Throughput-Enhanced Relay Placement Mechanism in WiMAX 802.16j Multihop Relay Networks

Chih-Yung Chang, *Member, IEEE*, Chao-Tsun Chang, Tzu-Chia Wang, and Ming-Hsien Li

***Abstract*—The IEEE 802.16j standard proposes a multihop relay network architecture that introduces relay stations aiming at increasing the network throughput or coverage. The deployment of the relay stations is one of the most important issues that determine the network throughput. In literature, some deployment strategies have been proposed. However, none of them follows the frame structure designed in IEEE 802.16j standard. Furthermore, they did not consider the fact that the bandwidth constraints of a base station, relay stations, and mobile stations (MSs) are highly related to the locations of relays. Given a base station, k relay stations, and a region that can be fully covered by the base station, this paper proposes a relay deployment mechanism aiming to determine the best deployed locations for relays so that the bandwidth requirement of MSs can be satisfied while the network throughput can be significantly improved. Performance study reveals that the proposed relay deployment mechanism can improve the network throughput as compared with the existing approaches.**

***Index Terms*—Deployment, IEEE 802.16j, mobile stations (MSs), relay stations (RSs), Worldwide Interoperability for Microwave Access (WiMAX).**

I. Introduction

In the IEEE 802.16e standard, a wireless metropolitan area network (WMAN) consists of a base station (BS), mobile stations (MSs), and subscriber stations (SSs). The BS serves as a gateway between the WMAN and external networks and supports MSs and relay stations (RSs) with Internet access. In literature, there have been several studies [1]–[3] developing QoS scheduling frameworks for IEEE 802.16e networks, aiming at improving the network throughput. However, the network performance still highly depends on the network deployment. In considering the deployment issue in IEEE 802.16e networks, the cost is expensive for achieving full coverage by deploying only BSs. Therefore, the relay station (RS) interconnected between the BS and MSs (or SSs) for WMANs is proposed in the new version of IEEE 802.16j standard [4]. The RSs are applied to extend the coverage range of a WMAN and hence to reduce the cost of infrastructure construction. Another, the RSs interconnecting the BS and MSs (or SSs) can reduce the communication distance and therefore improve the network capacity by applying efficient modulations.

In the 802.16j networks, the BS is in charge of bandwidth assignment [5], [6] and routing [7] tasks for RSs, MSs, and SSs. The standard of 802.16j defines two types of RSs, including transparent and nontransparent RSs. A transparent RS can be served as a forwarding agent to improve the transmission efficiency between the BS and MSs, but it does not participate in the scheduling task in a WMAN. The BS has to announce and handle all control messages, including dedicated channel, downlink/uplink (DL/UL) map, and DL/UL channel descriptors, for MSs [8]. This indicates that a transparent RS is only in charge of forwarding the data. On the contrary, a nontransparent RS can perform all the tasks of transparent RS plus the transmission scheduling [9]. Therefore, the MSs that are located outside the coverage area of the BS can still be served by the BS due to the help of relaying operations from the nontransparent RSs.

The deployment of the RSs is crucial for improving the network capacity of Worldwide Interoperability for Microwave Access networks. In literature, a number of studies [10]–[12] have proposed related approaches for RS placement. In [10], Bender's decomposition method is applied to break down the original problem to a sequence of small 0–1 integer problems. By applying Bender's decomposition, the proposed integer programming determines the minimal number of required RSs and the way of RS deployment. However, the frame constraint of the IEEE 802.16j standard and the bandwidth constraints between the BS and an RS and between an RS and an MS were not considered.

In [11], Yu *et al.* proved that the throughput-maximization RS deployment problem is NP-hard. To reduce the computational cost, Yu *et al.* [11] proposed another heuristic scheme for minimizing the deployment cost by randomly selecting the candidate sites based on the integer programming technique. In this scheme, some test points are treated as MSs with different bandwidth requirements, whereas BSs and RSs are considered randomly deployed at the candidate sites. The proposed deployment scheme then determines the feasible number of BSs and RSs. The mechanism proposed in [11] aims to find the feasible candidate sites for deploying the BSs and RSs with minimal deployment cost. However, the deployment cost may not be minimal because the candidate sites are randomly selected. Furthermore, the bandwidth calculation did not take into consideration the frame structure of IEEE 802.16j standard.

Manuscript received April 20, 2013; revised March 4, 2014; accepted March 19, 2014. Date of publication May 9, 2014; date of current version June 18, 2015. This work was supported by the National Science Council of the Republic of China under Contract NSC 100-2632-E-032-001-MY3 and Contract NSC 100-2221-E-007-054-MY3.

C.-Y. Chang, T.-C. Wang, and M.-H. Li are with the Department of Computer Science and Information Engineering, Tamkang University, Taipei 25137, Taiwan (e-mail: cychang@mail.tku.edu.tw; tcwang@mail.tku.edu.tw; minghsienli@gmail.com).

C.-T. Chang is with the Department of Information Management, Hsiuping University of Science and Technology, Taichung 41280, Taiwan (e-mail: cctas@mail.hust.edu.tw).

Digital Object Identifier 10.1109/JSYST.2014.2314011

In [12], Wang *et al.* aimed to maximize the network capacity by considering the network throughput and the signal strength. In their study, all RSs are regularly deployed around the BS with a specific distance where the transmission rate from an MS to the BS through an RS can be improved. The BS makes decision for either one-hop (MS–BS) or two-hop (MS–RS–BS) communications by considering their transmission rates. They assume that MSs with similar transmission requirements are evenly distributed in the network. Then, an approach is proposed to determine a distance for deploying all RSs. However, all RSs deployed in a way that they have the same distance to the BS might lead to inefficient performance. This occurs because the MSs served by different RSs might have different traffic requirements and the RSs might need to be deployed with different distances to the BS. Therefore, the network throughput can be further improved. In addition, the bandwidth measurement did not consider the frame structure of IEEE 802.16j standard.

In [13], Yu *et al.* aimed to minimize the RS deployment cost by developing an integer programming formulation to select as few as possible positions for RS deployment while the transmission rates required by the MSs are satisfied. Yang *et al.* [14] studied the RS placement problem and deployed a minimum number of RSs to satisfy all of the data rate requests from the MSs via cooperative communications. A heuristic algorithm was proposed in [15] to determine the number and locations of the RS deployments with a given MS distribution and deployment budget. Previous schemes [10]–[15] emphasized the network throughput enhancement by deploying the RSs. They investigated the impact of RS location on the data rate, aiming to determine the better locations for deploying RSs. However, they did not consider the bandwidth constraint of each RS, possibly leading to two problems: load unbalance and bandwidth starvation.

1) Without considering the bandwidth constraint of an RS, load unbalance might occur among RSs. Once an RS is deployed at a candidate position, which is closed to many MSs, the RS cannot support all the requirements of MSs, whereas the other deployed RSs might have low bandwidth utilization.
2) As aforementioned, once the RS cannot support the required bandwidth resource of some MSs, these MSs are difficult to access the wireless resource from the RS and easily encounter bandwidth starvation.

This paper investigates the relay deployment issue in the 802.16j mobile multihop relay network. A heuristic RS placement scheme is proposed to determine the k feasible locations for given k RSs. The contributions of this paper are summarized as follows.

1) This paper takes into account the frame structure of 802.16j standard. The transmission latency from an MS to a BS and from an MS to a BS through an RS can be accurately estimated, as compared with the related works [11], [12]. As a result, a better decision for deploying relays can be made based on the accurate estimation of transmission latency.
2) To cope with the aforementioned load unbalance and bandwidth starvation, this paper first divides the given region into k partitions with equal traffic demands for balancing the loads of k RSs in the partitioning phase. After that, this paper evaluates the available transmission intervals of MSs and the bandwidth constraint of RSs. As a result, the proposed approach can prevent the occurrences of load unbalance and bandwidth starvation.
3) To reduce the computational complexity, this paper proposes the bright-region identification phase and the candidate-region identification phase to identify the better locations for deploying relays. Compared with the related work [11], our approach does not require predefined candidate locations while obtaining better results in terms of QoS satisfactory.
4) This paper proposes the candidate-location identification phase, which determines the best location of each candidate region for deploying the RS. Simulation results further verifies that the proposed mechanism achieves better network throughputs.
5) The RS deployment achieved by the proposed scheme guarantees that the bandwidth requirements of MSs can be satisfied while the network capacity can be significantly improved.

The rest of this paper is organized as follows. Section II proposes the considered network model and problem formulation. The developed relay placement mechanism is proposed in Section III. Section IV investigates the performance improvement of the proposed mechanism against the existing works. Section V simply concludes this paper.

II. System Model and Problem Formulation

This paper aims to develop a relay placement scheme for enhancing the network capacity for given k relays in an IEEE 802.16j WiMAX network. A region A that can be fully covered by a BS is considered the deployment region for a given set of k RSs, $R = \{\mathrm{RS}_1, \mathrm{RS}_2, \ldots, \mathrm{RS}_k\}$. Region A is assumed comprised of n subregions $A_1, A_2, \ldots, A_n$. In reality, region A can be treated as a city that comprises n districts. The sizes and shapes of all subregions can be different. The traffic requirement of each subregion is assumed known at the network deployment phase. This assumption is reasonable since the traffic requirement can be estimated according to historical traffic of Internet usage. Without loss of generality, let P_i be the representative point of each subregion A_i and the traffic demands of subregion A_i be the traffic demand of P_i. Let $P_i(x, y)$ denote the location of P_i. The deployed location of the BS is assumed known and is denoted $\mathrm{BS}(0, 0)$.

The following explains that it is reasonable to use the traffic requirement of P_i to represent the traffic requirements of all MSs in subregion A_i. Assume that there are q users in subregion A_i. Assume that each user MS_j has traffic demand d_j and the transmission rate from MS_j to the BS is r_j, for $1 \leq j \leq q$. In order to guarantee that traffic demand d_j of user MS_j can be satisfied by applying the transmission rate r_j, the required

transmission time allocated to MS_j should be at least d_j/r_j. Let T_i denote the total required transmission time that satisfies the requirements of all users in subregion A_i. The value of T_i can be calculated by applying the following:

$$T_i = \sum_{j=1}^{q} \frac{d_j}{r_j}. \tag{1}$$

Let w_i be the total traffic demands of subregion A_i, i.e.,

$$w_i = \sum_{j=1}^{q} d_j. \tag{2}$$

Assume that there exists a point P_i located in subregion A_i. Let x_i be the transmission rate of point P_i. No matter where the P_i locates, its maximal achievable throughput during the time period T_i can be calculated as $x_i \times T_i$. If its transmission rate x_i is fast enough, the BS can allocate the transmission time T_i for P_i to achieve the same throughput w_i. If the traffic demands of P_i are satisfied, the total traffic demands of all users in subregion A_i can be also satisfied, as depicted in the following:

$$w_i = x_i \times T_i. \tag{3}$$

Next, we intend to explain that a point P_i satisfying (3) exists. Let $r_{\max}$ denote the maximal transmission rate of q MSs. Without loss of generality, assume a point P_i is picked in subregion A_i. Because the value of d_j/r_j is equal to or larger than that of $d_j/r_{\max}$, $x_i \le r_{\max}$ can be derived as follows:

$$x_i = \frac{w_i}{T_i} = \frac{w_i}{\sum_{j=1}^{q} \frac{d_j}{r_j}} \le \frac{w_i}{\sum_{j=1}^{q} \frac{d_j}{r_{\max}}}$$

$$= \frac{w_i}{(\sum_{j=1}^{q} d_j)/r_{\max}} = r_{\max}. \tag{4}$$

Let $r_{\min}$ denote the minimal transmission rate of q MSs. On the contrary, the value of d_j/r_j is equal to or less than that of $d_j/r_{\min}$; the following derives $r_{\min} \le x_i$:

$$x_i = \frac{w_i}{T_i} = \frac{w_i}{\sum_{j=1}^{q} \frac{d_j}{r_j}} \ge \frac{w_i}{\sum_{j=1}^{q} \frac{d_j}{r_{\min}}}$$

$$= \frac{w_i}{\left(\sum_{j=1}^{q} d_j\right)/r_{\min}} = r_{\min}. \tag{5}$$

Since

$$r_{\min} \le x_i \le r_{\max} \tag{6}$$

is satisfied, there exists a point P_i in subregion A_i satisfying the argument that using the traffic requirement of P_i can represent the traffic requirements of all MSs in subregion A_i.

The transmission rate between a sender and a receiver depends on the signal-to-interference-plus-noise ratio (SINR) at the receiver side, which highly depends on the distance between the sender and the receiver. IEEE 802.16 standard employs modulation and coding schemes (MCSs) for adjusting the transmission rate according to the channel condition. Generally, stations closer to the BS apply a higher MCS level. The MCS level will be decreased with the distance from the BS to the device. Let d_B^A denote the distance between A and B, where A and B might be a BS, an RS, or an MS. Table I summarizes the seven MCS levels and their corresponding transmission rates.

TABLE I
SEVEN LEVELS OF MCSS [12], [16]

Level	Modulation & Coding Rate	Received SNR (dB)	Data Rate (bit/symbol)	Distance Range (m)
1	BPSK 1/2	6.4	0.5	$7400 \le d_B^A$
2	QPSK 1/2	9.4	1	$5220 \le d_B^A \le 7399$
3	QPSK 3/4	11.2	1.5	$4250 \le d_B^A \le 5219$
4	16-QAM 1/2	16.4	2	$2320 \le d_B^A \le 4249$
5	16-QAM 3/4	18.2	3	$1900 \le d_B^A \le 2319$
6	64-QAM 2/3	22.7	4	$1120 \le d_B^A \le 1899$
7	64-QAM 3/4	24.4	4.5	$d_B^A \le 1119$

TABLE II
NOTATION LIST

u_a^b	Throughput between stations a and b. (bps)
l_a^b	1: A link is existed between stations a and b. 0: Otherwise.
r_a^b	The transmission rate between a and b. (*bps*)
$r_{P_i}^{past}$	The history traffic pattern of A_i. (*bps*)
C	The network throughputs of DL and UL per frame. (*bits/frame*)
$t_{P_i\text{-}BS}$	The latency of data transmission from P_i to BS. (s)
$t_{P_i\text{-}RS_j\text{-}BS}$	The latency of data transmission from P_i through RS_j to BS. (s)
$t_{P_i}^{UL}$	The length of UL interval of P_i in a frame. (s)
$t_{RS_j}^{UL}$	The length of UL interval of RS_j in a frame. (s)
$t_{P_i}^{DL}$	The length of DL interval of P_i in a frame. (s)
$t_{RS_j}^{DL}$	The length of DL interval of RS_j in a frame. (s)
t_{access}^{DL}	The length of DL access zone in a frame. (s)
t_{relay}^{DL}	The length of DL relay zone in a frame. (s)
t_{access}^{UL}	The length of UL access zone in a frame. (s)
t_{relay}^{UL}	The length of UL relay zone in a frame. (s)

Let r_B^A denote maximal achievable transmission rate from A to B, which highly depends on distance d_B^A. Function δ can query Table I and output the best MCS level by giving d_B^A as the input, i.e.,

$$\delta\left(d_{P_i}^{\mathrm{BS}}\right) = r_{P_i}^{\mathrm{BS}} \quad \forall i \in 1, 2, \ldots, n \tag{7}$$

$$\delta\left(d_{\mathrm{RS}_j}^{\mathrm{BS}}\right) = r_{\mathrm{RS}_j}^{\mathrm{BS}} \quad \forall j \in 1, 2, \ldots, k \tag{8}$$

$$\delta\left(d_{P_i}^{\mathrm{RS}_j}\right) = r_{P_i}^{\mathrm{RS}_j} \quad \forall i \in 1, 2, \ldots, n;\ \forall j \in 1, 2, \ldots, k. \tag{9}$$

The main purpose of this paper is to determine k RS deployment locations $L_{\mathrm{RS}} = \{\mathrm{RS}_1(x, y), \mathrm{RS}_2(x, y), \ldots, \mathrm{RS}_k(x, y)\}$ such that the network capacity can be maximized. Table II lists a set of notations that will be used to present the network model and the problem statement.

Fig. 1 depicts the frame structure developed in IEEE 802.16j standard. Each frame is comprised of *DL* and *UL subframes*. The DL subframe is further partitioned into *access zone* and

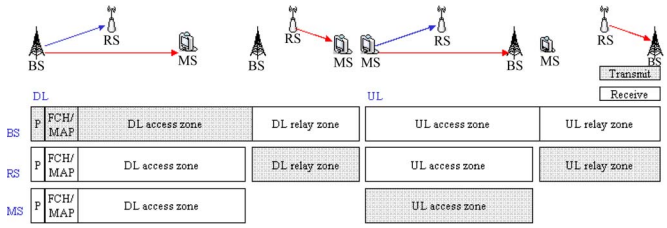


Fig. 1. Frame structure defined in IEEE 802.16j standard.

relay zone. The DL access zone can be only allocated for a BS transmitting data to RSs and MSs, whereas the relay zone can be allocated for RSs forwarding data to MSs. Similar to the DL access zone, the UL subframe is partitioned into the access zone and the relay zone.

According to the frame structure, the network throughputs of UL and DL should be calculated separately. Let notation U^{downlink} denote the MSs' DL throughput that refers to the data volume received by MSs. The DL throughput U^{downlink} is mainly composed of two parts, including the data volume directly transmitted from the BS to MSs and the data volume transmitted from RSs to representative points. The following reflects this argument:

$$\text{Downlink throughput } U^{\text{downlink}}$$

$$\sum_{i=1}^{n} u_{\text{BS}}^{P_i} + \sum_{j=1}^{k}\sum_{i=1}^{n} u_{\text{RS}_j}^{P_i} = \sum_{i=1}^{n} l_{P_i}^{\text{BS}} r_{P_i}^{\text{BS}} t_{P_i}^{\text{DL}} + \sum_{j=1}^{k}\sum_{i=1}^{n} l_{P_i}^{\text{RS}_j} r_{P_i}^{\text{RS}_j} t_{\text{RS}_j}^{\text{DL}}. \quad (10)$$

Similarly, let notation U^{uplink} denote the BS's UL throughput, which refers to the data volume that a BS actually received. The following calculates the UL throughput U^{uplink}, which is composed of two parts, i.e., the data volume of the BS directly received from representative points and the data volume of the BS received from RSs:

$$\text{Uplink throughput } U^{\text{uplink}}$$

$$\sum_{i=1}^{n} u_{P_i}^{\text{BS}} + \sum_{j=1}^{k} u_{\text{RS}_j}^{\text{BS}} = \sum_{i=1}^{n} l_{P_i}^{\text{BS}} r_{P_i}^{\text{BS}} t_{P_i}^{\text{UL}} + \sum_{j=1}^{k} l_{P_i}^{\text{RS}_j} r_{\text{RS}_j}^{\text{BS}} t_{\text{RS}_j}^{\text{UL}}. \quad (11)$$

The major goal of this paper is to maximize the network throughput. Equation (12) depicts that the network throughput C is the summation of DL and UL throughputs that are derived by (10) and (11), respectively.

Goal: Maximize (C)*:*

$$\text{where} \quad C = U^{\text{uplink}} + U^{\text{downlink}} = \left(\sum_{i=1}^{n} u_{\text{BS}}^{P_i} + \sum_{j=1}^{k}\sum_{i=1}^{n} u_{\text{RS}_j}^{P_i}\right) + \left(\sum_{i=1}^{n} u_{P_i}^{\text{BS}} + \sum_{j=1}^{k} u_{\text{RS}_j}^{\text{BS}}\right). \quad (12)$$

Equations (13)–(17) represent the constraints for achieving the goal of this paper. Equation (13), gives the *RS location constraint,* which represents the fact that none of the RSs can be deployed at the same location.

RS Location Constraint:

$$\text{RS}_i(x, y) \neq \text{RS}_j(x, y) \quad \forall i \in 1, 2, \ldots, k; \quad \forall j \in 1, 2, \ldots, k, i \neq j. \quad (13)$$

Equation (14) gives the *link constraint*, which constrains that a representative point P_i cannot communicate with both the BS and an RS.

Link Constraint:

$$l_{P_i}^{\text{RS}_j} + l_{P_i}^{\text{BS}} = 1 \quad \forall i \in 1, 2, \ldots, n; \quad \forall j \in 1, 2, \ldots, k. \quad (14)$$

The following *latency constraint* asks that the data forwarding by relay RS_j requires smaller latency than that required for directly transmission from P_i to the BS.

Latency Constraint:

$$t_{P_i\text{–RS}_j\text{–BS}} < t_{P_i\text{–BS}} = 1 \quad \forall i \in 1, 2, \ldots, n; \quad \forall j \in 1, 2, \ldots, k. \quad (15)$$

Another *minimal bandwidth constraint* is required to verify whether the assigned bandwidth satisfies the bandwidth requirement $r_{P_i}^{\text{past}}$ of P_i. Because P_i can transmit data to the BS with or without the help of RS_j, the data of P_i will be transmitted through the best way with maximal transmission rate such that the bandwidth requirement of P_i can be satisfied. In the case that relay R_i is considered, the rate of the bottleneck link would be either $r_{P_i}^{\text{RS}_j}$ or $r_{\text{RS}_j}^{\text{BS}}$. Equation (16) presents the *minimal bandwidth constraint*.

Minimal Bandwidth Constraint:

$$\max\left(r_{P_i}^{\text{BS}}, \min\left(r_{P_i}^{\text{RS}_j}, r_{\text{RS}_j}^{\text{BS}}\right)\right) \geq r_{P_i}^{\text{past}} \quad \forall i \in 1, 2, \ldots, n; \quad \forall j \in 1, 2, \ldots, k. \quad (16)$$

The last constraint given in (17) guarantees that the volume of the incoming data is equal to that of the outgoing data for any RS_j.

Transmitting and Receiving Equivalence Constraint:

$$\sum_{i=1}^{n} r_{P_i}^{\text{RS}_j} \times t_{P_i}^{\text{UL}} = r_{\text{RS}_j}^{\text{BS}} \times t_{\text{RS}_j}^{\text{UL}}, \forall i \in 1, 2, \ldots, n; \forall j \in 1, 2, \ldots, k. \quad (17)$$

As depicted in (12), the goal of this paper is to maximize the channel utilization, while constraints given in (13)–(17) should be satisfied.

III. RPM

A. Overview of RPM

The following presents the major concept of the proposed relay placement mechanism with maximal network capacity (RPM). The RPM majorly consists of four phases, called

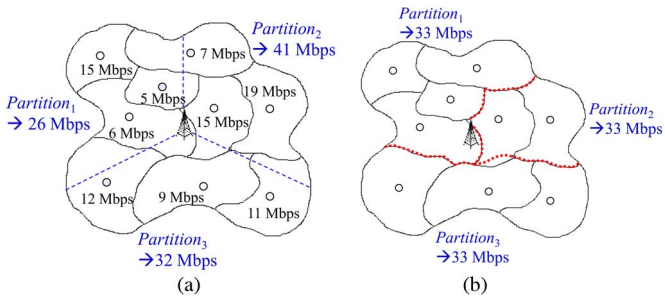


Fig. 2. Example of executing the partitioning phase for $k = 3$. (a) Initial partitions. (b) Adjusted partitions.

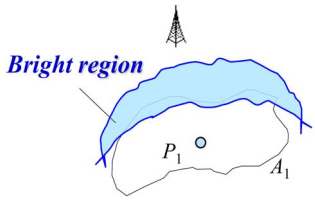


Fig. 3. *Bright region* of P_i is a promising region for RS deployment.

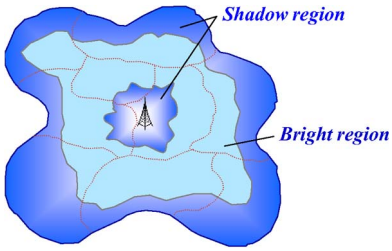


Fig. 4. Bright region that is obtained by the union of bright regions of n subregions A_i.

partitioning, *bright-region identification*, *candidate-region identification,* and *candidate-location identification*. Initially, the RPM in *partitioning phase* geographically partitions the given region A into k partitions with equal size. Then, the size of each partition will be further adjusted such that the subregion with higher traffic requirements can be migrated to the neighboring partition with lower traffic requirement; hence, the traffic demands of all partitions can be identical. Fig. 2 gives an example of $k = 3$. Initially, region A is equally partitioned into three partitions, as shown in Fig. 2(a). Then, the partitions will be further adjusted according to the traffic demand of each subregion A_i. Fig. 2(b) depicts the adjusted partitions.

Since an RS will be deployed in each partition, the next step is to find a promising region for deploying an RS in a partition. Therefore, the next three phases would be executed repeatedly for each partition. The *bright-region identification* phase aims to identify a bright region that is a promising region for deploying the RS. For subregion A_i, it is improper to deploy an RS very close to the BS or P_i. In case that the RS is deployed very close to the BS, the rate between the RS and P_i will be reduced significantly, which increases the transmission delay from the MSs to the RS. On the other hand, if the RS is deployed far away from the BS, the rate between the RS and the BS will be poor. Fig. 3 depicts that the bright region of A_1 is the promising zone for deploying the relay. Fig. 4 depicts the union of bright regions of n subregions.

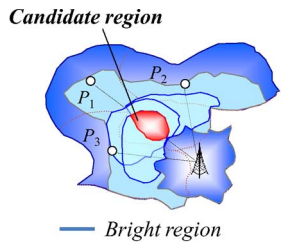


Fig. 5. Example of the candidate region that is marked by the red ink.

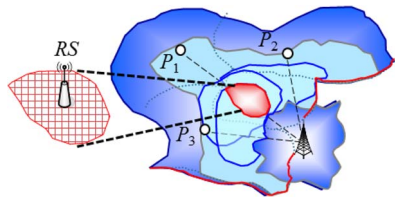


Fig. 6. Candidate region is partitioned into several grids and a location will be selected from the grids for deploying the RS.

The next phase, which is called the *candidate-region identification* phase, aims to further identify a candidate region that is a subregion of a bright region and covered by the maximal number of bright regions. The execution of the *candidate-region identification* phase can further reduce the computational cost for solving the relay deployment problem. Fig. 5 gives an example of the candidate region, which is marked by the red boundary.

It is worth mentioning that deploying the RS at any point of the bright region can satisfy the traffic demand of each A_i. However, there exists the best location that deploying the RS at this location can lead to maximal network capacity. The last phase, namely, the *candidate-location identification* phase, aims to find the best location for deploying the RS such that the network capacity can be maximal. In this phase, the candidate region will be partitioned into several small and equal-sized grids. The proposed *RPM* will calculate the network capacity grid by grid and identify the best location for deploying the RS. The grid size can be determined according to the constraint of computational cost. A larger grid will reduce the computational cost but will find a location with lower network capacity. Fig. 6 gives an example where the candidate region is partitioned into several grids and a location will be identified for deploying the relay for achieving the maximal network capacity.

B. Proposed RPM

Here, the details of the four phases designed in the proposed RPM are presented.

1) Partitioning Phase: Given k RSs, this phase aims to divide the given region A into k partitions with equal traffic demands for balancing the loads of RSs. Two steps will be executed in this phase. Recall that the region A can be fully covered by the BS, which is located at the central point of A. Region A consists of n subregions $A_1, A_2, \ldots, A_n$. Subregion A_i can be considered the district belonged to region A. For example, region A is a city, and each A_i denotes a district of this city. First, region A is equally partitioned into k partitions

$\zeta_1, \zeta_2 \ldots, \zeta_k$. The location of the BS will be treated as the central point of a circle containing region A. The circle will be divided into k partitions by k partitioning lines, so that the angle between every two consecutive lines will be $360/k^\circ$. Fig. 2(a) depicts an example of dividing the region into three partitions. In this step, each subregion A_i can be only included in exactly one partition. In case that a subregion A_i is separated by a partitioning line, subregion A_i belongs to the partition that contains the representative point P_i of subregion A_i. As a result, all subregions can be well divided into k partitions.

The second step is to adjust the size of each partition so that all partitions have equal traffic demands. The total traffic demands $u(\zeta_j)$ of the partition ζ_j will be calculated as follows:

$$u(\zeta_j) = \sum_{\forall P_i \in \zeta_j} r_{P_i}^{\text{past}}. \tag{18}$$

This step aims to adjust the partitioning line such that the total traffic demands of all partitions are equal. The adjustment for each partitioning line will be executed sequentially from ζ_1 to ζ_k along a clockwise direction. In case that the total traffic demands of ζ_j exceed the average traffic demand, it will release some subregions that are close to the partition ζ_{j+1}. Otherwise, ζ_j obtains some subregions from ζ_{j+1} so that its total traffic demands approach to the average traffic demand. Fig. 2(b) depicts an example of the partitions after applying the adjustment step.

2) Bright-Region Identification Phase: This phase aims to identify a bright region that is a promising region for deploying an RS. Herein, a location in region A is said to be a bright point if an RS deployed at this location can improve the transmission rate between the BS and P_i. Otherwise, the point is said to be a shadow point. The bright and shadow regions are the collection of all bright and shadow points in region A, respectively.

Since a bright region contains infinite bright points, it is impossible to obtain the bright region by considering every point in region A. To cope with this problem, the presentation of this phase is divided into two parts. The first part aims to present the latency evaluation approach, which takes frame structure into consideration. That is, given a location h, this part proposes a decision function that makes decision if it is favorable to deploy an RS at location h. The second part further presents *bright-region identification mechanism*, which constructs a bright region for deploying an RS. The key concept of the mechanism is to utilize the latency evaluation approach to make decision whether a given MCS combination is appreciate when the MCS combination is applied on the links l_h^{BS} and $l_{P_i}^h$.

a) Latency evaluation by considering frame structure: Given location h, this part helps make a decision whether deploying an RS at location h can reduce the latency of transmission directly from P_i to the BS. Compared with the existing deployment schemes [10]–[12], the proposed *latency evaluation by considering frame structure* takes into consideration the IEEE 802.16j frame structure.

In order to improve the network throughput through RS deployment, the *latency constraint* depicted in (15) should be satisfied. However, the evaluation of transmission latency is more complicated since the frame structure designed in IEEE 802.16j standard should be taken into consideration. The length of UL interval allocated to P_i is proportional to the traffic demand of P_i. Let the traffic demand of P_i be u_j. The transmission interval $t_{P_i}^{\text{UL}}$ allocated to P_i during each UL access zone can be derived by

$$t_{P_i}^{\text{UL}} = u_i \times t_{\text{access}}^{\text{UL}} / \sum_{j=1}^{n} u_j. \tag{19}$$

The evaluation of the transmission latency can be further considered in two cases, depending on whether the traffic is forwarded by the RS. For the ease of transmission latency calculation, it is assumed that the transmission latency is counted from the start of UL access zone. First, we consider that the data are directly transmitted from P_i to the BS. In case that the data can be completely transmitted within one frame, the latency will be equal to the transmission time. Otherwise, the transmission of u_j requires several frames. Let t_{ttg} and t_{rtg} denote the duration of transmit/receive transition gap and the duration of receive/transmit transition gap, respectively. Let t_{frame} denote the duration of one IEEE 802.16j frame. The duration of t_{frame} can be represented as follows:

$$t_{\text{frame}} = t_{\text{access}}^{\text{UL}_{k\text{th}}} + t_{\text{relay}}^{\text{UL}_{k\text{th}}} + t_{\text{ttg}} + t_{\text{access}}^{\text{DL}_{(k+1)\text{th}}} + t_{\text{relay}}^{\text{DL}_{(k+1)\text{th}}} + t_{\text{rtg}}. \tag{20}$$

Let $f_{P_i}^{\text{BS}} + 1$ denote the number of frames required for P_i to transmit data volume u_j from itself to the BS. The value of $f_{P_i}^{\text{BS}}$ can be calculated as

$$f_{P_i}^{\text{BS}} = \left\lfloor \frac{u_i}{r_{P_i}^{\text{BS}} \times t_{P_i}^{\text{UL}}} \right\rfloor. \tag{21}$$

Let $t_{P_i-\text{BS}}^{\text{last\_}f}$ denote the transmission time of the remaining traffic in the $(f_{P_i}^{\text{BS}} + 1)$th frame. Based on (21), the value of $t_{P_i-\text{BS}}^{\text{last\_}f}$ can be evaluated by

$$t_{P_i-\text{BS}}^{\text{last\_}f} = \frac{u_i - r_{P_i}^{\text{BS}} \times \left(t_{P_i}^{\text{UL}} \times f_{P_i}^{\text{BS}}\right)}{r_{P_i}^{\text{BS}}}. \tag{22}$$

According to (20)–(22), the transmission latency $t_{P_i-\text{BS}}$ will be the time duration of $f_{P_i}^{\text{BS}}$ frames plus the transmission time $t_{P_i-\text{BS}}^{\text{last\_}f}$ of the remaining traffic. Equation (23) depicts the evaluation of the transmission latency $t_{P_i-\text{BS}}$ when the value of $t_{P_i-\text{BS}}^{\text{last\_}f}$ is larger than zero. In case that the value of $t_{P_i-\text{BS}}^{\text{last\_}f}$ is equal to zero, it indicates that P_i completes its data transmission at the end of $t_{P_i}^{\text{UL}}$ in the $(f_{P_i}^{\text{BS}})$th frames. Therefore, the transmission latency is the time duration of the $f_{P_i}^{\text{BS}} - 1$ frames plus the transmission time.

$$t_{P_i-\text{BS}} = \begin{cases} f_{P_i}^{\text{BS}} \times t_{\text{frame}} + t_{P_i-\text{BS}}^{\text{last\_}f}, & \text{if } t_{P_i-\text{BS}}^{\text{last\_}f} > 0 \\ \left(f_{P_i}^{\text{BS}} - 1\right) \times t_{\text{frame}} + t_{P_i}^{\text{UL}}, & \text{if } t_{P_i-\text{BS}}^{\text{last\_}f} = 0. \end{cases} \tag{23}$$

Figs. 7 and 8 illustrate the case that the data transmission from P_i to the BS cannot be completed during a single frame. Let one frame consist of 256 slots. As shown in Fig. 7, the

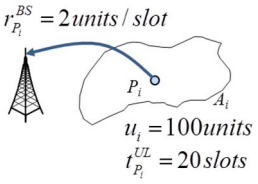


Fig. 7. Example of data transmission from P_i to the BS.

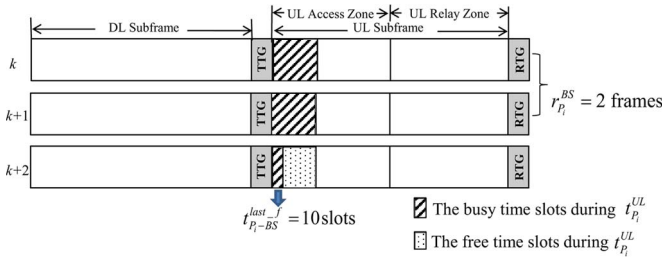


Fig. 8. Calculation of transmission latency of the example given in Fig. 7.

traffic demand u_i and the length of $t_{P_i}^{\mathrm{UL}}$are 100 units and 20 slots, respectively. Let transmission rate $r_{P_i}^{\mathrm{BS}}$ between P_i and the BS be 2 units/slot. Fig. 8 shows that P_i cannot complete its transmission in the single frame. By applying (21), the value of $f_{P_i}^{\mathrm{BS}}$ can be obtained as follows:

$$f_{P_i}^{\mathrm{BS}} = \left\lfloor \frac{100}{2 \times 20} \right\rfloor = 2 \text{ frames}.$$

By applying (22), the value of $t_{P_i-\mathrm{BS}}^{\mathrm{last\_}f}$ can be calculated as follows:

$$t_{P_i-\mathrm{BS}}^{\mathrm{last\_}f} = \frac{100 - 2 \times (20 \times 2)}{2} = 10 \text{ slots}.$$

Therefore, the total traffic demand of P_i will be completed within three frames and the total latency $t_{P_i-\mathrm{BS}}$ is $522(256 \times 2 + 10)$ slots.

Alternatively, the following considers the case that P_i transmits its data to the BS through RS_j, which is deployed at a given location h_i. Let there be k RSs connecting to the BS. Let the interval of UL relay zone $t_{\mathrm{relay}}^{\mathrm{UL}}$ be equally partitioned into k segments, and each segment is allocated to one RS. Therefore, the available transmission interval $t_{\mathrm{RS}_j}^{\mathrm{UL}}$ of RS_j can be calculated by

$$t_{\mathrm{RS}_j}^{\mathrm{UL}} = \frac{t_{\mathrm{relay}}^{\mathrm{UL}}}{k}. \tag{24}$$

With the help of RSs, time period $t_{P_i}^{\mathrm{UL}}$ given in (19) can be allocated for transmitting data from P_i to RS_j. Afterward, RS_j will forward the received data to the BS during the time period $t_{\mathrm{RS}_j}^{\mathrm{UL}}$. For the ease of calculation, as shown in Fig. 9, assume that transmission time $t_{P_i}^{\mathrm{UL}}$ and forwarding transmission time $t_{\mathrm{RS}_j}^{\mathrm{UL}}$ are allocated at the end of the UL access zone and in the beginning of the UL relay zone, respectively.

Since RS_j is deployed at the given location h_i, the rates $r_{P_i}^{\mathrm{RS}_j}$ from P_i to RS_j and $r_{\mathrm{RS}_j}^{\mathrm{BS}}$ from RS_j to BS can be derived by (8) and (9), respectively.

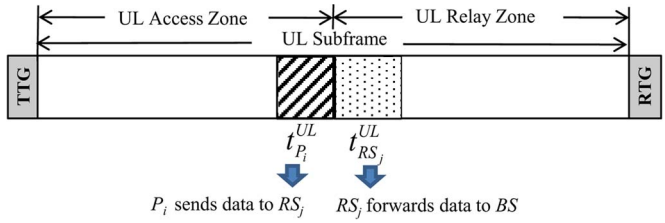


Fig. 9. Example of data transmissions from P_i to RS_j, and then the data are forwarded from RS_j to BS.

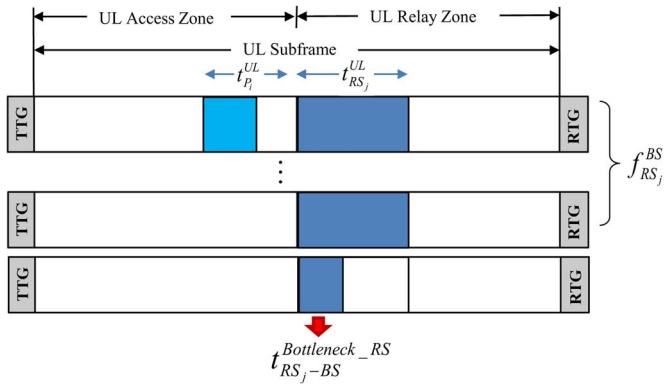


Fig. 10. Link capacity of $l(\mathrm{RS}_j, \mathrm{BS})$ is the bottleneck of the path from P_i to the BS through RS_j.

Let $f_{P_i}^{\mathrm{RS}_j} + 1$ denote the number of frames required for data volume u_i transmitted from P_i to RS_j. The value of $f_{P_i}^{\mathrm{RS}_j}$ can be calculated as follows:

$$f_{P_i}^{\mathrm{RS}_j} = \left\lfloor \frac{u_i}{r_{P_i}^{\mathrm{RS}_j} \times t_{P_i}^{\mathrm{UL}}} \right\rfloor. \tag{25}$$

Similarly, let $f_{\mathrm{RS}_j}^{\mathrm{BS}} + 1$ denote the number of frames required for RS_j to forward the received data volume u_i to BS. The following calculates the value of $f_{\mathrm{RS}_j}^{\mathrm{BS}}$:

$$f_{\mathrm{RS}_j}^{\mathrm{BS}} = \left\lfloor \frac{u_i}{r_{\mathrm{RS}_j}^{\mathrm{BS}} \times t_{\mathrm{RS}_j}^{\mathrm{UL}}} \right\rfloor. \tag{26}$$

Let $l(a, b)$ denote the link between nodes a and b. For the data transmission from P_i to the BS through RS_j, the capacity bottleneck link is either $l(P_i, \mathrm{RS}_j)$ or $l(\mathrm{RS}_j, \mathrm{BS})$. The following discusses the calculations of the transmission latency $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ in two cases.

The first case is that the value of $f_{\mathrm{RS}_j}^{\mathrm{BS}}$ is larger than that of $f_{P_i}^{\mathrm{RS}_j}$. It indicates that link capacity of $l(\mathrm{RS}_j, \mathrm{BS})$ is the bottleneck link and the traffic demand of u_i will be completed in the $(f_{\mathrm{RS}_j}^{\mathrm{BS}} + 1)$th frame. Let $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}$ denote the transmission time of the remaining traffic of RS_j in the $(f_{\mathrm{RS}_j}^{\mathrm{BS}} + 1)$th frame. The following depicts the calculation of $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}$:

$$t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}} = \frac{u_i - r_{\mathrm{RS}_j}^{\mathrm{BS}} \times \left(t_{\mathrm{RS}_j}^{\mathrm{UL}} \times f_{\mathrm{RS}_j}^{\mathrm{BS}}\right)}{r_{\mathrm{RS}_j}^{\mathrm{BS}}}, \quad \text{if } f_{P_i}^{\mathrm{RS}_j} < f_{\mathrm{RS}_j}^{\mathrm{BS}}. \tag{27}$$

Fig. 10 gives an example that link $l(\mathrm{RS}_j, \mathrm{BS})$ is the bottleneck link. As shown in Fig. 10, P_i can transmits all of its data

to RS_j within the allocated interval $t_{P_i}^{\mathrm{UL}}$. However, RS_j cannot complete the data forwarding to the BS within $t_{\mathrm{RS}_j}^{\mathrm{UL}}$. Therefore, RS_j needs $f_{\mathrm{RS}_j}^{\mathrm{BS}}$ frames plus transmission time $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}$ to complete the data forwarding. The following depicts the transmission latency $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ when $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}$ is larger than zero:

$$t_{P_i-\mathrm{RS}-\mathrm{BS}} = \begin{cases} t_{P_i}^{\mathrm{UL}} + f_{\mathrm{RS}_j}^{\mathrm{BS}} \times t_{\mathrm{frame}} + t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}, \\ \quad \text{if } f_{P_i}^{\mathrm{RS}_j} < f_{\mathrm{RS}_j}^{\mathrm{BS}}, t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}} > 0 \\ t_{P_i}^{\mathrm{UL}} + \left(f_{\mathrm{RS}_j}^{\mathrm{BS}} - 1\right) \times t_{\mathrm{frame}} + t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}, \\ \quad \text{if } f_{P_i}^{\mathrm{RS}_j} < f_{\mathrm{RS}_j}^{\mathrm{BS}}, t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}} = 0. \end{cases} \tag{28}$$

Alternatively, (28) also depicts evaluation of $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ when the value of $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}$ is equal to zero. Similar to the concept of (23), in the case that the value of $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}$ is equal to zero, the data transmission of P_i can be completed at the end of $t_{\mathrm{RS}_j}^{\mathrm{UL}}$ in the $f_{\mathrm{RS}_j}^{\mathrm{BS}}$th frame. Therefore, the transmission latency is the time duration of the $f_{\mathrm{RS}_j}^{\mathrm{BS}} - 1$ frames plus the transmission time $t_{\mathrm{RS}_j}^{\mathrm{UL}}$. Equation (28) depicts the evaluation of the total transmission latency $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ for the case that the link capacity of $l(\mathrm{RS}_j, \mathrm{BS})$ is the bottleneck.

Alternatively, in the case that the value of $f_{P_i}^{\mathrm{RS}_j}$ is larger than or equal to that of $f_{\mathrm{RS}_j}^{\mathrm{BS}}$, the link capacity of $l(P_i, \mathrm{RS}_j)$ is the bottleneck of the data transmission from P_i to the BS. The traffic demand of P_i will be completed in the $(f_{P_i}^{\mathrm{RS}_j} + 1)$th frame. Let $t_{P_i-\mathrm{RS}_j}^{\mathrm{last\_}f}$ denote the transmission time of the remaining traffic transmitted from P_i to RS_j in the $(f_{P_i}^{\mathrm{RS}_j} + 1)$th frame. The following depicts the value of $t_{P_i-\mathrm{RS}_j}^{\mathrm{last\_}f}$:

$$t_{P_i-\mathrm{RS}_j}^{\mathrm{last\_}f} = \frac{u_i - r_{P_i}^{\mathrm{RS}_j} \times \left(t_{P_i}^{\mathrm{UL}} \times f_{P_i}^{\mathrm{RS}_j}\right)}{r_{P_i}^{\mathrm{RS}_j}}. \tag{29}$$

Although link capacity of $l(\mathrm{RS}_j, \mathrm{BS})$ is not the bottleneck link, RS_j needs to forward data that are received from P_i in the access zone of each frame, containing the $(f_{P_i}^{\mathrm{RS}_j} + 1)$th frame. Let $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_}P}$ denote the transmission time of RS_j's data forwarding in the $(f_{P_i}^{\mathrm{RS}_j} + 1)$th frame. Different from the transmission time depicted in (27), if the value of $f_{P_i}^{\mathrm{RS}_j}$ is larger than or equal to that of $f_{\mathrm{RS}_j}^{\mathrm{BS}}$, then the value of $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_}P}$ is evaluated by the following:

$$t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_}P} = \frac{u_i - r_{P_i}^{\mathrm{RS}_j} \times \left(t_{P_i}^{\mathrm{UL}} \times f_{P_i}^{\mathrm{RS}_j}\right)}{r_{\mathrm{RS}_j}^{\mathrm{BS}}}, \text{ if } f_{P_i}^{\mathrm{RS}_j} \geq f_{\mathrm{RS}_j}^{\mathrm{BS}}. \tag{30}$$

Consequently, the total transmission latency $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ of this case can be evaluated as follows:

$$t_{P_i-\mathrm{RS}-\mathrm{BS}} = \begin{cases} t_{P_i}^{\mathrm{UL}} + f_{P_i}^{\mathrm{RS}_j} \times t_{\mathrm{frame}} + t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_}P}, \\ \quad \text{if } f_{P_i}^{\mathrm{RS}_j} \geq f_{\mathrm{RS}_j}^{\mathrm{BS}}, t_{P_i-\mathrm{RS}_j}^{\mathrm{last\_}f} > 0 \\ t_{P_i}^{\mathrm{UL}} + \left(f_{P_i}^{\mathrm{RS}_j} - 1\right) \times t_{\mathrm{frame}} + t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_}P}, \\ \quad \text{if } f_{P_i}^{\mathrm{RS}_j} \geq f_{\mathrm{RS}_j}^{\mathrm{BS}}, t_{P_i-\mathrm{RS}_j}^{\mathrm{last\_}f} = 0. \end{cases} \tag{31}$$

Let the Boolean variable $x_{P_i}^{\mathrm{access}}$ represent whether the $f_{P_i}^{\mathrm{RS}_j}$ is no less than $f_{\mathrm{RS}_j}^{\mathrm{BS}}$, i.e.,

$$x_{P_i}^{\mathrm{access}} = \begin{cases} 1, & \text{if } f_{P_i}^{\mathrm{RS}_j} \geq f_{\mathrm{RS}_j}^{\mathrm{BS}} \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

Let $x_{\mathrm{RS}_j}^{\mathrm{relay}}$ represent the logical negation value of $x_{P_i}^{\mathrm{access}}$. Furthermore, let Boolean variable x^{nonzero} represent whether the values of $t_{P_i-\mathrm{RS}_j}^{\mathrm{last\_}f}$ or $t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}$ are not equal to zero, i.e.,

$$x^{\mathrm{nonzero}} = \begin{cases} 1, & \text{if } t_{P_i-\mathrm{RS}_j}^{\mathrm{last\_f}} \neq 0 \text{ or } t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{33}$$

Based on (28)–(34), shown at the bottom of the page, depicts the derivation of transmission latency $t_{P_i-\mathrm{RS}-\mathrm{BS}}$.

For a given location h_i, the latency $t_{P_i-\mathrm{BS}}$ and $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ can be evaluated by applying (23) and (34), respectively. As a result, constraint (15) can be applied as a decision function to verify whether the latency constraint is satisfied. If yes, the given location can be identified as a bright point.

As aforementioned, although (23) and (34) can verify whether a given location is bright point, however, it is not feasible to take into consideration every possible location in the serving region A because region A contains infinite points. The following presents the *bright-region identification mechanism,* which can efficiently construct a bright region of P_i.

b) Bright-region identification mechanism: The *bright-region identification mechanism* aims to efficiently identify a bright region for deploying the relay. The key concept of the mechanism is to utilize the latency evaluation approach to make a decision whether a given MCS combination is appreciate when the MCS combination is applied on links $l_h^{\mathrm{BS}}$and $l_{P_i}^h$. The proposed mechanism aims to identify the bright regions with reasonable computing complexity.

According to the distance depicted in Table I, a given MCS (e.g., Modulation $= 64$ QAM, Cording rate $= 1/2$) can be applied to the MSs in a particular area. Let M be the number of predefined MCS levels in IEEE 802.16j standard. There are M possible MCS levels applied to links $l(\mathrm{BS}, \mathrm{RS}_j)$ or $l(\mathrm{RS}_j, P_i)$. Therefore, there are $M \times M$ combinations of MCS levels that

$$t_{P_i-\mathrm{RS}-\mathrm{BS}} = \begin{cases} \max\left(f_{P_i}^{\mathrm{RS}_j}, f_{\mathrm{RS}_j}^{\mathrm{BS}}\right) \times t_{\mathrm{frame}} + t_{P_i}^{\mathrm{UL}} + x_{P_i}^{\mathrm{access}} \times t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_}P} + x_{\mathrm{RS}_j}^{\mathrm{relay}} \times t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_RS}}, & \text{if } x^{\mathrm{nonzero}} = 1 \\ \left(\max\left(f_{P_i}^{\mathrm{RS}_j}, f_{\mathrm{RS}_j}^{\mathrm{BS}}\right) - 1\right) \times t_{\mathrm{frame}} + t_{P_i}^{\mathrm{UL}} + x_{P_i}^{\mathrm{access}} \times t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{Bottleneck\_}P} + x_{\mathrm{RS}_j}^{\mathrm{relay}} \times t_{\mathrm{RS}_j-\mathrm{BS}}^{\mathrm{UL}}, & \text{if } x^{\mathrm{nonzero}} = 0 \end{cases} \tag{34}$$

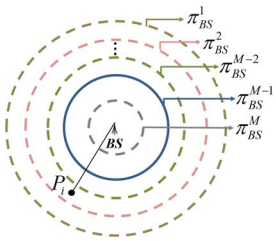


Fig. 11. Serving area of BS is partitioned into M concentric circles.

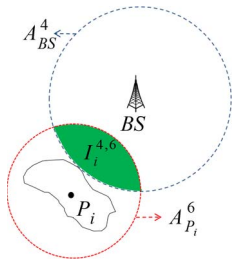


Fig. 12. Example of the intersection area $I_i^{4,6}$ of A_{BS}^4 and $A_{P_i}^6$.

can be applied to links $l(\mathrm{BS},\ \mathrm{RS}_j)$ and $l(\mathrm{RS}_j,\ P_i)$. Two steps will be executed in the proposed mechanism. First, considering $M \times M$ combinations of MCS levels, the mechanism explores the feasible combinations of MCS levels, which satisfy the latency constraint as proposed in (15). Then, the second step constructs the bright region of P_i based on the feasible combinations of MCS levels that are derived in step 1.

In the first step, given a combination of MCS levels, the corresponding transmission rates of links $l(\mathrm{BS},\ \mathrm{RS}_j)$ or $l(\mathrm{RS}_j,\ P_i)$ can be obtained from Table I. Let x and y denote two MCS levels that can be applied to $l(\mathrm{BS},\ \mathrm{RS}_j)$ and $l(\mathrm{RS}_j,\ P_i)$, respectively. Let notation m_{A-B}^x denote the modulations and coding rates of MCS level x applied on the link between stations A and B. Therefore, the MCS pair $(m_{\mathrm{BS}-\mathrm{RS}_j}^x,\ m_{\mathrm{RS}_j-P_i}^y)$ represents the pair of two modulations (with coding). Let rate pair $r_i^{xy} = (r_{\mathrm{RS}_j}^{\mathrm{BS}},\ r_{P_i}^{\mathrm{RS}_j})$ denote the corresponding rates of MCS pair $(m_{\mathrm{BS}-\mathrm{RS}_j}^x,\ m_{\mathrm{RS}_j-P_i}^y)$. For each rate pair, the transmission latency $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ can be derived by applying (34). In case the transmission latency $t_{P_i-\mathrm{RS}-\mathrm{BS}}$ satisfies the latency constraint (15), the MCS pair $(x,\ y)$ is a feasible MCS pair. By applying the aforementioned calculations, all feasible combinations of MCS levels that satisfy the latency constraint (15) can be obtained. Let Ψ_i denote the set of all feasible MCS pairs, i.e.,

$$\Psi_i = \left\{ (m_{\mathrm{BS}-\mathrm{RS}_j}^x,\ m_{\mathrm{RS}_j-P_i}^y) | r_i^{xy} \text{ satisfies (15)} \ \forall\, x,\ y \in M \right\}. \tag{35}$$

According to the feasible MCS set Ψ, the second step aims to further construct the bright region of P_i. Recall that each MCS level corresponds to a particular area where any MS in this area can apply the MCS level to successfully transmit data to the BS. For each $(m_{\mathrm{BS}-\mathrm{RS}_j}^x,\ m_{\mathrm{RS}_j-P_i}^y)$ in feasible MCS set Ψ, the second step aims to construct the intersection region of the particular areas of MCS levels x and y. Since pair x and y are feasible MCS levels that can be applied to links $l(\mathrm{BS},\ \mathrm{RS}_j)$ and $l(\mathrm{RS}_j,\ P_i)$, respectively, the intersection region can be treated as a bright region. The following further presents the second step in detail.

The second step aims to further construct the bright region of P_i. As shown in Fig. 11, the serving region of BS can be partitioned into a set of M concentric circles $\Pi_{BS} = \{\pi_{BS}^1, \pi_{BS}^2, \ldots, \pi_{BS}^M\}$ according to the M levels of MCS as shown in Table I. Let $d_{\max}^x$ denote the maximal transmission distance of the MCS level x (for example, $d(m_{\max}^6) = 1899$) where the SINR value at a given receiver reaches to the lower bound of SINR value of MCS level x. Let π_{BS}^x be the concentric circle with radius $d(m_{\max}^x)$. Let A_{BS}^x denote circle region of π_{BS}^x. Obviously, A_{BS}^x is totally covered by A_{BS}^{x-1}. Similarly, the area of P_i's communication range can be partitioned into M concentric circles $\Pi_{P_i} = \{\pi_{P_i}^1, \pi_{P_i}^2, \ldots, \pi_{P_i}^M\}$. Let $A_{P_i}^x$ denote circle region of $\pi_{P_i}^x$. According to a given feasible MCS pair $(m_{BS-RS_j}^x, m_{RS_j-P_i}^y)$, an *intersection area* $I_i^{x,y}$ of A_{BS}^x and $A_{P_i}^y$ can be constructed. As shown in Fig. 12, the area marked with green color gives an example of the *intersection area* $A_{BS}^4 \cap A_{P_i}^6$ between the BS and P_i, which is constructed based on the feasible MCS pair (16QAM 1/2, 64QAM 2/3).

Let BR_i denote the bright region of P_i. Let the notation $|I_i^{x,y}|$ denote the size of the intersection area $I_i^{x,y}$. Because each feasible MCS pair contributes an intersection area, the area $I_i^{x,y}$ with maximal size will be the bright region of P_i which can be derived by

$$BR_i = \underset{\forall\, m_x, m_y \in \Psi_i}{\arg\max} \{|I_i^{x,y}|\}. \tag{36}$$

Based on the *Latency Criteria*, the bright region can be constructed by applying Exps. (19)–(36). The *bright-region* guarantees that deploying a relay at any point in the bright region can improve both the transmission rate and latency.

3) Candidate-Region Identification Phase: This phase aims to identify a smaller region from the bright region for further reducing the computational complexity. Two or more bright regions might be overlapped. If the RS is deployed in the overlapped regions, it can improve the transmission rate of these regions. In this phase, the overlapped region, also referred to as the *candidate region*, that is covered by the maximal number of bright regions will be selected as the candidate region for deploying an RS. As shown in Fig. 5, the candidate region is marked by red ink. Let O_Φ denote an *overlapped region* constructed by a set of *bright region*s $\mathrm{BR}_1, \mathrm{BR}_2, \ldots, \mathrm{BR}_x$, where $\Phi = \{1, 2, \ldots, x\}$. Let notation $|O_\Phi|$ denote the *relay benefit* that represents the reduced transmission time if an RS is deployed in O_Φ. The following evaluates the value of $|O_\Phi|$ by calculating the sum of relay benefit:

$$|O_\Phi| = \sum_{\forall\, i \in \Phi} (t_{P_i-\mathrm{BS}} - t_{P_i-\mathrm{RS}-\mathrm{BS}}). \tag{37}$$

Let CR_j denote the candidate region where deploying an RS can reduce the most transmission delay time in the partition ζ_j, where $1 \le j \le k$. We have

$$\mathrm{CR}_j = \underset{\forall\, O_\Phi}{\arg\max} \{|O_\Phi|\}. \tag{38}$$

4) Candidate-Location Identification Phase: This phase further partitions the candidate region into a set of equal-sized grids and then evaluates the improved network capacity of each grid if a relay is deployed at that grid. This phase aims to determine the best grid for deploying an RS. A fine-grain partition of the candidate region will increase the computational cost but obtain a location with higher network capacity. By giving a bound of computational resource, the grid size can be determined. Let $\bar{P}_{g_x}$ denote the set of P_i whose bright region covers grid g_x. Let T_{g_x} denote the throughput from all $P_i \in \bar{P}_{g_x}$ to the BS through RS_j, which is deployed at grid g_x. That is, the throughput obtained from the traffic passing through relay RS_j can be measured by

$$T_{g_x} = \sum_{\forall P_i \in \overline{P}_{g_x}} \min\left(\delta\left(d_{P_i}^{g_z}\right) \times t_{P_i}^{\mathrm{UL}}, \delta\left(d_{g_s}^{\mathrm{BS}}\right) \times t_{\mathrm{RS}_j}^{\mathrm{UL}}\right)$$

$$= \sum_{\forall P_i \in \overline{P}_{g_x}} \min\left(r_{P_i}^{\mathrm{RS}_j} \times t_{P_i}^{\mathrm{UL}}, r_{\mathrm{RS}_j}^{\mathrm{BS}} \times t_{\mathrm{RS}_j}^{\mathrm{UL}}\right),$$

$$1 \le j \le k. \tag{39}$$

Let q_j denote the number of total grids in candidate region CR_j in partition ζ_j. In candidate region CR_j, let $g_{x,j}^{\mathrm{best}}$ denote the grid where deploying an RS can obtain the maximal value of T_{g_x}. The following derives the best grid where deploying an RS can obtain the maximal throughput improvement:

$$g_{x,j}^{\mathrm{best}} = \underset{\forall g_x \in \mathrm{CR}_j}{\arg\max}\{T_{g_x}\}, \qquad 1 \le x \le q_j. \tag{40}$$

C. Algorithm of RPM

Here, the proposed *RPM* algorithm that consists of a main algorithm and several subprocedures is presented. There are four phases designed in the main algorithm. In the *partitioning* phase, step 1 calls $\mathrm{Partition}(\cdot)$ procedure to geographically partition the whole region into k equal-sized partitions. However, the traffic demands of k partitions might be different, resulting in the situation that the loads of k relays are unbalanced. Steps 2 further calls subroutine $\mathrm{Partition\_Adjust}(\cdot)$ to adjust the region of each partition so that all partitions have similar traffic demands. Then, a relay will be allocated to each partition.

The second phase, which is called the *bright-region identification* phase, mainly aims to identify the proper areas for deploying the RS in each partition region. In Steps 3–5, the *RPM* identifies the *bright region* of each subregion by calling $\mathrm{BR}(\cdot)$ subroutine, which constructs the *bright regions*. In the *candidate region identification* phase, steps 6–8 derive the overlapped region covered by the maximal number of bright regions in the same partition. The overlapped region is called the *candidate region*. It means that placing an RS at any position of the *candidate region* could obtain maximal benefits in terms of throughput. Finally, in steps 9–20, the *RPM* partitions the *candidate region* into a number of grids and picks out the best grid for deploying an RS.

The Algorithm of RPM

Input: The location of BS, a region A consists of n subregions $A_1, A_2, \ldots, A_n$, a set of traffic demands $r_{P_i}^{\mathrm{past}}$ of each subregion A_i, and the number of RSs k.
Output: k locations L_{RS} for deploying k RSs.

Initial: $L_{\mathrm{RS}} = \varnothing$;
//Phase I: ***Partitioning***
01. $Partition(A)$;
02. $Partition\_\mathrm{Adjust}(A)$;
// Phase II: ***Bright-Region Identification***
03. **For** $i = 1$ to n
04. $\mathrm{BR}(A_i)$;
05. **End For**
// Phase III: ***Candidate-Region Identification***
06. **For** $j = 1$ to k // k partitions
07. Select the overlapped region that has the maximal relay benefit as candidate region CR_j in the partition ζ_j;
08. **End For**
// Phase IV: ***Candidate-Location Identification***
09. **For** $j = 1$ to k // k partitions
10. $\max T = 0$;
11. **For** $x = 1$ to q_j
12. Calculate T_{g_x} by applying the (39);
13. **If** $T_{g_x} > \max T$ // referred to (40)
14. Set the center of grid g_x as the deployment location of $\mathrm{RS}_j(x, y)$;
15. Update $\mathrm{RS}_j(x, y)$;
16. **End If**
17. **End For**
18. Add $\mathrm{RS}_j(x, y)$ to L_{RS};
19. **End For**
20. **Return** L_{RS}

Partition(A)

Input: Region A consists of n subregions $A_1, A_2, \ldots, A_n$
Output: k partitions that consist of n subregions

01. Let the location of the BS be the center of a circle that contains region A.
02. Equally partition the circle into k partitions by k partitioning lines, so that the angle between every two consecutive lines will be $360/k^{\circ}$.
03. **For** $i = 1$ to k
04. **For** $j = 1$ to n
05. **If** the P_i of A_j is located in the ζ_j
06. Add A_j to the ζ_i;
07. **End If**
08. **End For**
09. **End For**

Partition_Adjust(ζ_i, ζ_{i+1})

Input: Partitions ζ_j and ζ_{j+1}
Output: Well-adjusted partitions ζ_i and ζ_{i+1}

01. **For** $j = 1$ to k

02. **While** $u(\zeta_j)! = (\sum_{j=1}^{k} u(\zeta_j))/k$
03. **If** $u(\zeta_j) < (\sum_{j=1}^{k} u(\zeta_j))/k$
04. Select subregion $A_m \in \zeta_{j+1}$ closest to ζ_j
05. Remove A_m from ζ_{j+1} and add A_m into ζ_j;
06. **else**
07. Select subregion $A_m \in \zeta_j$ closest to ζ_{j+1}
08. Release A_m from ζ_j and add A_m into ζ_{j+1};
09. **End If**
10. **End While**
11. **End For**

BR(A_i)

Input: A subregions A_i
Output: The bright region BR_i of A_i

// Bright Region Identification Mechanism
// Step1
01. **For** $x = 1$ to M
02. **For** $y = 1$ to M
03. Apply the ***Latency Evaluation by Considering Frame Structure*** to check whether $(m^x_{\mathrm{BS-RS}_j}, m^y_{\mathrm{RS}_j-P_i})$ satisfies (15).
//referred to (19)–(34)
04. **If** $(m^x_{\mathrm{BS-RS}_j}, m^y_{\mathrm{RS}_j-P_i})$ satisfies (15)
05. Add $(m^x_{\mathrm{BS-RS}_j}, m^y_{\mathrm{RS}_j-P_i})$ to Ψ_i //referred to (35)
06. **End If**
07. **End For**
08. **End For**
// Step2
09. Partition the serving region of the BS into M concentric circles $\Pi_{\mathrm{BS}} = \{\pi^1_{\mathrm{BS}}, \pi^2_{\mathrm{BS}}, \ldots, \pi^M_{\mathrm{BS}}\}$ according to the M levels of MCS.
// referred to Table I
10. Partition the communication range of the P_i into M concentric circles $\Pi_{P_i} = \{\pi^1_{P_i}, \pi^2_{P_i}, \ldots, \pi^M_{P_i}\}$ according to the M levels of MCS.
11. **For each** $(m^x_{\mathrm{BS-RS}_j}, m^y_{\mathrm{RS}_j-P_i})$ where m_x, $m_y \in \Psi_i$
12. Construct the intersection area $I^{x,y}_i$ of A^X_{BS} and $A^y_{P_i}$
13. **End For**
14. Calculate $\mathrm{BR}_i = \arg\max\{|I^{x,y}_i|\}$ where m_x, $m_y \in \Psi_i$
15. **Return** BR_i

D. Considering IEEE 802.16j Frame Structure

The following presents an example to clearly specify that the importance of considering the 802.16j frame structure.

Consider the example of Fig. 13 where the size of data required to be transmitted from MS to BS is 3600 units. The data rates of links $l(\mathrm{MS}, \mathrm{BS})$, $l(\mathrm{MS}, \mathrm{RS})$ and $l(\mathrm{RS}, \mathrm{BS})$ are 7, 8, and 10 units/slot, respectively. Related works [11], [12] evaluate the total transmission time of two-hop links $l(\mathrm{MS}, \mathrm{RS})$ and $l(\mathrm{RS}, \mathrm{BS})$ as $(300/8 + 3600/10) = 450 + 360 = 810$slots. In addition, the average data rate of two-hop links $l(\mathrm{MS}, \mathrm{RS})$ and $l(\mathrm{RS}, \mathrm{BS})$ is $3600/810 = 4.44$ units/slot, which is smaller than the data rate of direct link $l(\mathrm{MS}, \mathrm{BS})$ (7 units/slot). As a result, the approaches proposed in [11] and [12] will determine that the MS cannot obtain the benefit of relay deployment. As a result, an improper decision will be made. In fact, deploying a relay can help reduce the total transmission time. The following presents how the proposed approach considers the 802.16j frame structure and then make the proper decision.

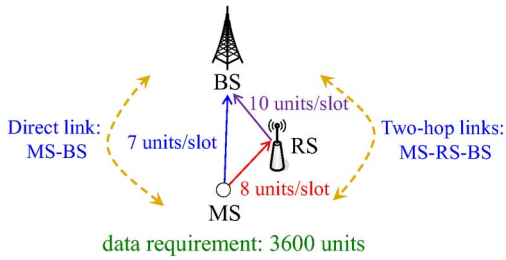


Fig. 13. Direct and two-hop links of the MS.

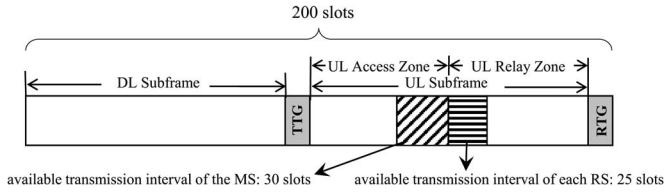


Fig. 14. Consideration of 802.16j frame structure.

Different from the related works, this paper considers the 802.16j frame structure, as shown in Fig. 14, to identify the transmission latency of the direct link and the two-hop links. Let one frame consist of 200 slots. Assume that the available transmission interval of the MS in the access zone is 30 slots. Assume that the available transmission interval of each RS in the relay zone is 25 slots.

1) Direct Link—Transmission Latency of MS: By applying (23), the transmission latency of the MS through the direct link is

$$t_{\mathrm{MS-BS}} = \left\lfloor \frac{3600}{(7\times 30)} \right\rfloor = 17 \text{ frames plus } \left\lceil \frac{(3600 - 7\times 30\times 17)}{7} \right\rceil = 5 \text{ slots.}$$

The total number of slots is $17 \times 200 + 5 = 3405$ slots.

2) Two-Hop Links—Transmission Latency of MS: By applying (28) and (31), the transmission latency of an MS through the two-hop links is

$$t_{\mathrm{MS-RS-BS}} = \max\left(\left\lfloor \frac{3600}{(8\times 30)} \right\rfloor, \left\lfloor \frac{3600}{(10\times 25)} \right\rfloor \right) = \left\lfloor \frac{3600}{(8\times 30)} \right\rfloor = 15 \text{ frames.}$$

The total number of slots is $15 \times 200 = 3000$ slots.

As a result, it is obvious that the MS can obtain the benefit of RS deployment by considering the 802.16j frame structure.

TABLE III
SIMULATION PARAMETERS

Parameter	Value
Carrier frequency	5 GHz
System bandwidth	20 MHz
BS radius	5000 m
BS transmission power	47 dBm
RS transmission power	37 dBm
Noise power	-102 dBm
BS/ *RS* antenna	Omni-directional
Frame duration	10 ms
Modulation coding schemes	BPSK 1/2, 4-QAM 1/2 and 3/4, 16-QAM 1/2 and 3/4, 64-QAM 2/3 and 3/4

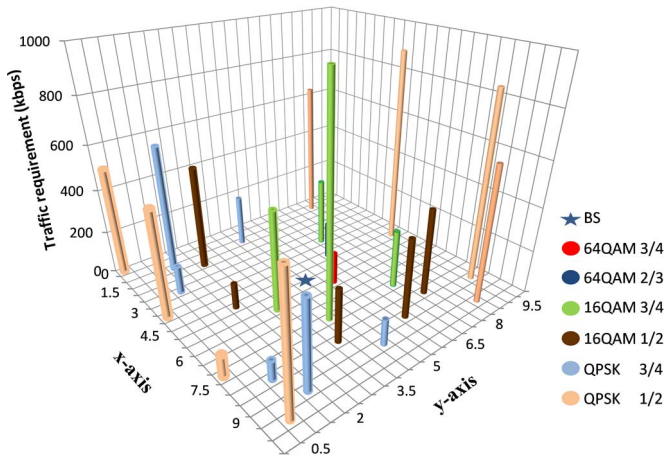


Fig. 15. Example of considered environment that contains 25 representative points.

IV. SIMULATION

This section presents the performance evaluation of the proposed RPM. The proposed RPM is compared with a random-based scheme (*Random*) and the existing scheme proposed in [12], which is referred to as the throughput-oriented scheme (*T-O*). The Random scheme randomly determines the deployment locations of RSs in the networks. The considered region is set by 10 km $\times$ 10 km. One BS is deployed at the center of the region. The region is partitioned into 25 unequal-size subregions. All users (MSs or SSs) are randomly deployed in the considered region. The service types only consider Unsolicited Grant Service (UGS) and real-time polling-service (rtPS) connections. The offered traffic of UGS service is 32 Kb/s, whereas the offered traffic of rtPS service is randomly generated from 100 to 500 Kb/s. Each user establishes one connection belonging to either UGS or rtPS services. After executing the compared RS deployment algorithms, the QoS scheduling algorithm proposed in [17] is adopted to compare the performance of the three compared deployment algorithms. Each result is obtained from the average of 200 independent runs. The 95% confidence interval is always smaller than 5% of the reported values. The MCSs used in the simulations refers to [4]. Table III gives the parameters and their values considered in the simulation.

Fig. 15 gives a scenario of the considered environment, consisting of 25 representative points. The x-axis and y-axis, ranging from 0 to 10 km, represent the coordinates system applied to the service region. The location of the BS is set at the center of the region and marked by a blue star symbol. The location of each P_i and its modulation are presented by bars with different colors, as shown in Fig. 15. Recall that P_i represents the representative point of each subregion A_i. The traffic requirement of P_i is the summation of the traffic requirements of users in subregion A_i. The z-axis denotes the traffic requirement of each representative point P_i. There are 200 users randomly deployed in the BS's serving region.

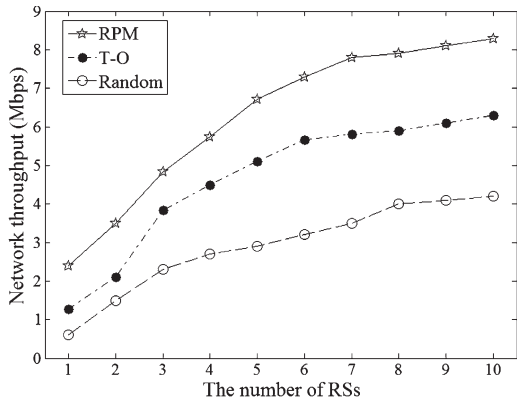


Fig. 16. Comparison of network throughput by varying the numbers of RSs.

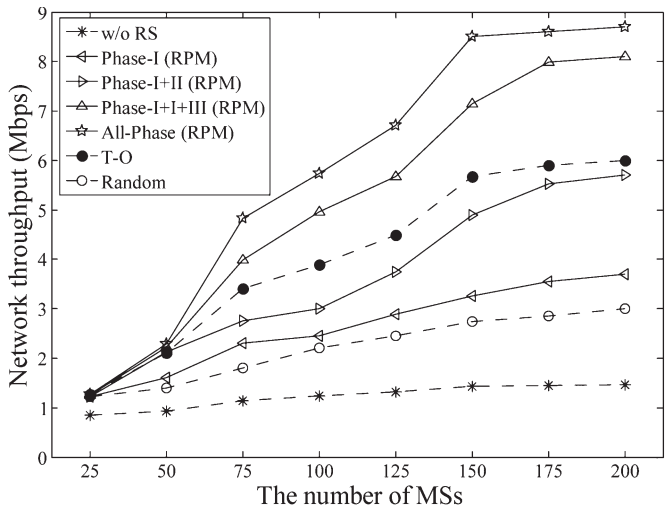


Fig. 17. Comparisons of each phase designed in the proposed *RPM* and the other mechanisms in terms of throughput by varying the number of MSs.

Fig. 16 compares the proposed *RPM* with the other two mechanisms in terms of network throughput. The number of RSs is varied, ranging from 1 to 10. There are 200 users (MSs or SSs) randomly deployed in the considered region. In general, the network throughputs of the three compared mechanisms increase with the number of RSs. This is because the RSs can enhance transmission rate between MSs and the BS; thus, deploying RSs can achieve higher network throughput. The proposed *RPM* determines the grids that deploying RSs can obtain the maximal throughput improvement. Compared with the proposed *RPM* scheme, the existing *T-O* scheme evenly deploys RSs around the BS in a way that all RSs have the same distance to BS. Hence, it might lead to the situation that the some MSs with a large amount of traffic demands cannot

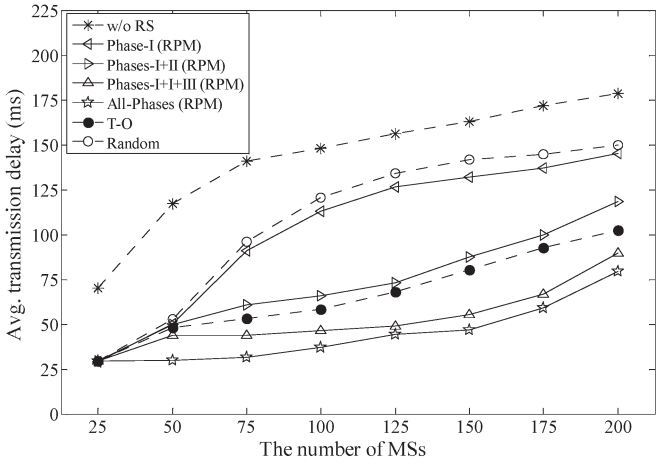


Fig. 18. Comparison of each phase of the proposed *RPM* and the other mechanisms in terms of average transmission delay by varying the number of MSs.

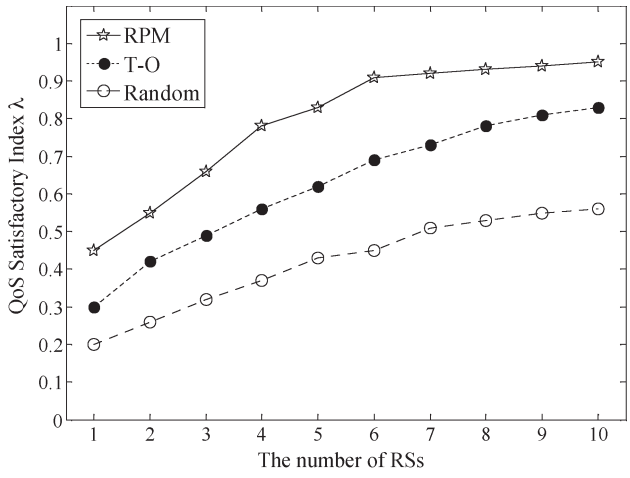


Fig. 19. Comparison of the proposed *RPM* and the other two schemes in terms of QoS satisfactory index λ versus the number of given RSs.

improve their transmission rates because they are far away from the RS. As a result, the network throughput of the *T-O* scheme is smaller than that of the proposed *RPM*. On the other hand, the *Random* scheme might determine infeasible deployment locations for RSs. It results in the worst network throughput. Consequently, as shown in Fig. 16, the proposed *RPM* mechanism outperforms the *Random* and *T-O* schemes in terms of the network throughput.

Fig. 17 compares the proposed *RPM* and other mechanisms in terms of network throughput. The number of MSs is varied ranging from 25 to 200. There are six RSs deployed in the network environment. The abbreviation *Phase_I(RPM)* only involves operations designed in Phase I. It divides the serving region of the BS into several equal-sized partitions and then adjusts the size of each partition so that all partitions have equal traffic demands. Subsequently, *Phase_I(RPM)* deploys one RS at a random selected location in each partition. The abbreviation $Phases\_I + II$(*RPM*) additionally involves Phase II, which aims to deploy RSs at the bright region of each subregion. The abbreviation $Phases\_I + II + III(RPM)$ additionally involves Phase III and thus considers the criteria of candidate region. The abbreviation *All_Phases(RPM)* involves all phases designed in the *RPM*. Finally, the abbreviation *w/o RS* did not deploy any relay in the serving region. In general, the network throughputs of all compared mechanisms increase with the number of MSs. However, *All_Phases(RPM)* achieves the upper bound of throughput when the number of MSs reaches to 150 because the bandwidth has been fully utilized. *All_Phases(RPM)* and $Phases\_I + II + III(RPM)$ have better performance than the others in terms of network throughput. The simulation results depict that each phase designed in *RPM* has its own contribution for improving the network throughput and Phase III has the most significant impact. In general, the proposed *RPM* schemes with *All_Phases(RPM)* and $Phases\_I + II + III(RPM)$ outperform the *T-O* scheme. The major reason is that the candidate region is determined during Phase III in the *RPM*. Then, an RS will be deployed in the candidate region for serving the MSs with superior data demands, increasing network throughput significantly.

Fig. 18 further investigates the impact of each phase of the proposed *RPM* and compares their performance with *w/o RS*, *Random,* and *T-O* mechanisms, in terms of average transmission delay. The number of MSs is varied ranging from 25 to 200. In general, the average transmission delays of all mechanisms increase with the number of MSs. The main reason is that the increment of the number of MSs leads to more traffic requirements. Hence, the BS needs to allocate more time to satisfy the total requirements. Compared with *Phase_I(RPM)*, $Phases\_I + II(RPM)$ significantly reduce the average transmission delay. This indicates that the Phase II, which aims to identify the bright region, has considerable contribution in delay reduction. The *All_Phases(RPM)* outperforms the other three mechanisms, *w/o RS*, *T-O,* and *Random*, in terms of average delay. This occurs because *All_Phases(RPM)* obtains the benefit by applying each phase, and finally, it can deploy RSs at the most feasible locations. The design of Phase I can balance the overhead of RSs to prevent the situation that an RS is deployed in a subregion where there are few traffic demands. By considering the reasonable computational cost, Phase II is used for identifying the bright regions where deploying an RS within the region can improve the transmission rate between P_i and the BS. Phases III and IV can significantly reduce the average transmission delay and obtain the maximal throughput because these phases determine the candidate region and grid and then deploy the RSs at that grid in the candidate region.

Fig. 19 further investigates the performance of three compared mechanisms in terms of the satisfactory index of QoS requirements of MSs. Let μ denote the total number of MSs in the network. Herein, the value of μ is set to 200. Let ω denote the number of MSs whose QoS requirements are satisfied. Equation (41) defines the QoS Satisfaction Index λ

$$\lambda = \frac{\omega}{\mu}. \tag{41}$$

The larger value of λ means that more traffic requirements of MSs are satisfied from the deployed RSs. In general, the value of λ is increased with the number of RSs. When the number

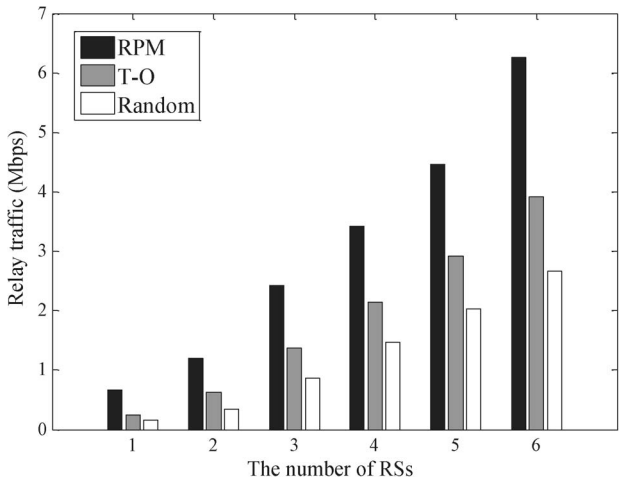


Fig. 20. Comparison of the proposed *RPM* and the other mechanisms in terms of the relay traffic by varying the number of RSs.

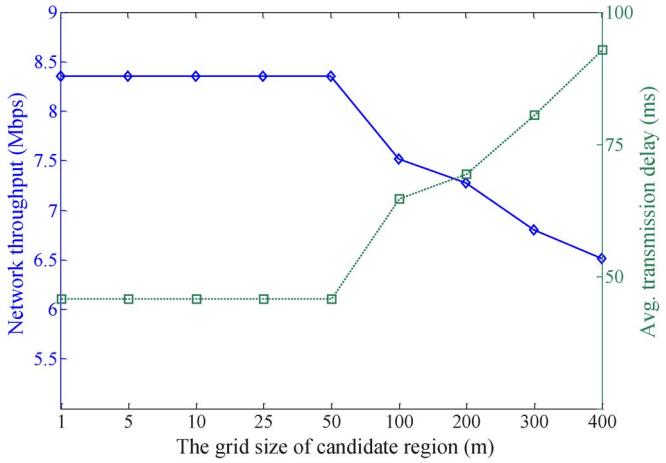


Fig. 21. Impact of grid size on the network throughput and transmission delay of the proposed *RPM*.

of deployed RSs is small, most QoS requirements of MSs cannot be satisfied. However, when the number of deployed RSs grows, the value of λ is close to one, indicating that the all traffic requirements are likely to be satisfied. In general, the proposed *RPM* mechanism outperforms the other two deployment schemes. This is because the *RPM* determines the locations for deploying RSs by taking (15) into consideration. Hence, the proposed RPM has the largest QoS Satisfaction Index λ. In addition, when the number of deployed RSs is more than six, the value of λ of *RPM* grows slowly and closes to one. This indicates that six RSs almost satisfy all traffic requirements. In addition, the *T-O* scheme did not consider uneven distribution of traffic demands and might cause the situation that some users with traffic demands cannot use a superior transmission rate. Therefore, the data might be dropped by some MSs because their transmission delays exceed the time constraint.

Let *relay traffic* represent the overall traffic forwarded through RSs to BS. Fig. 20 compares the proposed *RPM* with the other two schemes in terms of the relay traffic by varying the number of deployed RSs from 1 to 6. In general, the relay traffic of the three mechanisms increases with the number of RSs. This is because that deploying more RSs can help forward more DL/UL traffic requested from the representative points. As shown in Fig. 20, the proposed *RPM* has the highest relay traffic than the other two approaches. The major reason is that the proposed *RPM* applies phases III and IV to identify the candidate regions and grids where deploying an RS can increase the most network throughput. However, the compared *T-O* and *Random* schemes might deploy RSs at improper locations, which are far from the MSs with a large amount of traffic demands. To reduce the transmission delay, MSs directly transmit data to the BS, leading to lower relay traffic.

Fig. 21 exams the impact of grid size on both throughput and average transmission delay. In Phase IV of the proposed *RPM*, the candidate region is partitioned into a set of equal-sized grids, and then the network throughput and transmission delay of each grid is evaluated. Based on the evaluations, an RS can be deployed at the most appreciate grid. In general, a smaller grid size will result in a higher network throughput and lower average transmission delay. However, when the grid size reaches to 50 m, the performance of throughput and average transmission delay drops off significantly. This also indicates that a careful decision on the grid size is important for increasing the network throughput and reducing the transmission delay.

V. Conclusion

Network planning is an important issue for maximizing the network throughput and capacity for a given WiMAX network. Given a network region and k available RSs, this paper has investigated the relay deployment strategy in an IEEE 802.16j WiMAX network. A relay placement mechanism, which is called *RPM*, is proposed for determining the deployment location that satisfies the traffic demand and maximizes the network capacity. The proposed *RPM* first partitions the network region into k partitions according to the historical traffic pattern. Then, the *RPM* takes into consideration the frame structure constraint and bandwidth constraint. For each partition, the *RPM* identifies the bright regions and chooses the best location for deploying an RS. Simulation study reveals that the proposed *RPM* outperforms the existing two mechanisms in terms of network throughput and transmission delay.

References

[1] S. M. Oh and J. H. Kim, "Application-aware design to enhance system efficiency for VoIP services in BWA networks," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 143–154, Feb. 2011.

[2] N. AbuAli, M. Hayajneh, and H. Hassanein, "Congestion-based pricing resource management in broadband wireless networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 8, pp. 2600–2610, Aug. 2010.

[3] A. K. A. Tamimi, C. S. In, and R. Jain, "Modeling and resource allocation for mobile video over WiMAX broadband wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 354–365, Apr. 2010.

[4] *IEEE 802.16 Working Group, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems-Multihop Relay Specification*, IEEE Std. 802.16, 2009.

[5] H. Wang, W. Jia, and G. Min, "Effective channel exploitation in IEEE 802.16j networks for maritime communications," in *Proc. IEEE ICDCS*, Jun. 2011, pp. 162–171.

[6] H. C. Lu and W. Liao, "On cooperative strategies in wireless relay networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 531–535.
[7] C. Y. Hong and A. C. Pang, "3-approximation algorithm for joint routing and link scheduling in wireless relay networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 856–861, Feb. 2009.
[8] K. Sundaresan and S. Rangarajan, "Efficient algorithms for leveraging spatial reuse in OFDMA relay networks," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 1539–1547.
[9] Y. Kim and M. L. Sichitiu, "Optimal max-min fair resource allocation in multihop relay-enhanced WiMAX networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3907–3918, Oct. 2011.
[10] A. So and B. Liang, "Optimal placement of relay infrastructure in heterogeneous wireless mesh networks by bender's decomposition," in *Proc. Int. Conf. QShine Wired/Wireless Netw.*, Aug. 2006, p. 22.
[11] Y. Yu, S. Murphy, and L. Murphy, "Planning base station and relay station locations in IEEE 802.16j multi-hop relay networks," in *Proc. IEEE CCNC*, Jan. 2008, pp. 922–926.
[12] J. H. Huang, L. C. Wang, C. J. Chang, and W. S. Su, "Design of optimal relay location in two-hop cellular systems," *Wireless Netw.*, vol. 16, no. 8, pp. 2179–2189, Nov. 2010.
[13] Y. Yu, S. Murphy, and L. Murphy, "A clustering approach to planning base station and relay station locations in IEEE 802.16j multi-hop relay networks," in *Proc. IEEE ICC*, Jan. 2008, pp. 2586–2591.
[14] D. Yang, X. Fang, G. Xue, and J. Tang, "Relay station placement for cooperative communications in WiMAX networks," in *Proc. IEEE GLOBECOM*, Dec. 2010, pp. 1–5.
[15] H. C. Lu, W. Liao, and F. Y. S. Liu, "Relay station placement strategy in IEEE 802.16j WiMAX networks," *IEEE Trans. Commun.*, vol. 59, no. 1, pp. 151–158, Jan. 2011.
[16] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.
[17] P. S. Mogre, M. Hollick, S. Dimitrov, and R. Steinmetz, "Incorporating spatial reuse into algorithms for bandwidth management and scheduling in IEEE 802.16j relay networks," in *Proc. IEEE 34th Conf. LCN*, Oct. 2009, pp. 384–391.

Chih-Yung Chang (M'05) received the Ph.D. degree in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 1995.

He is currently a Full Professor with the Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan. His current research interests include Internet of Things, cyber–physical systems, wireless sensor networks, ad hoc wireless networks, and Worldwide Interoperability for Microwave Access broadband technologies.

Dr. Chang is a member of the IEEE Computer and Communication Societies. He served as an Area Chair for IEEE AINA'2005, TANET'2000, TANET'2010; as a Vice Chair for IEEE WisCom'2005, EUC'2005, IEEE ITRE'2005, and IEEE AINA 2008; as a Program Cochair for IEEE MNSA'2005, UbiLearn' 2006, WASN'2007, ACM SAMnet'2008, IEEE AHUC'2008, iCube'2010, iCube'2011; as a Workshop Cochair for MSEAT'2003, MSEAT'2004, IEEE INA'2005, ICS'2008, NCS'2009, IEEE VCNA'2009; and as a Publication Chair for MSEAT'2005 and SCORM'2006. He served as an Associate Guest Editor for many SCI-indexed journals, including *International Journal of Ad Hoc and Ubiquitous Computing* (2011–2013), *International Journal of Distributed Sensor Networks* (2012), *IET Communications* (2011), *Telecommunication Systems* (2010), *Journal of Information Science and Engineering* (2008), and *Journal of Internet Technology* (2004 and 2008).

Chao-Tsun Chang received the Ph.D. degree in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 2006.

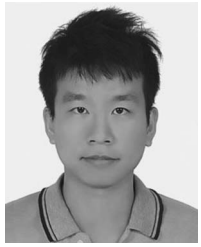
He is currently an Associate Professor with the Department of Information Management, Hsiuping University of Science and Technology, Taichung, Taiwan. In the recent ten years, he has directed 12 researching projects, including five national National Science Council projects and seven information system developments. He is the author of 14 SCI-indexed journal papers, which are published on the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE SENSORS, *Computer Networks*, *ACM/Baltzer Journal of Wireless Networks*, and *Journal of Parallel and Distributed Computing*. His current research interests include ad hoc wireless networks, wireless sensor networks, Bluetooth radio networks, and mobile computing.

Dr. Chang is a member of the IEEE Computer and Communication Societies.

Tzu-Chia Wang received the B.S. and M.S. degrees in computer science and information engineering in 2005 and 2009, respectively, from Tamkang University, Taipei, Taiwan. He is currently working toward the Ph.D. degree in the Department of Computer Science and Information Engineering, Tamkang University.

His research interests include Internet of Things, cyber–physical systems, wireless sensor networks, ad hoc wireless networks, vehicular ad hoc networks, and Worldwide Interoperability for Microwave Access broadband technologies.

Mr. Wang has been a recipient of several scholarship grants in Taiwan and has participated in many projects related to wireless sensor networks, vehicular ad hoc networks, Internet of things, and WiMAX broadband networks.



Ming-Hsien Li received the B.S. degree in computer science and information engineering from Aletheia University, Taipei, Taiwan, in 2006 and the M.S. and Ph.D. degrees in computer science and information engineering from Tamkang University, Taipei, Taiwan, in 2008 and 2013, respectively.

His current research interests include Internet of Things, wireless sensor networks, ad hoc wireless networks, and Worldwide Interoperability for Microwave Access broadband technologies.

Mr. Li received several scholarship grants in Taiwan and has participated in many wireless sensor network projects.