Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 6975890, 15 pages
https://doi.org/10.1155/2020/6975890

Research Article

A Multidepth Load-Balance Scheme for Clusters of Congested Cells in Ultradense Cellular Networks

Wei-Kuang Lai,[1] Ya-Ju Yu,[2] Pei-Lun Tsai,[1] and Meng-Han Shen[1]

[1]*Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan*
[2]*Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan*

Correspondence should be addressed to Ya-Ju Yu; yjyu@nuk.edu.tw

Received 17 February 2020; Revised 23 August 2020; Accepted 2 September 2020; Published 30 September 2020

Academic Editor: Tuan M. Nguyen

Densely deploying small cells will be a solution to provide explosive data requirements in fifth-generation networks. Because users often cluster together in popular locations of an urban area, congested small cells in the popular sites are also gathered. Traditional load balancing schemes generally only consider neighboring cells when offloading users are not suitable for clusters of overloaded cells. This paper considers groups of overloaded cells in the load balancing problem. The objective is to maximize the quality-of-service satisfaction ratio. To solve the problem, we propose a multidepth offloading algorithm with the consideration of the radio resource allocation. The proposed multidepth offloading algorithm can be applied no matter that congested small cells are gathered or not. Compared with a previous offloading algorithm and a baseline, the simulation results show that our proposed algorithm can increase 16% QoS satisfaction ratio in a real user distribution and 13% QoS satisfaction ratio in a clustered user placement.

1. Introduction

Advances of mobile devices and wireless communications bring the rapid development of various network applications so that mobile data traffic explosively grows. According to the Cisco forecast, 8.4 billion mobile devices would be in use, and total mobile data traffic in 2022 is about 77 exabytes per month, which is a seven-fold increase over 2017 [1]. Densely deploying small cells will be a solution to provide explosive data requirements in fifth-generation (5G) networks. Using small cells is one of the offloading methods to moderate traffic loads of macrocells [2], and the performance improvement of the network with small cells has been analyzed in [3].

Although small cells can be densely deployed to provide the above advantages, each small cell can only serve a few users because of their low capacities. The loading distribution among small cells will be significantly affected by user locations. Users often cluster together in popular sites of an urban area. Therefore, 3GPP specifies a clustered user placement model following the observation that users generally gather in certain *hot-zones* in urban areas [4, 5]. Clustered users in popular locations will typically connect to the same small cells. Therefore, congested cells in famous places are also grouped. As a result, some small cells will quickly exhaust their radio resources while other small cells operate on light loads. The imbalanced situation will decrease the quality-of-service (QoS) satisfaction ratio and waste the resources of lightly loaded small cells.

In 5G cellular networks, telecom operators and service providers have changed their focus from network quality of service (QoS) to user quality of experience (QoE). Various emerging services indicate that the overall network performance should be improved from a user's subjective view [6]. A general relationship between QoE and QoS is the exponential hypothesis [7]. When the QoS of video applications cannot be satisfied, a user's QoE will be decreased exponentially [7, 8]. Therefore, to increase the QoE of users in cellular networks, maximizing the QoS satisfaction ratio is essential. Improving the QoE of users motivates us to maximize the QoS satisfaction ratio as our objective function. This paper proposes a concept, namely, the *multidepth offloading*

mechanism, to solve the problem mentioned above in ultradense cellular networks.

The idea is described as follows.

Considering an overloaded small cell, namely, a *root cell*, traditional offloading schemes find its neighboring cells called *1-depth cells*. Because the offloading algorithms only consider 1-depth cells, they are called *1-depth offloading mechanisms*. However, when users gather in popular locations, 1-depth cells in the areas may also be overloaded. Therefore, traditional algorithms cannot offload the traffic of the root cell. We can firstly offload the users of 1-depth cells to their adjacent small cells called 2-depth cells. When loads of 1-depth cells have been considerably lightened, the root cell can offload its traffic to the 1-depth cells. The offloading procedure considering the 2-depth cells is called the 2-depth offloading.

With a similar concept, when 1-depth and 2-depth cells with full loads are clustered, we can execute 3-depth offloading to firstly lighten loads of 2-depth cells and secondly reduce loads of 1-depth cells; then, the root cell can offload traffic to the 1-depth cells. Consequently, when more overloaded cells are clustered, we can rely on a deeper offloading depth. Because small cells are densely deployed in ultradense cellular networks, we can search for high depth cells with light loads when overloaded cells are clustered.

This paper studies the load balancing problem in ultradense cellular networks. The objective is to maximize the QoS satisfaction ratio under a radio resource constraint. The contributions of this paper are summarized as follows.

(i) This paper considers clusters of overloaded cells in ultradense cellular networks and observes that traditional approaches for the load balancing problem with the consideration of radio resource allocation are not effective in clusters of overloaded cells.

(ii) We propose a multidepth offloading algorithm with the considerations of resource allocation and load balance of neighboring cells to maximize the QoS satisfaction ratio. The proposed algorithm can be used no matter that overloaded small cells are grouped or not.

(iii) We conduct a series of simulations. In our simulations, we consider a real user distribution, where user locations are extracted from Tokyo data sets and a clustered user placement with the user distribution model specified by 3GPP. Compared with the offloading algorithm designed by [9] and a baseline, the simulation results show that our proposed algorithm can improve 16% QoS satisfaction ratio in a real user distribution and 13% QoS satisfaction ratio in a clustered user placement.

The rest of the paper is organized as follows. In Section 2, we review the related works. In Section 3, we represent the system model and problem formulation. Section 4 explains the proposed algorithm.

Simulation results are presented in Section 5. Section 6 concludes this paper.

2. Related Works

Recently, many papers have studied the load balancing problem, and the proposed methods can be classified into four classes: The first direction studied cell breathing techniques, which adjust the transmit power to shrink (or expand) the coverage area of each congested base station, e.g., [10, 11]. The second direction adapts handover parameters to achieve load balance, e.g., [12, 13]. The third way uses device-to-device communications to offload users in overloaded cells, e.g., [14]. Recently, Farzi et al. proposed a game-theoretic approach with the consideration of two-tier cellular networks to improve load distribution among heterogeneous cells [15]. Each macrocell and its associated picocells play a game to distribute the load among themselves. However, the work does not consider the radio resource constraint and the resource allocation.

This paper targets the fourth class, which investigates the resource allocation centrally controlled by a controller for load balance in multicell scenarios, e.g., [9, 16–19]. In [16, 17], a controller can collect related information and transmits a user association decision to each base station operating on a fractional frequency reuse mechanism. Each base station can inform users to execute the handover procedure. Then, in each cell, the works studied how to allocate frequency under a fractional frequency reuse policy for each user to balance spectrum utilization. In [18], unlicensed bands are utilized to offload traffic of licensed bands, and the particle swarm optimization algorithm is adopted for channel selection and power allocation to maximize throughput. In [19], Du et al. considered that a controller is unaware of the offloading performance of small cells for a macro base station and designed contract-based traffic offloading and resource allocation methods. In [20], Chung and Cho considered Wi-Fi access points to reduce traffic loads of macrocells in heterogeneous networks. They designed a load balancing approach based on coalition game theory. The most related work of our paper is [9]. Hasan et al. [9] defined a resource block-utilization ratio to measure a cell's load and adopted an adaptive threshold for defining an overloaded cell in multicell networks. The objective is to minimize the standard deviation of resource block utilization among the cells. In [9], the proposed algorithm determined the order of base stations to execute the user offloading procedure by considering loads of cells and the order of edge users to perform the handover by rating their signal strengths.

However, the related works have not considered clusters of overloaded cells in ultradense cellular networks. Table 1 summarizes the comparison between related works and our proposed method, where homogeneous and heterogeneous are, respectively, abbreviated to homo and hetero for short.

3. System Model and Problem Formulation

3.1. System Model. In cellular networks, each small cell has a limited number of radio resource blocks. A resource block (RB) consists of 12 subcarriers in the frequency domain and multiple orthogonal frequency-division multiplexing

Table 1: Comparison between the related works and the proposed method.

Work	Scenario	Method	Multidepth offloading	Features
[16]	Homo.	An offline and an online algorithm		A fractional frequency reuse mechanism
[17]	Homo.	Considering intra- and intercell handover		A fractional frequency reuse mechanism
[18]	Homo.	Considering channel and power allocation		Unlicensed and licensed bands
[19]	Hetero.	Considering asymmetric information		A controller is unaware of the offloading performance of small cells
[20]	Hetero.	Considering fairness		Wi-Fi access points to reduce traffic loads of macrocells
[9]	Homo.	Considering RB utilization ratio, cells' loads, and signal qualities		Multicell networks
Proposed	Homo.	Considering load balance of neighboring cells	✓	Clusters of congested cells in ultradense cellular networks

(OFDM) symbols in the time domain. A resource block can provide a higher data rate to a user when the user closer to the small cell has a higher signal-to-interference-plus-noise ratio (SINR). Each user has a data rate requirement for receiving the requested downlink data. We define that a user is satisfied when the user can receive his/her required data rate. Therefore, different small cells should consume different numbers of resource blocks for satisfying the data requirement of a user depending on the link quality between a small cell and a user.

Each small cell can collect the information on resource block consumption for serving each user and send the information to a centralized controller. When the ratio of the allocated resource blocks to the total resource blocks is higher than a predefined threshold, a small cell can be seen as an overloaded cell. Because users often gather, burdened small cells are also grouped [4], where a scenario is shown in Figure 1. In Figure 1, considering an overloaded cell as a root cell, the neighboring cells of the root cell are called the 1-depth cells, and the adjacent cells of the 1-depth cells are named 2-depth cells. Because the 1-depth cells and the root cell are overloaded, we should firstly offload traffic of the 1-depth cells to the 2-depth cells by executing 2-depth offloading. If 1-depth and 2-depth cells are overloaded, we can perform 3-depth offloading.

A controller can connect with each small cell via the S1 interface and can collect the information from each small cell to make a handover decision for each small cell. A small cell can exchange handover messages with each neighboring cell via the X2 interface. Figure 2 shows the handover procedure with a controller. The steps are described as follows.

(1) In step 1, a serving cell transmits the measurement control message to each user. Then, each user will measure the reference signal receiving power (RSRP) from the cell.

(2) In step 2, each user reports the instantaneously measured RSRP value (i.e., the channel state information) to its serving cell. Link qualities can be periodically reported by users or proactively requested by its serving cell.

(3) In step 3, the serving cell relays the RSRP of each user and its loading information to a controller.

(4) In step 4, the controller executes a load balancing algorithm based on the loading information of each small cell and the instantaneous RSRPs of the users. Then, the controller can deliver a handover decision to each small cell.

(5) In step 5, the serving cell transmits the handover request to a neighboring cell decided by the offloading algorithm.

(6) In step 6, the neighboring cell replies the acknowledgment for the handover request.

(7) In step 7, the serving cell informs an edge user to perform the handover procedure.

Based on the above procedure, the information required by the controller from each device only is the RSRP value. In step 2, the channel quality report is a snapshot value. No matter the device's speed, the device will report the RSRP value based on its current situation. Therefore, considering user mobility does not impact the load balancing problem, and this paper considers that users are stationary.

3.2. Problem Formulation. In this paper, we study the load balancing problem in ultradense cellular networks considering clustered congestion cells. The objective is to maximize the QoS satisfaction ratio under overloaded cells. The target problem is formulated as follows.

We consider a set of small cells (small base stations) B to serve a set of users U. Let U_j denote the set of users in the coverage area of small cell j. B_j^A expresses the set of adjacent small cells of small cell j. The set of small cells B_i^C covers user i. Small cell j can offload user $i \in U_j$ to one of neighboring small cells $j \in B_i^C$, where $B_i^C \subseteq B_j^A$.

Small cell j has $\mathcal{N}_j$ resource blocks. According to Shannon capacity, small cell j using a resource block can provide $R_{i,j}$ data rate for user i. $R_{i,j}$ will be updated according to the measurement report of each user when each serving

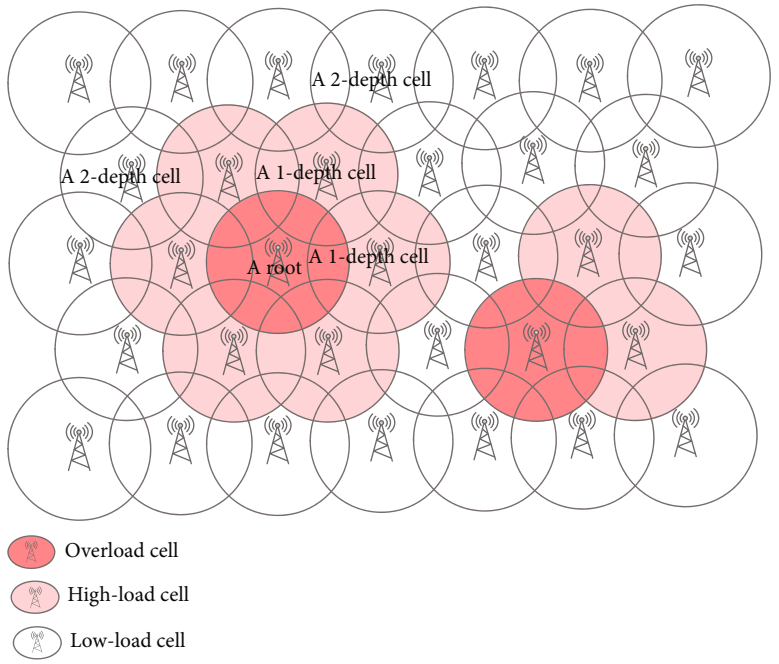

Figure 1: A scenario of clustered cells with heavy loads under densely deployed small cells.

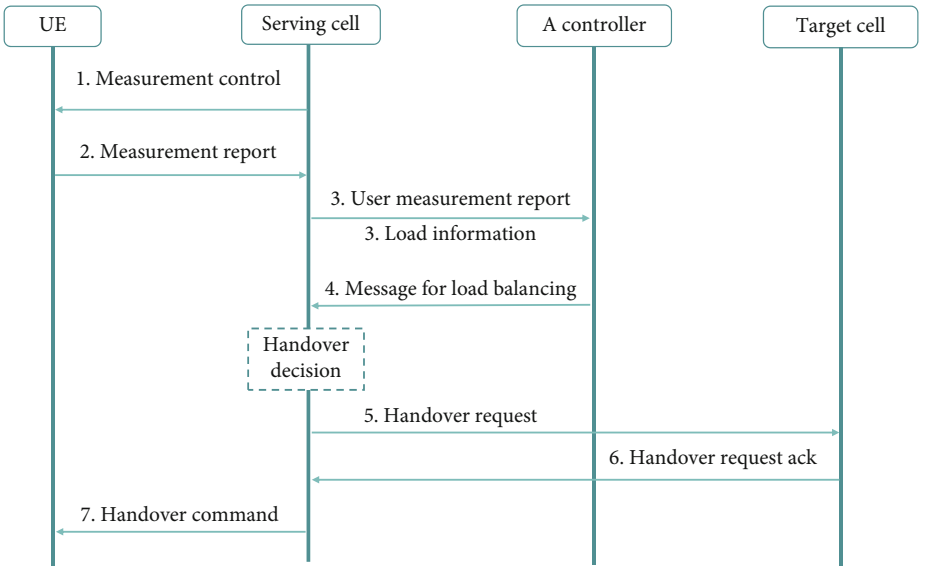

Figure 2: Procedure with a controller for handover.

cell sends the measurement control signal as shown in Figure 2 and is defined as follows.

$$R_{i,j} = W \log_2\left(1 + \mathrm{SINR}_{i,j}\right), \tag{1}$$

$$\mathrm{SINR}_{i,j} = \frac{\mathrm{RSRP}_{i,j}}{\omega + \sum_{j' \in B_i^C \{j\}} \mathrm{RSRP}_{i,j'}}, \tag{2}$$

where $\mathrm{SINR}_{i,j}$ is the signal-to-interference-plus-noise ratio between small cell j and user i, and W is the bandwidth of a resource block. ω is a noise power value. $\mathrm{RSRP}_{i,j}$ is the reference signal receiving power between small cell j and user i.

The data rate requirement of user i is D_i. Small cell j should consume at least $N_{i,j}$ resource blocks to satisfy the data rate requirement of user i. $N_{i,j}$ is represented as follows.

$$N_{i,j} = \left\lceil \frac{D_i}{R_{i,j}} \right\rceil. \tag{3}$$

$N_{i,j} = \infty$ if small cell j cannot serve user i. We define a utility function $U(\mathscr{R}_{i,j} - D_i)$ as follows.

$$U\left(\mathscr{R}_{i,j} - D_i\right) = \begin{cases} 1, & \text{if } \mathscr{R}_{i,j} - D_i \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$\mathscr{R}_{i,j}$is the total data rate allocated by serving small cell j for user i. $U(\mathscr{R}_{i,j} - D_i) = 1$ means that the data requirement of user i is satisfied, and 0 otherwise.

The set of overloaded small cells is B^O. We define that an overloaded small cell $j \in B^O$ means that a load of a small cell is higher than a threshold ρ_{TH}, where $\rho_{TH} \leq 1$. The threshold can avoid exhausting all the resource blocks of a small cell to serve users reconnecting from other small cells. This paper attempts to offload the excess traffic of each overloaded small cell $j \in B^O$ to lightly loaded small cells such that the number of satisfied users is maximized. A load balancing solution is feasible if the following constraints are met.

Resource Constraint: Each small cell j cannot allocate the number of resource blocks over its number of available resource blocks.

$$\sum_{i \in U_j} X_{i,j} N_{i,j} \leq \mathscr{N}_j, \forall j \in B. \tag{5}$$

Serving Constraint: Each user can receive data from at most one small cell.

$$\sum_{j \in B} X_{i,j} \leq 1, \forall i \in U. \tag{6}$$

We now define the load balancing problem.

Input Instance: Consider a set of small cells B to serve a set of users U. Small cell j can serve a set of users U_j. The adjacent small cell set of small cell j is B_j^A. The set of small cells B_i^C covers user i. Each small cell j has $\mathscr{N}_j$ resource blocks. Small cell j using a resource block can provide data rate $R_{i,j}$ for user i. The set of overloaded small cells is B^O when the loading threshold is ρ_{TH}.

Output Variable: $X_{i,j}$.

Objective: Our objective is to balance loads of the overloaded small cells such that the QoS satisfaction ratio under the overloaded cells is maximized. We state our objective function formally as follows:

$$\max \frac{\sum_{j \in B^O} \sum_{i \in U_j} U\left(\mathscr{R}_{i,j} - D_i\right)}{\sum_{j \in B^O} \sum_{i \in U_j} X_{i,j}}, \tag{7}$$

subject to constraints (5) and (6). The variables used in this problem are summarized in Table 2.

3.3. An Illustrative Example. Figure 3 illustrates an example of the multidepth offloading concept. There are three cells and three users. As shown in Figure 3(a), initially, cells A and B are a cluster of clustered cells. Cell A (a root cell) serves user equipment 1 (UE1) and does not remain any available resource block to serve UE3. Cell B (1-depth cell) serves UE2 and remains two resource blocks. Cell C (2-depth cell) does not serve any UE and has five resource blocks. The number of required resource blocks for serving each user by each cell is shown in Table 3. Cell A cannot offload UE1 to neighboring cell B because cell B is also overloaded. In other words, 1-depth offloading is not valid in this situation. As shown in Figure 3(b), we should execute 2-depth offloading in this example. Cell B should firstly offload UE2 to cell C and can release two resource blocks. Cell C serves UE2 consuming three resource blocks. As shown in Figure 3(c), cell A can offload UE1 to cell B consuming three resource blocks to serve UE1. Thus, cell A has two resource blocks to serve UE3. The example shows that the proposed multidepth offloading concept is required when overloaded cells are gathered.

Table 2: Summary of notations.

Symbol	Description
B	The set of small cells
U	The set of users
U_j	The set of users that can be served by small cell j
B_j^A	The set of adjacent small cells of small cell j
B_i^C	The set of small cells covers user i
$\mathscr{N}_j$	The number of resource blocks of small cell j
$X_{i,j}$	The indicator function, which is 1 if small cell j serves user i, and 0 otherwise
$R_{i,j}$	The data rate of a resource block allocated by small cell j to user i
$\mathscr{R}_{i,j}$	The total data rate provided by small cell j to user i
D_i	The data rate requirement of user i
$N_{i,j}$	The number of consumed resource blocks of small cell j for satisfying the data rate requirement of user i
B^O	The set of overloading small cells
ρTH	The loading threshold of being an overloaded cell

4. Multidepth Offloading Algorithm

In this section, we propose a multidepth offloading algorithm (MDOA) to maximize the QoS satisfaction ratio. The proposed algorithm is based on a recursive procedure and addresses two decision problems. (1) We should choose which overloaded cell as a root cell to offload users firstly. (2) Then, after determining an overloaded small cell (a root cell), we should select some users served by the overloaded small cell reconnecting to the neighboring cells of the overloaded cell.

4.1. Selection of Overloaded Small Cells. We define a utility function Γ_j for each small cell j. The utility function is designed for selecting which overloaded cell should be first offloaded. Utility function Γ_j simultaneously considers the remaining resource blocks of each neighboring cell of cell j

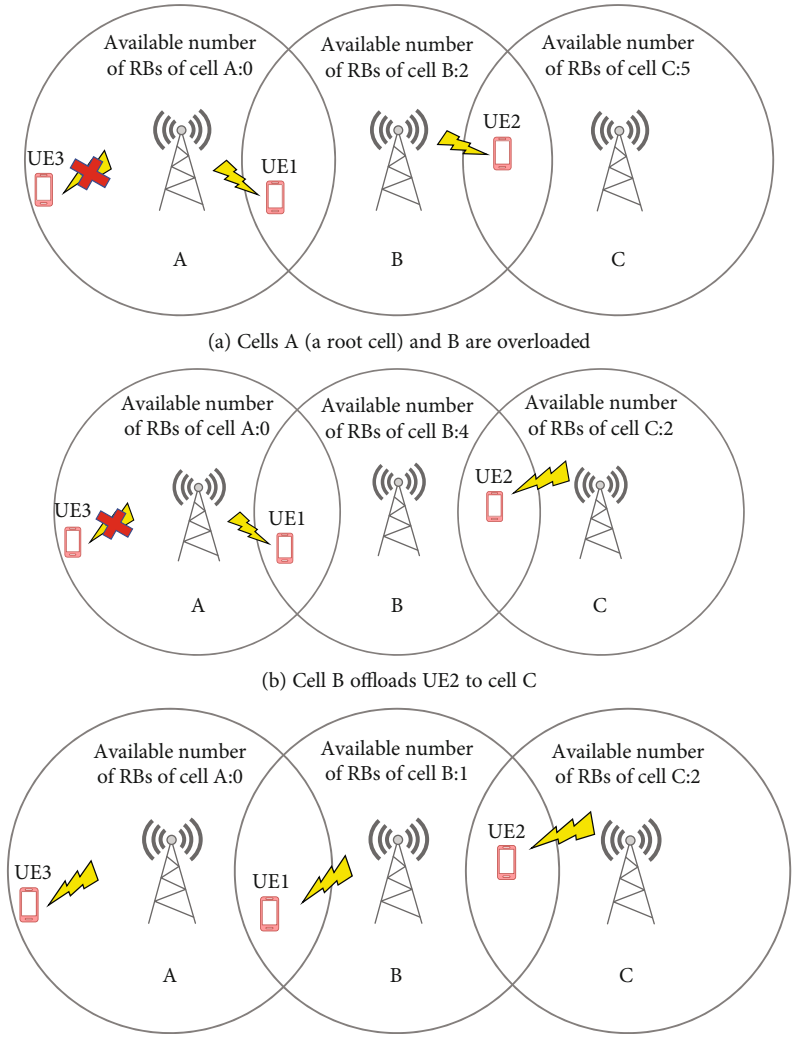

Figure 3: An example of the multidepth offloading concept.

Table 3: The number of required RBs for each user served by each cell.

	RBs consumed by cell A	RBs consumed by cell B	RBs consumed by cell C
UE1	2	3	∞
UE2	∞	2	3
UE3	2	∞	∞

and link qualities between the neighboring cells of cell j and the users of cell j. This design can achieve load balance of neighboring cells. Utility function Γ_j is shown in Eq. (8).

$$\Gamma_j = \frac{\sum_{i \in U_j} \sum_{j \in B_i^C} N_{i,j} / \left| B_i^C \right| + \epsilon}{\sum_{j \in B_j^A} \widehat{\mathcal{N}}_j / \left| B_j^A \right| + \epsilon},$$
$$\widehat{\mathcal{N}}_j = \max \left(\mathcal{N}_j - \sum_{i \in U_j} X_{i,j} N_{i,j}, 0 \right). \tag{8}$$

$\widehat{\mathcal{N}}_j$ is the remaining resource blocks of small cell j after serving the users in set U_j. In Eq. (8), the numerator term is the average resource block consumption of all the neighboring small cells for serving the users of cell j. If the numerator value is small, it means that the neighboring small cells can use fewer resource blocks for serving the users, and we can successfully offload more users to maximize the QoS satisfaction ratio. In other words, the neighboring cells and the users have better link qualities. ϵ is a constant value to avoid the denominator being zero.

In the denominator term of Eq. (8), we evaluate the average remaining resource blocks of all the adjacent small cells of overloaded small cell j. If the denominator value is large, the adjacent small cells have more remaining resource blocks. When the denominator value is larger, we can offload more users to the cells and can avoid exhausting the resource blocks of specific cells. Therefore, we will select the overloaded small cell with the smallest Γ_j to execute the load balancing procedure first.

Then, we select an overloaded small cell j^o in B^O as a root cell to offload traffic based on Eq. (9), which addresses the first decision problem. Small cell j^o with the smallest Γ_j among all the overloaded cells will be selected first. It means that the neighboring cells of small cell j^o have more remaining resource blocks and consume fewer resource blocks for serving users.

$$j^o = \arg \min_{j \in B^O} \Gamma_j. \tag{9}$$

4.2. Selection of Offloaded Users and Target Cells. After selecting the overloaded cell (the root cell) j^o based on Eq. (9), we should address the second decision problem. We define a utility function $m_{i,j}$ for each user i of overloaded cell j^o and each neighboring cell j of overloaded cell j^o as shown in Eq. (10). The utility function is designed for selecting which target user i of overloaded cell j^o should be offloaded to which neighboring cell j of overloaded cell j^o. Eq. (10) can equally and efficiently use the resource blocks of each small cell shown as follows.

$$m_{i,j} = \frac{N_{i,j}}{\widehat{\mathcal{N}}_j + \epsilon}, \tag{10}$$

In Eq. (10), $m_{i,j}$ simultaneously considers the consumed resource blocks of small cell $j \in B^A_{j^o}$ for serving user i in the numerator and the remaining resource blocks of small cell $j \in B^A_{j^o}$ in the denominator.

If user i consumes fewer resource blocks of small cell j, and small cell j has more available resource blocks; user i has a higher probability that can successfully handover to small cell j. This design will not always select the same small cell to avoid exhausting the radio resource blocks of a specific small cell. Thus, if some users are covered by only one neighboring small cell, the users also have chances to be offloaded. When user i cannot be served by small cell j, $m_{i,j}$ is set as ∞.

Because overloaded cell j^o serves $|U_{j^o}|$ users and has $|B^A_{j^o}|$ neighboring cells, the overloaded cell has a $|U_{j^o}| \times |B^A_{j^o}|$ utility matrix M_{j^o} as follows.

$$M_{j^o} = \begin{bmatrix} m_{1,1} & \cdots & m_{1,\left|B^A_{j^o}\right|} \\ \vdots & \ddots & \vdots \\ m_{\left|U_{j^o}\right|,1} & \cdots & m_{\left|U_{j^o}\right|,\left|B^A_{j^o}\right|} \end{bmatrix}. \tag{11}$$

According to utility matrix M_{j^o}, we can choose a target user i^t of overloaded cell j^o offloading to a neighboring cell j^t of overloaded cell j^o based on Eq. (12). The pair of target user i^t and neighboring cell j^t has the smallest utility value $m_{i,j}$ in matrix M_{j^o}. It means that target user i^t consumes fewer resource blocks of neighboring cell j^t which has more remaining resource blocks.

$$\left(i^t, j^t\right) = \arg \min_{i \in U_{j^o}, j \in B^A_o} m_{i,j}. \tag{12}$$

After user i^t is selected, we should determine whether user i^t can associate with small cell j^t or not. If the load of the small cell is less than or equal to the loading threshold ρ_{TH}, the user can successfully handover to the small cell, i.e., $\rho_{j^t} + \rho_{i^t,j^t} \le \rho_{TH}$, where

$$\rho_{i^t,j^t} = \frac{N_{i^t,j^t}}{\mathcal{N}_{j^t}}, \tag{13}$$

$$\rho_{j^t} = \frac{\mathcal{N}_{j^t} - \widehat{\mathcal{N}}_{j^t}}{\mathcal{N}_{j^t}}. \tag{14}$$

If target user i^t successfully connects to target small cell j^t, m_{i^t,j^t} will be set as ∞ to avoid that we select the same user again. Finally, we will update the load information of overloaded small cell j^o and target small cell j^t as well as matrix M_{j^o} and M_{j^t}. If $m_{i,j} = \infty$, $m_{i,j}$ will not be updated.

4.3. Algorithm Description. Algorithm 1 shows the pseudocode for the proposed multidepth offloading algorithm. The proposed algorithm is executed based on recursion. When a root cell cannot offload a user to a target neighboring cell, which is also overloaded, the algorithm will perform further depth offloading for the target neighboring cell. The recursion is terminated when the target depth of $\mathcal{D}$ for the target neighboring cell is reached. If the root cell is still overloaded, the algorithm will search for the next pair of a target user and a target neighboring cell to release the load of the root cell.

Based on Eq. (9), we find an overloaded small cell (a root cell) $j^o \in B^o$, while there is at least one overloaded small cell (i.e., $B^o \neq \varnothing$) (Lines 1-2). The selected cell j^o is removed from set B^o to avoid choosing the same cell again (Line 3). Variable d is the offloading depth of d and is initially set as 0 (Line 4). In Line 5, for an overloaded small cell, $\mathcal{B}_{\text{prev}}$ is used to record the previous serving cells of users for avoiding the ping-pong effect and is initially set as an empty set. Then, we try to offload the traffic of the overloaded small cell, until the load of the small cell is less than 1 (i.e., $\rho_{j^o} \le 1$) (Lines 6-13).

In Lines 7-8, if there exists a user that can be offloaded to a neighboring small cell (i.e., $\exists m_{i,j} \neq \infty$), we select the pair of a target user i^t and a target small cell j^t with the smallest $m_{i,j}$ according to Eq. (12). Then, we set m_{i^t,j^t} as ∞ to avoid that we pick the same pair again (Line 9). Cell j^o is added to set $\mathcal{B}_{prev}$ to avoid that users offloaded to neighboring cells reconnect back to cell j^o (Line 10). We call function Next_Depth() to check whether user i^t can be offloaded to small cell j^t, and the next offloading depth is needed or not. If the user of the root cell cannot be offloaded to the neighboring cell, we will recursively execute further depth offloading to release the load of the neighboring small cell. Otherwise, if the user can be offloaded to the neighboring cell, we will not need to call

Input: $B, U, L, U_j, B_j^A, B_i^C, \mathcal{N}_j, R_{i,j}, B^o$
Output: $X_{i,j}$
1 **while** $B^o \neq \varnothing$ **do**
2 $j^o = \arg\min_{j \in B^o} \Gamma_j$
3 $B^o = B^o \setminus \{j^o\}$
4 $d = 0$
5 $\mathscr{B}_{prev} = \phi$
6 **while** $\rho_{j^o} > 1$ **do**
7 **if** $\exists m_{i,j} \neq \infty, \forall i \in U_{j^o}, j \in B_{j^o}^A$ **then**
8 $(i^t, j^t) = \arg\min_{i \in U_{j^o}, j \in B_{j^o}^A} m_{i,j}$
9 $m_{i^t,j^t} = \infty$
10 $\mathscr{B}_{prev} = \mathscr{B}_{prev} \cup \{j^o\}$
11 Next_Depth$(j^o, i^t, j^t, d+1)$
12 **else**
13 break
14 **return** $X_{i,j}, \forall i, j$

Algorithm 1: Multidepth offloading algorithm.

1 **Function** Next_Depth (j^o, i^t, j^t, d):
2 **while** $\rho_{j^t} + \rho_{i^t j^t} > \rho_{TH}$ **do**
3 **if** $d + 1 \leq \mathscr{D}$ **then**
4 **while** *True* **do**
5 **if** $\exists m_{i,j} \neq \infty, \forall i \in U_{j^t}, j \in B_{j^t}^A$ **then**
6 $(i^d, j^d) = \arg\min_{i \in U_{j^t}, j \in B_{j^t}^A} m_{i,j}$
7 **if** $j^d \notin \mathscr{B}_{prev}$ **then**
8 $\mathscr{B}_{prev} = \mathscr{B}_{prev} \cup \{j^t\}$
9 $m_{i^d,j^d} = \infty$
10 Next_Depth$(j^t, i^d, j^d, d+1)$
11 break
12 **else**
13 **return** store the updated results
14 **else**
15 **return** restore the updated results
16 Let user i^t associate with small cell j^t
17 Update X_{i^t,j^t} and X_{i^t,j^o}
18 Update $M_{j^o}, M_{j^t}, \rho_{j^o}$, and ρ_{j^t}
19 **return**

Algorithm 2:

function Next_Depth() (Line 11). If the root cell is still overloaded, we will find the next pair of a target user and a target small cell with the smallest $m_{i,j}$ in Line 7 until the loading threshold of the root cell is less than 1. When we cannot find any user to be offloaded from overloaded small cell j^o, we break the while loop and select the next overloaded small cell (Lines 12-13).

Function Next_Depth() takes overloaded small cell j^o, target user i^t, target small cell j^t, and a depth of d as inputs. In Line 2, if target small cell j^t can serve target user i^t (i.e., $\rho_{j^t} + \rho_{i^t,j^t} \leq \rho_{TH}$), user i^t can associate with target small cell j^t (Line 16). The corresponding functions, X_{i^t,j^t}, X_{i^t,j^o}, M_{j^o}, M_{j^t}, ρ_{j^o}, and ρ_{j^t}, are updated (Lines 17-18). Because the user can be offloaded, we do not require further depth offloading and return this function (Line 19).

When neighboring cell j^t is overloaded (i.e., $\rho_{j^t} + \rho_{i^t,j^t} > \rho_{TH}$), the neighboring cell does not have sufficient resource blocks to serve target user i^t. In Line 3, if the $(d+1)$-th depth is less than or equal to the predefined depth of $\mathscr{D}$, we will execute $(d+1)$-depth offloading to release traffic of target small cell j^t to its neighboring cells (i.e., $(d+1)$-depth cells). When the target depth is reached, we terminate the offloading process for neighboring cell j^t and restore the corresponding functions which have been updated before (Lines 14-15). In the while loop of Lines 4-13, if there exists a user of cell j^t that can be offloaded (i.e., $\exists m_{i,j} \neq \infty, \forall i \in U_{j^t}, j \in B_{j^t}^A$), we will find the combination of a user i^d reconnecting to a neighboring small cell j^d with the minimum $m_{i,j}$ based on Eq. (12). Otherwise, because no user can be offloaded, we terminate the offloading process for neighboring cell j^t and restore the corresponding functions updated before (Lines 12-13).

If small cell j^d resides in set $\mathscr{B}_{prev}$, it means that user i^d will reconnect back to the previous one of overloaded cells. Therefore, we will find the next pair of a user and a small cell in Lines 4-5. If small cell $j^d \notin \mathscr{B}_{prev}$, target small cell j^t is added to set $\mathscr{B}_{prev}$, and m_{i^d,j^d} is set as ∞. Then, we call function Next_Depth() to decide whether small cell j^d can serve user i^d or we need to execute the next depth offloading procedure for releasing the load of cell j^d (Line 10). When function Next_Depth() returns, we break the while loop of Line 4 and go to the while loop of Line 2 to check whether the load of target small cell j^t is lower than the threshold ρ_{TH}. If yes, user i^t can associate with small cell j^t, and the corresponding functions are updated. If not, we continuously find the next pair of a user and a neighboring cell for offloading the load of target small cell j^t in Line 5.

Theorem 1. *The time complexity of Algorithm 1 is* $O(\bar{U}^2\bar{B})^{\mathscr{D}}$, *where* $\bar{U} = \max_{j \in B} U_j$ *and* $\bar{B} = \max_{j \in B} B_j^A$.

Proof. For each overloaded small cell j^o, because matrix 1 M_{j^o} has at most $\bar{U} \times \bar{B}$ elements, it takes $O(\bar{U}\bar{B})$ to find a pair of a user and a neighboring small cell. Since there are at most $\bar{U}$ users in an overloaded small cell j^o, the while loop of Lines 6-11 takes $O(\bar{U}^2\bar{B})$. If the target neighboring small cell does not have sufficient resource blocks for serving a user, we will recursively execute Next_Depth() function. The time complexity of the function is the same with Algorithm 1. When the target depth is $\mathscr{D}$, the time complexity of Algorithm 1 is $O(\bar{U}^2\bar{B})^{\mathscr{D}}$.

5. Performance Evaluation

5.1. Simulation Setups. In this section, we developed a simulation to evaluate our proposed algorithm. The proposed

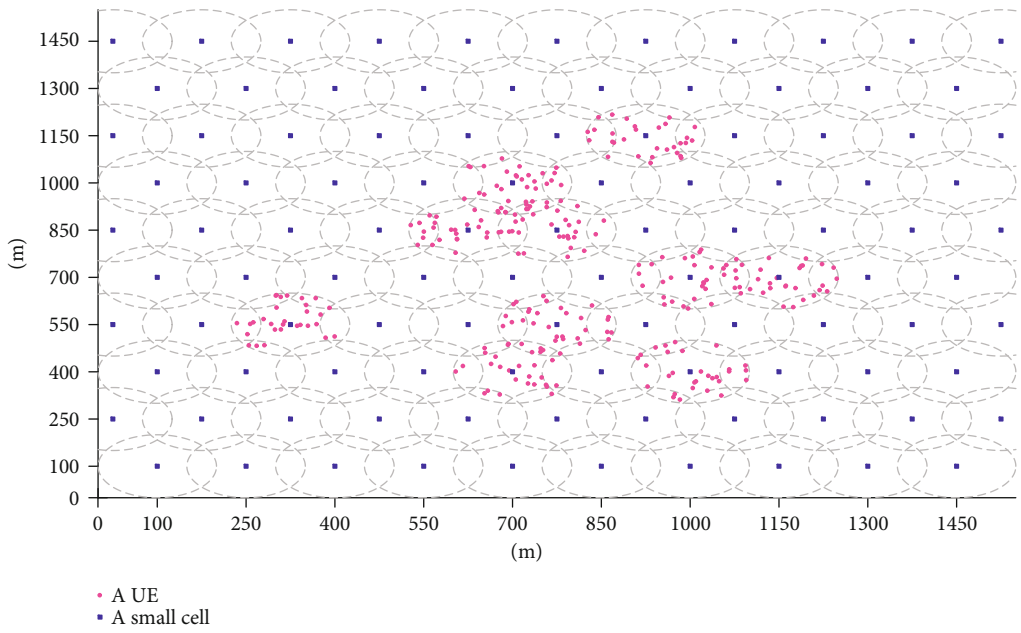

Figure 4: A simulation topology with a regular deployment.

multidepth offloading algorithm, denoted as *MDOA*, is compared with two baselines. *MDOA(1-D)* and *MDOA(2-D)*, respectively, mean that the proposed algorithm executes 1-depth offloading and 2-depth offloading. The first baseline is named *Max-RSRP*. Under *Max-RSRP*, each user directly connects to the small cell with the best signal quality without considering a load balancing mechanism. The second baseline, denoted as *AMLB*, is designed by [9] to minimize the standard deviation of resource block utilization. The order of small cells for offloading users is from the small cell with the heaviest load to that with the lightest load. *AMLB* also considers the signal quality for the order of offloading the users. The user having the best signal quality will be firstly selected. The selected user will reconnect to the neighboring cell with the best signal quality.

We consider a topology with densely deployed small cells, as shown in Figure 4. In Figure 4, the small cells are regularly deployed with an intersite distance of 150 meters between two cells. There are 105 small cells in the topology. We randomly select ten small cells as hot-zones. We adopt the clustered user placement model specified by 3GPP [4] to distribute users for the observation that users often group in hot-zones of urban areas [5]. The model is shown as follows.

$$H = \left\lfloor p \times \frac{|U|}{|B|} \right\rfloor. \tag{15}$$

H is the number of users in a hot-zone area. p is the ratio of all hot-zone users to the total number of users in the network. $|U|$ and $|B|$, respectively, are the total number of users and cells. We set the number of users in a hot-zone as 30 (i.e., $H = 30$) by adjusting p. After arranging the hot-zone users, other users are randomly placed in the topology. We additionally consider a real user distribution shown in Figure 5, where user locations are extracted from Tokyo data sets in the authors' website [21]. If a user's distance is far away from its serving base station between 67 and 100 meters, the user is called an edge user. We will evaluate the performance of each algorithm in terms of edge users' throughput. Each user requires a constant bit rate of 512 kbps. The bandwidth of a small cell is 10 MHz, and the number of resource blocks of a small cell is 50. The transmit power of a small cell is set as 30 dBm [22]. We adopt the pass loss model as $140.7 + 36.7\log_{10}(\delta)$ [23], where δ is in kilometers. The thermal noise is -174 dBm/Hz [23]. We set the threshold of being an overloading small cell is 0.9. The parameter settings are listed in Table 4.

5.2. Simulation Results

(1) *Clustered User Placement:* Figure 6 shows the satisfaction ratio derived by our proposed algorithm for 1-depth offloading to 3-depth offloading under all the small cells. 0-depth offloading (0-D) means that we do not offload any user for each overloaded cell. When the depth is deeper, the proposed algorithm will try to release the resource blocks of more cells for an overloaded small cell. Therefore, we can get a higher satisfaction ratio. Compared with 1-depth offloading, 2-depth offloading has visible performance improvement under 1500 and 2000 users. The performance improvement between 2-depth and 3-depth offloading is marginal because our simulation settings make a cluster of a root cell and 1-depth cells overloaded so that 3-depth offloading is not needed. Therefore, the proposed algorithm only adopt 1depth and 2-depth offloading in our simulation

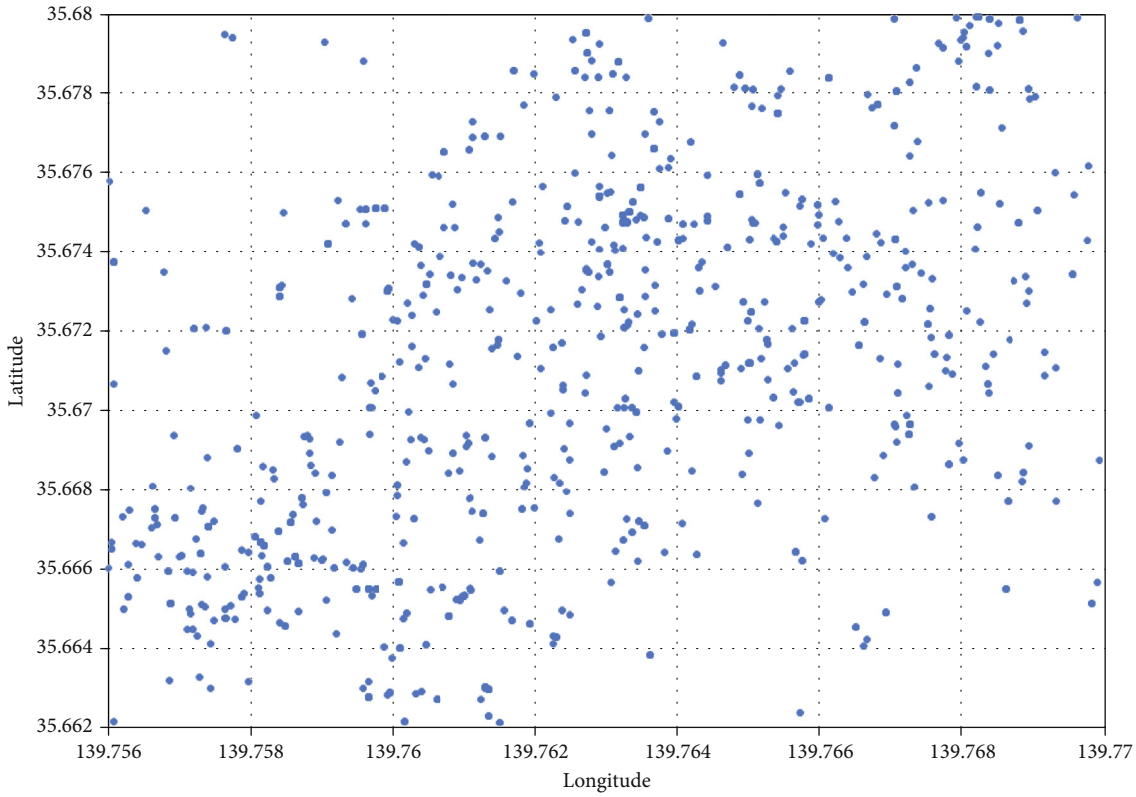

Figure 5: A real user distribution in Tokyo.

Table 4: Parameter Settings.

Parameter	Value
The bandwidth of a cell	10 MHz
The number of small cells	105
The radius of a small cell	100 meters
The transmit power of a small cell	30 dBm
The intersite distance	150 meters
The number of users	[500, 3000]
The thermal noise	-174 dBm/Hz
The overloading threshold ρ_{TH}	0.9

Figure 7 illustrates the impacts of the number of users on the satisfaction ratio. When the number of users increases, the satisfaction ratio decreases. When the number of users is 3000, the three algorithms have the same satisfaction ratio. Because the resource blocks of the small cells have been exhausted, no load balancing policy is valid. Our proposed *MDOA(1-D)* can significantly outperform *AMLB* for the satisfaction ratio because our proposed algorithm simultaneously considers the consumed resource blocks for satisfying each user's requirement and the remaining resource blocks of each neighboring cell. Therefore, *MDOA* does not exhaust the resource blocks of a specific cell and can handover more users to other cells to achieve load balance. However, *AMLB* does not consider loads of neighboring cells. Thus, under *AMLB*, the resource blocks of some specific neighboring small cells may be quickly exhausted. Compared with *MDOA(1-D)*, *MDOA(2-D)* can increase more the satisfaction ratio because 2-depth offloading can offload the users of 1-depth cells to 2-depth cells for an overloaded cell and satisfy more users' requirements when 1-depth cells are also overloaded. Compared with the two baselines, the proposed algorithm can improve the satisfaction ratio by up to 13%.

Figure 8 shows the impacts of the number of users in a hot-zone on the satisfaction ratio under the overloaded cells with 1500 users. When the number of users increases in a hot-zone, the satisfaction ratio decreases because more users in a hot-zone will consume more resource blocks. When there are more users in a hot-zone, the performance improvement under our proposed algorithm is more evident. This is because the proposed algorithm can offload more uses by simultaneously considering the QoS requirement of each user and the remaining resource blocks of neighboring cells. In other words, the proposed algorithm will avoid using up all the resource blocks of a small cell. Moreover, compared with *MDOA(1-D)*, *MDOA(2-D)* can increase more satisfaction ratio when there are more users in a hot-zone.

Figure 9 demonstrates the impacts of the number of users on the spectrum efficiency (bits/s/Hz). When the number of users increases, the spectrum efficiency under the three algorithms increases. The reason is that the three algorithms will select users connecting to the small cells with better signal qualities. Thus, when there are more users, the three algorithms have more selections of users with higher spectrum efficiencies. Compared with *Max-RSRP* and *AMLB*, our

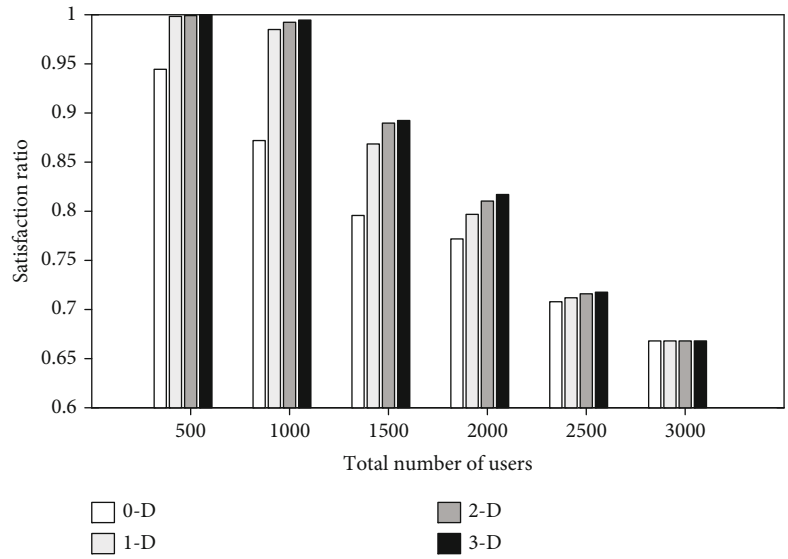

Figure 6: Satisfaction ratio under different offloading depths.

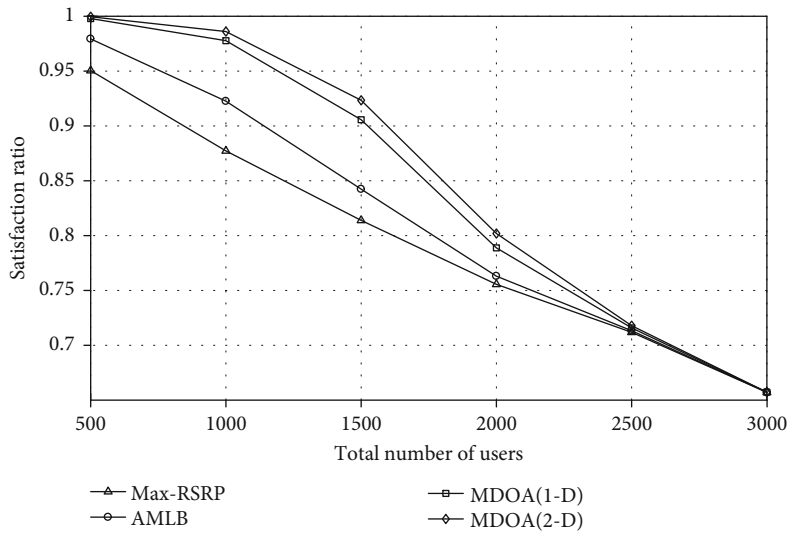

Figure 7: Impacts of the number of users on the satisfaction ratio.

proposed algorithm has lower spectrum efficiency because the proposed algorithm does not always choose the small cells with the best signal quality while balancing loads of neighboring cells. Therefore, our proposed algorithm should allocate more resource blocks to a user to satisfy the data requirement. This phenomenon also explains why the average number of remaining resource blocks under our proposed algorithm is lower than that under the two baselines, as shown in Figure 10.

Figure 11 illustrates the impacts of the number of users on the number of overloaded small cells after executing a load balancing algorithm. When the number of users increases, the number of overloaded small cells increases. This result is expected because more small cells will be overloaded by serving the users. The proposed algorithm can reduce more the number of overloaded small cells than *Max-RSRP* and *AMLB* because the proposed algorithm can offload more users to neighboring small cells if 1-depth and 2-depth cells have remaining resource blocks. Moreover, the proposed algorithm can also balance loads of neighboring cells and avoid depleting the resource blocks of a cell. Therefore, the proposed algorithm can efficiently reduce the number of overloaded small cells to increase the QoS satisfaction ratio.

Figure 12 shows the impacts of the number of users on the edge users' throughput under the clustered user placement. When the number of users increases, the edge users' throughput decreases. The result is because the limited radio resource blocks should be shared with users. In order to satisfy more data rate requirements of users, the radio resource

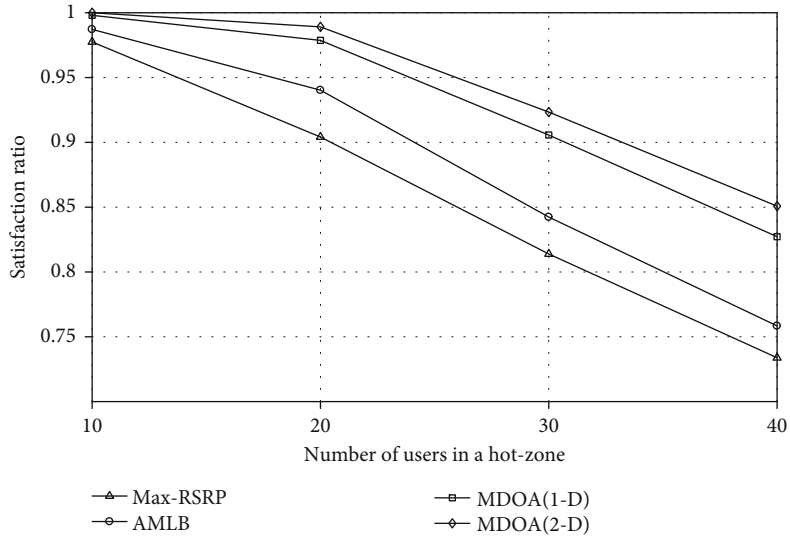

Figure 8: Impacts of the number of users in a hot-zone on the satisfaction ratio under 1500 users.

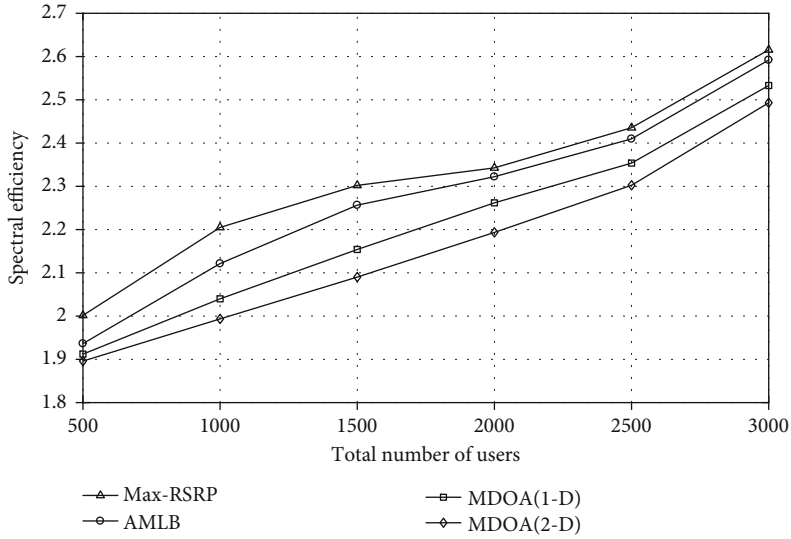

Figure 9: Impacts of the number of users on the spectrum efficiency (bits/s/Hz).

blocks will be first allocated to the users (near the serving base station) with better channel qualities, because the data rate requirements of the users can be satisfied by fewer resource blocks. When there are more users, more edge users will be sacrificed so that the edge users' throughput decreases. Compared with the two baselines, the proposed *MDOA* still can increase the edge users' throughput.

(2) *Real User Placement:* Table 5 demonstrates the edge users' throughput under the real user placement with 2000 users. Comparing Table 5 with Figure 12 at 2000 users, we can observe that the performance improvement in terms of the edge users' throughput coming from our proposed algorithm is more evident under the real user placement than under the clustered user placement. The reason is that because many cells do not serve any user under the real user distribution, the proposed concept of the multidepth offloading algorithm can offload more edge users to the cells with light loads for improving the edge users' throughput. However, the edge users' throughput under the real user placement is lower than that under the clustered user placement. Because we first arrange the hot-zone users, other users are placed with a uniform distribution so that each cell can serve some users. On the other hand, in the real user placement, some cells without serving users do not contribute the network throughput

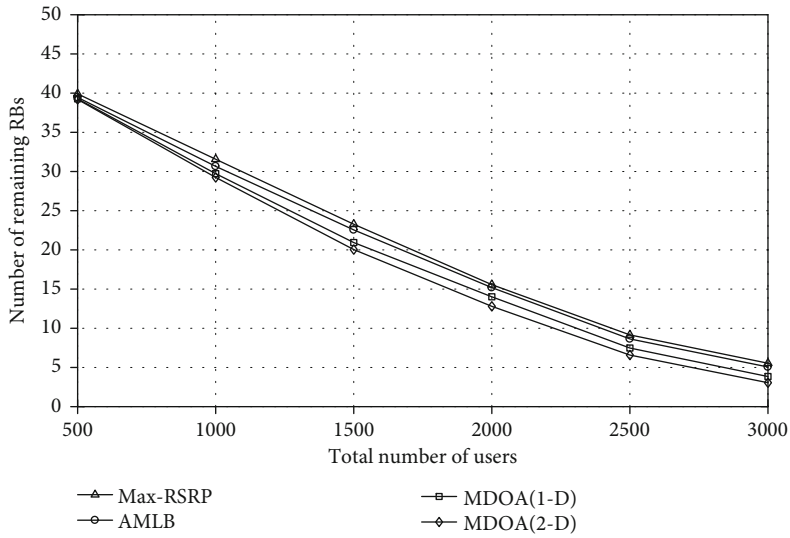

Figure 10: Impacts of the number of users on the number of remaining resource blocks.

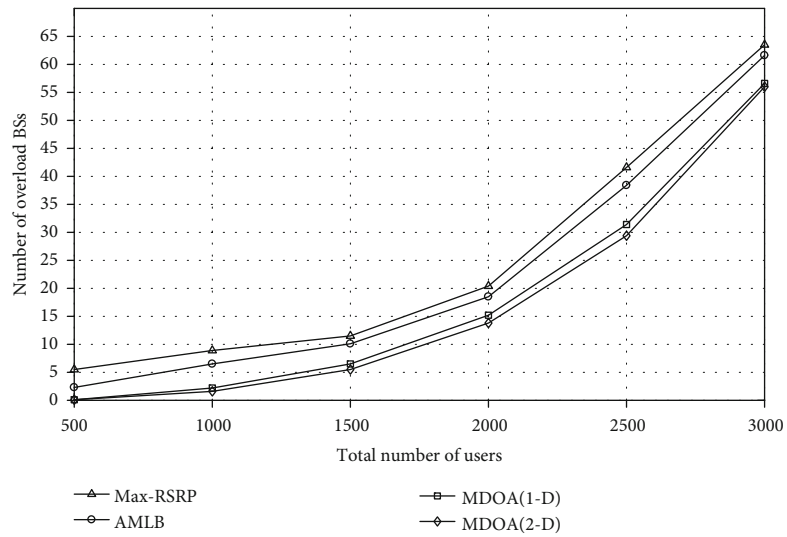

Figure 11: Impacts of the number of users on the number of overloaded cells.

Table 6 shows the edge users' satisfaction ratio under the real user placement with 2000 users. Comparing Table 6 with Figure 7 at 2000 users, the increase of the satisfaction ratio is more obvious when executing our proposed algorithm in the real user placement than in the clustered user placement. However, the satisfaction ratio under the real user placement is lower than that under the clustered user placement. The reasons are similar to the simulation result of Table 5. The two simulation results provide an insight that the proposed multidepth offloading algorithm can perform better in the real user distribution than in the clustered user placement, where loads of small cells are more unbalanced. Compared with the two baselines, the proposed algorithm can increase at most 16% satisfaction ratio in the real user distribution.

6. Conclusion

In this paper, we have studied the load balancing problem in ultradense cellular networks. The objective is to maximize the QoS satisfaction ratio under the limited radio resource blocks of each small cell. We observe that traditional offloading algorithms only considered signal qualities of neighboring cells and 1-depth cells for offloading users. The algorithms are not appropriate for the scenario that overloaded cells are clustered together when users gather together in popular locations of urban areas. The designs will quickly exhaust the resource blocks of specific small cells with better signal qualities. Therefore, this paper proposes an efficient multidepth offloading algorithm by simultaneously considering the consumed resource blocks for serving each user and

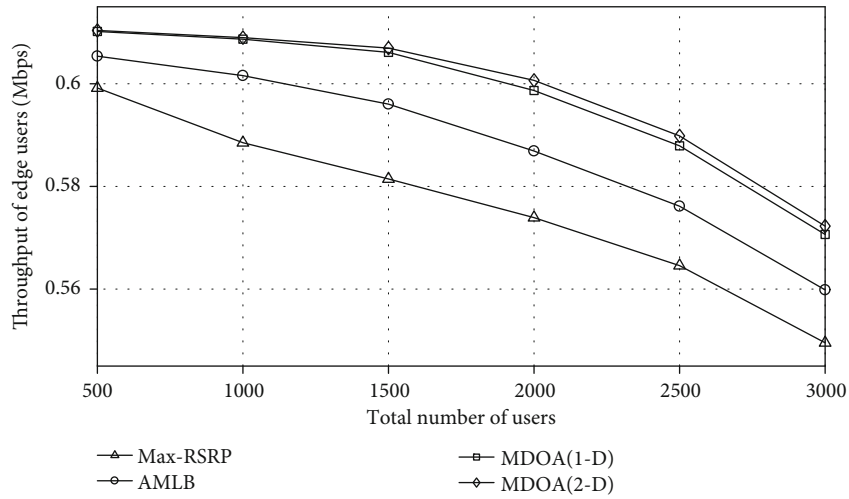

Figure 12: Impacts of the number of users on the edge users' throughput.

Table 5: Edge users' throughput.

Method	Max-RSRP	AMLB	MDOA(1-D)	MDOA(2-D)
Throughput (Mbps)	0.4	0.43	0.45	0.46

Table 6: Edge users' satisfaction ratio.

Method	Max-RSRP	AMLB	MDOA(1-D)	MDOA(2-D)
Satisfaction ratio	0.653	0.71	0.746	0.762

the remaining resource blocks of each neighboring cell. The proposed algorithm can abstain from draining the resource blocks of some specific cells and is suitable for clustered cells with full loads. Compared with a baseline and a previous offloading algorithm, the simulation results provide several insights for the load balancing algorithm design in densely deployed small cells and show that the proposed multidepth offloading algorithm can perform better when loads of small cells are more unbalanced. The results also show that our proposed algorithm can increase 16% QoS satisfaction ratio in a real user distribution and 13% QoS satisfaction ratio in a clustered user placement.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by MOST of Taiwan under contracts 108-2221-E-110-017, 109-2221E-110-049-MY2, and 107-2218-E-390-003-MY3.

References

[1] Cisco, "Cisco visual networking index: global mobile data traffic forecast update 2017-2022," 2019.

[2] H. Zhou, H. Wang, X. Li, and V. C. M. Leung, "A survey on mobile data offloading technologies," *IEEE Access*, vol. 6, no. 1, pp. 5101–5111, 2018.

[3] M. F. Khan, F. A. Bhatti, A. Habib et al., "Analysis of macro user offloading to femto cells for 5G cellular networks," in *International Symposium on Wireless Systems and Networks (ISWSN)*, Lahore, Pakistan, November 2017.

[4] 3GPP, "Further Advancements for E-UTRA Physical Layer Aspects," 2010, TR 36.814 V9.0.0.

[5] X. Ge, J. Ye, Y. Yang, and Q. Li, "User mobility evaluation for 5G small cell networks based on individual mobility model," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 528–541, 2016.

[6] K. Miyanabe, K. Suto, Z. M. Fadlullah et al., "A cloud radio access network with power over fiber toward 5G networks: QoE-guaranteed design and operation," *IEEE Wireless Communications*, vol. 22, no. 4, pp. 58–64, 2015.

[7] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[8] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "Quantifying QoS requirements of network services: a cheat-proof framework," in *ACM Conference on Multimedia Systems*, pp. 81–92, San Jose, CA, USA, February 2011.

[9] M. M. Hasan, S. Kwon, and J.-H. Na, "Adaptive mobility load balancing algorithm for LTE small-cell networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2205–2217, 2018.

[10] X. H. Chen, "Adaptive traffic-load shedding and its capacity gain in CDMA cellular systems," *IEE Proceedings - Communications*, vol. 142, no. 3, pp. 186–192, 1995.

[11] S. V. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1332–1340, 1995.

[12] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: modeling, analysis, and design insights," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, 2013.

[13] M. M. Hasan and S. Kwon, "Cluster-based load balancing algorithm for ultra-dense heterogeneous networks," *IEEE Access*, vol. 8, pp. 2153–2162, 2020.

[14] A. Laya, K. Wang, A. A. Widaa, J. Alonso-Zarate, J. Markendahl, and L. Alonso, "Device-to-device communications and small cells: enabling spectrum reuse for dense networks," *IEEE Wireless Communications*, vol. 21, no. 4, pp. 98–105, 2014.

[15] S. Farzi, S. Yousefi, J. Bagherzadeh, and B. Eslamnour, "Zone-based load balancing in two-tier heterogeneous cellular networks: a game theoretic approach," *Telecommunication Systems*, vol. 70, no. 1, pp. 105–121, 2019.

[16] K. Son, S. Chong, and G. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 7, pp. 3566–3576, 2009.

[17] Z. Li, H. Wang, Z. Pan, N. Liu, and X. You, "Dynamic load balancing in 3GPP LTE multi-cell fractional frequency reuse networks," in *IEEE Vehicular Technology Conference (VTC)*, pp. 1–5, Quebec City, QC, Canada, 2012.

[18] G. Liu and H. Zhao, "Power allocation and channel selection in small cell networks based on traffic-offloading," in *2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS)*, Harbin, China, June 2017.

[19] J. Du, E. Gelenbe, C. Jiang, H. Zhang, and Y. Ren, "Contract design for traffic offloading and resource allocation in heterogeneous ultra-dense networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2457–2467, 2017.

[20] B. C. Chung and D.-H. Cho, "Mobile data offloading with almost blank subframe in LTE-LAA and Wi-Fi coexisting networks based on coalition game," *IEEE Communications Letters*, vol. 21, no. 3, pp. 608–611, 2017.

[21] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45no. 1, pp. 129–142, 2015.

[22] A. Imran and A. Zoha, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, 2014.

[23] H. Klessig, D. Ohmann, A. I. Reppas et al., "From immune cells to self-organizing ultra-dense small cell networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 800–811, 2016.