is the Wasserstein distance of $p=\alpha$, namely

$$W_\alpha^\alpha\left(f,\sum_{i=1}\kappa_i\delta_i\right)=\min_{\mathcal{D}_i}\sum_{i=1}^{M}\iint_{\mathcal{D}_i}L_i(x,y)O(x,y)\mathrm{d}x\mathrm{d}y.\tag{A2}$$

According to (32), (33) and (A2), the problem (P2) can be transformed into

$$(P3)\quad\min_{(\kappa_i)_{i\in\mathcal{M}}}\left\{\Lambda\left(\kappa_1,\kappa_2,\ldots,\kappa_M\right)+\sum_{i=1}^{M}\eta\left(\kappa_i\right)\right\},\tag{A3}$$

$$\text{s.t. } \mathcal{D}_p \cap \mathcal{D}_q = \emptyset,\ \forall p \neq q \in \mathcal{M}, \tag{A3a}$$

$$\bigcup_{i\in\mathcal{M}} \mathcal{D}_i = \mathcal{D}, \tag{A3b}$$

where $\eta(\kappa) = \xi \frac{\kappa^3 U^3 \omega^3 (\iint_{\mathcal{D}} O(x,y)\mathrm{d}x\mathrm{d}y)^3}{((1-\tau)T)^2}$.

According to Lemma 1 and $\eta(\kappa)$ is a continuous function of κ, when $\alpha \geqslant 1$, the expression (A3) is a continuous function defined on the non-empty compact set $\mathcal{K}$. So there exists $(\hat{\kappa}_i)_{i\in\mathcal{M}}$ that satisfies the problem (P3). According to Proposition 1, it can be deduced that there is an optimal division scheme $(\hat{\mathcal{D}}_i)_{i\in\mathcal{M}}$ of BSs service areas, and problem (P2) can be proved. That is, there is an optimal division scheme of BSs service areas in problem (P1).

(b) Uniqueness. We assume that there are two different optimal service areas division schemes $(\mathcal{D}_{1,i})_{i\in\mathcal{M}}$ and $(\mathcal{D}_{2,i})_{i\in\mathcal{M}}$ for the BSs in problem (P2). Due to the delineation of the area, the proportion of the partition can be calculated

$$\kappa_{1,i} = \frac{\iint_{\mathcal{D}_{1,i}} O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}} O(x,y)\mathrm{d}x\mathrm{d}y},\ \forall i \in \mathcal{M}, \tag{A4}$$

$$\kappa_{2,i} = \frac{\iint_{\mathcal{D}_{2,i}} O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}} O(x,y)\mathrm{d}x\mathrm{d}y},\ \forall i \in \mathcal{M}, \tag{A5}$$

$F_1(\boldsymbol{v}) = \sum_{i=1}^{M} i\mathbf{1}_{\mathcal{D}_{1,i}}(\boldsymbol{v})$ represents the optimal transport map from UDs whose spatial distribution satisfies $f(x,y)$ to BSs whose spatial distribution satisfies $\sum_{i=1}^{M} \kappa_{1,i}\delta_i$, and $F_2(\boldsymbol{v}) = \sum_{i=1}^{M} i\mathbf{1}_{\mathcal{D}_{2,i}}(\boldsymbol{v})$ represents the optimal transport map from UDs whose spatial distribution satisfies $f(x,y)$ to BSs whose spatial distribution satisfies $\sum_{i=1}^{M} \kappa_{2,i}\delta_i$. When $p \geqslant 1$, i.e., $\alpha \geqslant 1$ in the problem (P2), according to Lemma 1 and the fact that $\eta(\kappa)$ is a continuous and strictly convex function about κ, we can obtain that $(\kappa_i^*)_{i\in\mathcal{M}}$ that minimizes the problem (P2) is unique. Since $(\kappa_{1,i})_{i\in\mathcal{M}}$ and $(\kappa_{2,i})_{i\in\mathcal{M}}$ correspond to the optimal BSs service areas division $(\mathcal{D}_{1,i})_{i\in\mathcal{M}}$ and $(\mathcal{D}_{2,i})_{i\in\mathcal{M}}$, respectively, so we have

$$\left(\kappa_i^*\right)_{i\in\mathcal{M}} = (\kappa_{1,i})_{i\in\mathcal{M}} = (\kappa_{2,i})_{i\in\mathcal{M}}. \tag{A6}$$

However, according to Lemma 2, there is a unique optimal transport map from continuous distribution to discrete distribution. Therefore, the condition for $(\kappa_{1,i})_{i\in\mathcal{M}} = (\kappa_{2,i})_{i\in\mathcal{M}}$ in (A6) is that two kinds of the BSs service areas division schemes are the same, which is contradictory to the assumption. Moreover, the optimal division scheme of BSs service areas for problem (P2) is unique. That is, there is a unique optimal BSs service area division scheme for problem (P1).

Appendix B Proof of Theorem 2

According to Theorem 1, there is an optimal BSs service areas division scheme $\mathcal{D}_i, \forall i \in \mathcal{M}$ for problem (P1). Consider two of the optimal partitions $\mathcal{D}_k$ and $\mathcal{D}_l$, and a point e_0 in the area $\mathcal{D}_k$ with coordinate (x_{k0}, y_{k0}). We let $\mathcal{D}_r(e_0)$ represent the circle domain with e_0 as the center and r as the radius, and limit the circle domain to be completely within the area $\mathcal{D}_k$, i.e., $\mathcal{D}_r(e_0) \subset \mathcal{D}_k$. According to the selected circular domain, a new cell partition is generated as follows:

$$\begin{cases} \tilde{\mathcal{D}}_k = \mathcal{D}_k \backslash \mathcal{D}_r(e_o), \\ \tilde{\mathcal{D}}_l = \mathcal{D}_l \cup \mathcal{D}_r(e_o), \\ \tilde{\mathcal{D}}_i = \mathcal{D}_i, i \neq l, k. \end{cases} \tag{B1}$$

Let $s_r = \iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y$, $\widetilde{s_i} = \iint_{\tilde{\mathcal{D}}_i} O(x,y)\mathrm{d}x\mathrm{d}y$ and $g(s) = \xi \frac{U^3\omega^3 s^3}{((1-\tau)T)^2}$. Since $\mathcal{D}_i, \forall i \in \mathcal{M}$ is the optimal BSs service areas division scheme, we have

$$\begin{aligned}
&\sum_{i=1}^{M} \left(\iint_{\mathcal{D}_i} L_i(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g\left(s_i\right) \right) \\
&\overset{(\Delta)}{\leqslant} \sum_{i=1}^{M} \left(\iint_{\tilde{\mathcal{D}}_i} L_i(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g(\widetilde{s_i}) \right) \\
&\Leftrightarrow \iint_{\mathcal{D}_l} L_l(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g\left(s_l\right) + \iint_{\mathcal{D}_k} L_k(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g\left(s_k\right) \\
&\leqslant \iint_{\tilde{\mathcal{D}}_l} L_l(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g(\widetilde{s_l}) + \iint_{\tilde{\mathcal{D}}_k} L_k(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g(\widetilde{s_k}) \\
&\Leftrightarrow \iint_{\mathcal{D}_r(e_0)} L_k(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g\left(s_k\right) - g(\widetilde{s_k}) \\
&\leqslant \iint_{\mathcal{D}_r(e_0)} L_l(x,y)O(x,y)\mathrm{d}x\mathrm{d}y + g(\widetilde{s_l}) - g\left(s_l\right),
\end{aligned} \tag{B2}$$

where $\overset{(\Delta)}{\leqslant}$ comes from the fact that $(\mathcal{D})_{i=1,\ldots,M}$ is optimal. Thus, any variation of that $(\tilde{\mathcal{D}})_{i=1,\ldots,M}$ cannot lead to a better solution. Now, we multiply both sides of the inequality in (B2) by $\frac{1}{s_r}$. Then, we take the limit when $s_r \to 0$, and we use the following equality:

$$\lim_{r\to 0} \frac{1}{s_r} \iint_{\mathcal{D}_r(e_0)} L_k(x,y)O(x,y)\mathrm{d}x\mathrm{d}y = \lim_{r\to 0} \frac{\iint_{\mathcal{D}_r(e_0)} L_k(x,y)O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}, \tag{B3a}$$

$$\lim_{r\to 0}\frac{1}{s_r}\iint_{\mathcal{D}_r(e_0)} L_l(x,y)O(x,y)\mathrm{d}x\mathrm{d}y=\lim_{r\to 0}\frac{\iint_{\mathcal{D}_r(e_0)} L_l(x,y)O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}, \tag{B3b}$$

$$\lim_{r\to 0}\frac{1}{s_r}\left(g\left(s_k\right)-g\left(\widetilde{s_k}\right)\right)=\lim_{r\to 0}\frac{g\left(s_k\right)-g\left(\widetilde{s_k}\right)}{s_r}, \tag{B3c}$$

$$\lim_{r\to 0}\frac{1}{s_r}\left(g\left(\widetilde{s_l}\right)-g\left(s_l\right)\right)=\lim_{r\to 0}\frac{g\left(\widetilde{s_l}\right)-g\left(s_l\right)}{s_r}. \tag{B3d}$$

According to Lemma 3, which is the extension of the double integral mean value theorem, Eqs. (B3a) and (B3b) can be obtained as follows:

$$\lim_{r\to 0}\frac{\iint_{\mathcal{D}_r(e_0)} L_k(x,y)O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}=\lim_{r\to 0}\frac{L_k(\mu_1,\eta_1)\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}, \tag{B4a}$$

$$\lim_{r\to 0}\frac{\iint_{\mathcal{D}_r(e_0)} L_l(x,y)O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}=\lim_{r\to 0}\frac{L_l(\mu_2,\eta_2)\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}, \tag{B4b}$$

where $(\mu_1,\eta_1)\in\mathcal{D}_r\left(e_0\right)$, $(\mu_2,\eta_2)\in\mathcal{D}_r\left(e_0\right)$.

Since the continuity of the function $L_i(x,y), i\in\{k,l\}$, it can be obtained according to (B4a) and (B4b)

$$\lim_{r\to 0}\frac{L_k(\mu_1,\eta_1)\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}=L_k(x_{k0},y_{k0}), \tag{B5a}$$

$$\lim_{r\to 0}\frac{L_l(\mu_2,\eta_2)\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)} O(x,y)\mathrm{d}x\mathrm{d}y}=L_l(x_{k0},y_{k0}). \tag{B5b}$$

Due to $g(s)$ is a continuously differentiable and nondecreasing convex function, it can be obtained according to (B3c) and (B3d)

$$\lim_{r\to 0}\frac{g\left(s_k\right)-g(\widetilde{s_k})}{s_r}=g'\left(s_k\right), \tag{B5c}$$

$$\lim_{r\to 0}\frac{g(\widetilde{s_l})-g\left(s_l\right)}{s_r}=g'\left(s_l\right). \tag{B5d}$$

Put (B5a)–(B5d) into (B2)

$$L_k(x_{k0},y_{k0})+g'\left(s_k\right)<L_l(x_{k0},y_{k0})+g'\left(s_l\right). \tag{B6}$$

Note that, Eq. (B6) provides the condition under which a point e_0 is assigned to the area $\mathcal{D}_k$ rather than the area $\mathcal{D}_l$. Therefore, the optimal BSs partition can be characterized as

$$\mathcal{D}_i^*=\left\{(x,y)|L_i(x,y)+\xi\frac{3U^3\omega^3{s_i}^2}{\left((1-\tau)T\right)^2}<L_j(x,y)+\xi\frac{3U^3\omega^3{s_j}^2}{\left((1-\tau)T\right)^2},\forall i\neq j\in\mathcal{M}\right\}. \tag{B7}$$

Putting (B7) into the problem (P2), the total energy consumption of the system in the area $\mathcal{D}$ can be obtained as

$$E_{\mathrm{OT}}=\sum_{i=1}^{M}\left(\iint_{\mathcal{D}_i^*} L_i(x,y)O(x,y)\mathrm{d}x\mathrm{d}y+\xi\frac{U^3\omega^3{s_i^*}^3}{\left((1-\tau)T\right)^2}\right). \tag{B8}$$