A Resource Allocation Scheme for Scalable Video Multicast in WiMAX Relay Networks

Jang-Ping Sheu, *Fellow*, *IEEE*, Chien-Chi Kao, *Student Member*, *IEEE*, Shun-Ren Yang, *Member*, *IEEE*, and Lee-Fan Chang

Abstract—This paper proposes the first resource allocation scheme in the literature to support scalable-video multicast for WiMAX relay networks. We prove that when the available bandwidth is limited, the bandwidth allocation problems of 1) *maximizing network throughput* and 2) *maximizing the number of satisfied users* are NP-hard. To find the near-optimal solutions to this type of maximization problem in polynomial time, this study first proposes a greedy weighted algorithm, *GWA*, for bandwidth allocation. By incorporating table-consulting mechanisms, the proposed *GWA* can intelligently avoid redundant bandwidth allocation and thus accomplish high network performance (such as high network throughput or large number of satisfied users). To maintain the high performance gained by *GWA* and simultaneously improve its worst case performance, this study extends *GWA* to a bounded version, *BGWA*, which guarantees that its performance gains are lower bounded. This study shows that the computational complexity of *BGWA* is also in polynomial time and proves that *BGWA* can provide at least $1/\rho$ times the performance of the optimal solution, where ρ is a finite value no less than one. Finally, simulation results show that the proposed *BGWA* bandwidth allocation scheme can effectively achieve different performance objectives with different parameter settings.

Index Terms—IEEE 802.16j, multicast, resource allocation, scalable video, WiMAX

1 Introduction

The IEEE 802.16j standard [1] for WiMAX has recently received considerable attention. The IEEE 802.16j amendment is fully compatible with the 802.16e standard [2] and enhances IEEE 802.16e by incorporating relay technology [6], [17], [19]. A typical IEEE 802.16j network consists of base stations (BSs), relay stations (RSs), and subscriber stations (SSs). The radio links between BSs and RSs are called *relay links*, while the links between BSs and SSs or between RSs and SSs are called *access links*. According to the channel qualities of these links, BSs and RSs can dynamically adapt the downlink modulation and coding schemes (MCSs) for data transmission. When RSs are deployed at appropriate locations between the BSs and SSs, the end-to-end channel qualities can be improved and the BSs and RSs can adopt high data-rate MCSs. Based on this improvement in data rate, IEEE 802.16j systems can offer higher throughput and serve more users than IEEE 802.16e systems.

Based on the performance enhancements above, IEEE 802.16j has the potential to provide real-time video multicast services such as mobile IPTV, live video streaming (e.g., athletic events), and online gaming. However, the BSs should allocate bandwidth efficiently to support such bandwidth-hungry services while guaranteeing the quality of user experience (QoE). The bandwidth allocation problems in IEEE 802.16j networks are more challenging than those in IEEE 802.16e networks because the BSs allocate bandwidth not only to the SSs, but also to the RSs. Multicasting also complicates the bandwidth allocation problems. In light of these factors, designing an efficient bandwidth allocation scheme for video multicast services in IEEE 802.16j networks is a critical issue.

Researchers have presented various bandwidth allocation approaches for video services in IEEE 802.16e networks (i.e., single-hop WiMAX systems). The approaches in [3], [4], [5], [16], [20], [33], and [35] allocate bandwidth by exploiting the common technology of scalable video coding (SVC) specified in the H.264/SVC standard [15], [21]. The H.264/SVC standard is extended from H.264/AVC [12], [13], [14] and can further split a video stream into a *base layer* for providing the basic video quality and multiple *enhancement layers* for providing better video quality layer-by-layer. Specifically, if a user already receives n-1 lower video layers, the nth enhancement layer will improve this user's video quality. Using this technology, the bandwidth allocation approaches can appropriately adapt the MCS of each video layer and properly determine the number of video layers to be transmitted. In contrast to transmitting a whole video stream, the flexibility of SVC effectively conserves bandwidth while providing satisfactory video quality. Although IEEE 802.16e bandwidth allocation approaches can achieve high performance in IEEE 802.16e networks, these approaches cannot be directly applied to IEEE 802.16j networks to achieve equivalent high performance. This is because IEEE 802.16e bandwidth allocation approaches do not consider multihop relay issues. When considering relay issues, the bandwidth allocation problem

- *J.-P. Sheu and S.-R. Yang are with the Department of Computer Science and Institute of Communications Engineering, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013, R.O.C. E-mail: {sheujp, sryang}@cs.nthu.edu.tw.*
- *C.-C. Kao and L.-F. Chang are with the Department of Computer Science, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013, R.O.C. E-mail: mickey@wmnet.cs.nthu.edu.tw, g9762504@oz.nthu.edu.tw.*

Manuscript received 30 Dec. 2010; revised 19 Oct. 2011; accepted 27 Oct. 2011; published online 15 Nov. 2011.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2010-12-0591.
Digital Object Identifier no. 10.1109/TMC.2011.250.

becomes more complex, since the bandwidth should be allocated not only to the SSs, but also to the RSs.

Researchers have presented numerous schemes to solve the multihop bandwidth allocation problems. The bandwidth allocation schemes proposed in [7], [10], [22], [23], [24], [25], [26], [27], [28], [29] were specifically designed for data unicast in IEEE 802.16j networks. These schemes allocate bandwidth in a dynamic manner with different performance objectives (e.g., delay minimization [24], [25] and throughput maximization [22], [23]). Several new bandwidth allocation schemes [8], [9], [11] have been developed in the last few years for data multicast in IEEE 802.16j networks. The scheme proposed in [8] conserves bandwidth by identifying the best multicast paths among the BSs, RSs, and SSs. Another scheme [9] allocates bandwidth in a heuristic manner, leading to a maximized number of service recipients. Nevertheless, almost all (if not all) of the existing bandwidth allocation schemes for IEEE 802.16j networks consider only simple data traffic. Without the consideration of SVC, existing schemes (for both data unicast and data multicast) allocate bandwidth inefficiently for real-time video services. Existing bandwidth allocation schemes also fail to consider the diversity of device capabilities. Not all video subscribers can display the video content with the highest quality due to their hardware limitations. Allocating a certain bandwidth for transmitting high-quality video contents to low-capability users would waste bandwidth. Moreover, the existing solutions to the multihop bandwidth allocation problems usually rely on sophisticated algorithms. When service providers take realism into account, some of these sophisticated algorithms become impractical due to their high computational complexities. Considering all these concerns, existing bandwidth allocation schemes still cannot satisfy the requirements of real-time video multicast services in IEEE 802.16j networks.

This study thoroughly investigates the bandwidth allocation problems of video multicast in IEEE 802.16j networks and provides a solution by designing a novel bandwidth allocation algorithm to fulfill different performance objectives. Under the constraint of limited available bandwidth, the proposed algorithm targets to maximize network performance, e.g., to maximize the network throughput or to maximize the number of satisfied users. Note that we prove that the maximization problems of network performance (i.e., maximization problem of network throughput and maximization problem of the number of satisfied users) are NP-hard. The proposed algorithm solves these NP-hard problems using a greedy weighted method. This greedy weighted method is a heuristic method that can find a suboptimal solution in polynomial time. Specifically, instead of enumerating all the possible choices to find the globally optimal solution, the greedy weighted method makes the locally optimal choice at each decision stage (according to a predefined weighted value). This approach significantly reduces the computational complexity. Unlike conventional greedy algorithms, the proposed algorithm applies table consulting in each greedy stage to determine whether any redundant bandwidth allocation exists. If yes, the algorithm removes this allocation to reclaim the bandwidth. Consequently, the proposed algorithm can achieve higher efficiency than previously proposed greedy bandwidth allocation algorithms. Depending on the performance objectives and operation profits, service providers can adopt the proposed algorithm with different parameter settings (i.e., different weighted values) for bandwidth allocation in WiMAX relay systems.

This study makes five main contributions:

1. this study is the first to investigate the bandwidth allocation problem of scalable video multicast in WiMAX relay systems;
2. this study proves that the maximization problems of the network throughput and of the number of satisfied users are NP-hard, and proposes a polynomial-time suboptimal solution to these problems;
3. this study provides detailed design, time-complexity analysis, and worst case analysis;
4. theoretical analysis shows that the worst case performance of the proposed bandwidth allocation algorithm is lower bounded by the approximation ratio of $2 \times D_{BS} \times D_{RS}$, where D_{BS} and D_{RS} are the degrees of the BS and RS, respectively;
5. extensive simulations demonstrate that the proposed algorithm significantly outperforms the naive heuristic and approximates the optimal solution. In other words, the performance of the proposed algorithm is at least 94 percent of that of the optimal solution.

The remainder of this paper is organized as follows: Section 2 reviews related work on bandwidth allocation schemes in WiMAX systems. Section 3 first models the bandwidth allocation problems for IEEE 802.16j networks and then proposes a novel bandwidth allocation scheme for different performance objectives. Section 4 proves the NP-hardness of the multicast bandwidth allocation problems and theoretically analyzes the performance of the proposed bandwidth allocation scheme. Section 5 evaluates the performance of the proposed bandwidth allocation scheme. Finally, Section 6 concludes the paper.

2 Related Work

This study investigates the bandwidth allocation problems of scalable video multicast in multihop WiMAX relay systems. Two categories of related work are classified and discussed as follows:

Scalable video multicast in single-hop WiMAX systems. The bandwidth allocation schemes for scalable video multicast in IEEE 802.16e (single-hop WiMAX) networks were proposed in [3], [4], [5], [33], [35] and the references therein. The authors of [4] presented a two-level bandwidth allocation scheme. In the first level, the bandwidth is allocated to transmit the base-layer videos using lower data-rate modulations, such as BPSK and QPSK. In the second level, the remaining bandwidth is allocated to transmit the enhancement layers using higher data-rate modulations, such as 16-QAM and 64-QAM. When these modulation schemes are determined appropriately, the two-level approach can efficiently allocate bandwidth to multicast the scalable videos. Huang et al. [3] argued for an enhanced system throughput using opportunistic multicasting, and introduced an opportunistic bandwidth allocation scheme for layered-video multicast services. Based on the results in [3], the same authors developed a complete opportunistic scheduling mechanism called Opportunistic

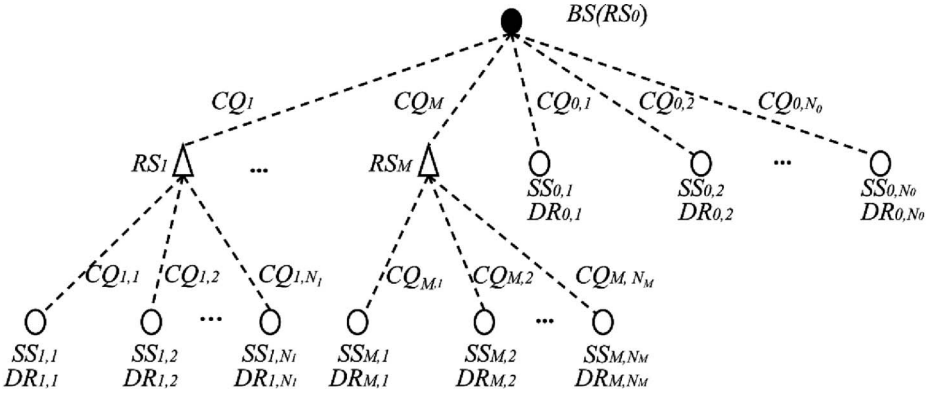

Fig. 1. An example of the proposed network model.

Layered Multicasting (OLM) [35]. This mechanism achieves high efficiency in terms of user satisfaction by sequentially determining the appropriate MCS and picking the proper video layer for each multicast. However, these bandwidth allocation schemes for IEEE 802.16e networks usually cannot accomplish equivalent performance in IEEE 802.16j networks. This is primarily because 1) the bandwidth should be allocated not only to the BS but also to the RSs, and 2) the RSs causes interference problems for data multicasting.

Simple data multicast in multihop WiMAX systems. The multihop (relay-based) bandwidth allocation schemes for IEEE 802.16j networks were presented in [8], [9], [11]. The previous works closest to ours are [8] and [9]. The authors of [8] introduced a dynamic station selection (DSS) algorithm designed to maximize the number of service recipients by effectively conserving bandwidth consumption. They first modeled the multihop network as the tree topology and then used DSS to decide the lowest bandwidth-consuming path for data multicast among the SSs, RSs, and BS. In [9], the same author designed the dynamic resource allocation (DRA) algorithm to solve the similar multicast problem. The DRA algorithm allocates bandwidth in a heuristic manner aiming to maximize the number of recipients. Although the above methods already consider the diversities of bandwidth budgets and channel qualities, they do not consider the diversity of the user's device capabilities. Moreover, these previous bandwidth allocation schemes are not suitable for scalable-video applications because video subscribers with diverse capabilities may request the same video with different data rates (i.e., different numbers of video layers). When using these simple-traffic mechanisms in [8], [9], [11] for scalable video applications, the bandwidth may be allocated to transmit high-quality video contents to low-capability users, which ultimately wastes bandwidth. In addition to the differences on user diversity and video scalability, the proposed bandwidth allocation scheme is also substantially different from those in the related work. The bandwidth allocation schemes in [8] and [9] employ multiple loops to examine the performance of the different combinations of recipients, which results in extremely high computational complexity. The bandwidth allocation scheme proposed in this study applies greedy methods to achieve low computational complexity while incorporating the table-consulting mechanisms to avoid redundant bandwidth allocation. Therefore, the proposed bandwidth allocation scheme can efficiently allocate bandwidth while maintaining low computational complexity. The concept behind the proposed bandwidth allocation scheme is substantially different from those in the related work.

3 Resource Allocation Scheme

This section first describes the assumptions of the network environment and then proposes an efficient scheme for resource allocation in the considered networks.

3.1 Network Model and Notation

This study considers the resource allocation problems in two-hop WiMAX relay networks similar to the existing research [7], [8], [9], [10], [11], [22], [23], [24], [25], [26], [27], [28], [29]. This study does not investigate multihop (more than two hops) problems because 1) more-than-2-hop scenarios usually make the resource allocation problems too complex and thereby the solutions too impractical; and 2) the network throughput performance usually decreases as the number of hops increases [17]. As illustrated in Fig. 1, the proposed model for two-hop WiMAX relay networks consists of one BS, M RSs, and N SSs. For consistency, the BS is regarded as the 0th RS and is denoted by RS_0 in the following discussion, while the RSs are denoted by RS_1 to RS_M. An SS can associate either with the BS or with one of the RSs, and the number of SSs associated with RS_m is denoted by N_m. The notation $SS_{m,n}$ represents the nth SS associated with RS_m. For each $SS_{m,n}$, $0 \leq m \leq M$ and $1 \leq n \leq N_m$.

In Fig. 1, CQ_m represents the channel quality of the link between the BS and RS_m while $CQ_{m,n}$ represents the channel quality between RS_m and $SS_{m,n}$. Assume that the video streams for the links with lower channel quality should be transmitted by the modulation schemes with higher reliability. This model considers four modulation schemes: BPSK, QPSK, 16-QAM, and 64-QAM. BPSK provides the highest reliability of these four schemes (making it suitable for links with bad channel quality) while 64-QAM provides the fastest transmission rate (making it suitable for links with good channel quality).

The data-rate requirement of $SS_{m,n}$ is denoted by $DR_{m,n}$ in Fig. 1. Assume that SSs with different device capabilities can request the same video with different video quality. The H.264/SVC standard [12], [15], [21] defines many video quality levels with their respective maximum and minimum

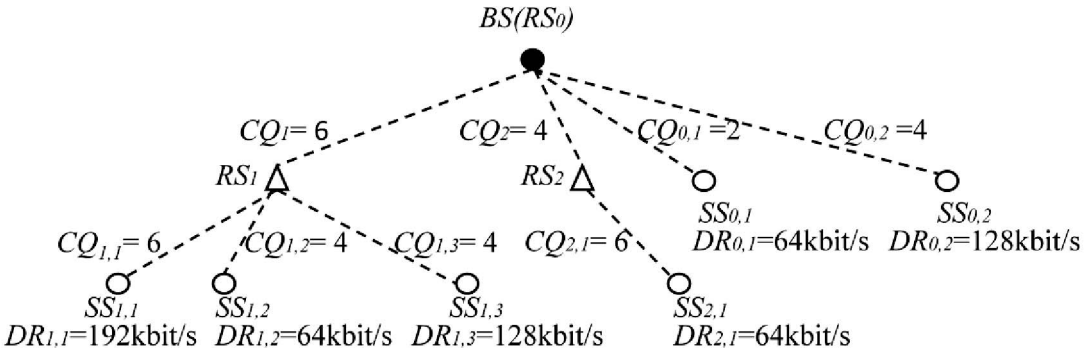

Fig. 2. An example of multicast network.

data-rate requirements. This paper considers six of them suitable for wireless applications. For the six video quality levels 1, 1b, 1.1, 1.2, 1.3, and 2 (see [21]), the proposed model uses the respective maximum bit rates, 64, 128, 192, 384, 768, and 2,048 kbit/s, as representative data rates. Note that these representative data rates are specified for convenience, and the proposed resource allocation scheme can also operate under any other data rates. An SS can select a quality level depending on its device capability. In the case when $SS_{m,n}$ requests a video under video quality level 1, the BS should guarantee a 64 kbit/s data rate to $SS_{m,n}$, and its $DR_{m,n}$ equals 64 kbit/s. Furthermore, to provide diverse data rates, H.264/SVC allows a video stream to be split into one base layer and multiple enhancement layers. This study assumes that a video can be split into six layers (one base layer and five enhancement layers) corresponding to the six video quality levels. For example, a user with the requirement of 64 kbit/s can be satisfied by receiving the base layer, while a user with the requirement of 128 kbit/s can be satisfied by receiving the base layer and one enhancement layer.

3.2 Concept of Multicast Resource Allocation Scheme in WiMAX Relay Networks

WiMAX relay networks make resource allocation decisions once per frame. An IEEE 802.16j frame consists of a downlink subframe and an uplink subframe. This study focuses on the downlink multicast problems. The downlink subframe can be divided into an access zone and a relay zone. In the access zone, the BS transmits the video data to its served RSs and SSs. In the relay zone, the RSs further relay the video data to their served SSs. To determine the data transmissions within each frame, the BS should make a scheduling decision at the beginning of each frame using an appropriate *resource allocation scheme*. Before specifying the proposed resource allocation scheme, this section first introduces some basic concepts: 1) bandwidth estimation, 2) multicast consideration, and 3) allocation decision.

This study uses the Shannon-Hartley theorem to estimate the bandwidth consumption [18]. This theorem states that $C \leqq 2B\log_2 L$, where C represents the channel capacity in bits per second, B represents bandwidth in Hertz, and L is the number of discrete signal elements for a modulation scheme. Following this inequality, we conservatively estimate the bandwidth consumption as $B = C/(\log_2 L)$. For instance, L's of BPSK, QPSK, 16-QAM, and 64-QAM are 2, 4, 16, and 64, respectively. If $SS_{m,n}$ links to the BS with BPSK and has the data-rate requirement $DR_{m,n} = 64$ kbit/s, the bandwidth consumption for serving $SS_{m,n}$ would be $64/(\log_2 2) = 64$ k Hertz. Note that the selection of an appropriate modulation scheme depends on the channel quality (i.e., CQ_m and $CQ_{m,n}$). To simplify this discussion, we divide CQ_m ($CQ_{m,n}$) into four classes with respect to their modulation schemes and quantify them as $\log_2 L$. Specifically, when CQ_m ($CQ_{m,n}$) equals 1, 2, 4, and 6, the corresponding modulation schemes BPSK, QPSK, 16-QAM, and 64-QAM should be adopted. Accordingly, the minimum bandwidth requirement $B_{m,n}$ for serving $SS_{m,n}$ can be estimated as $DR_{m,n}/CQ_m + DR_{m,n}/CQ_{m,n}$, where $DR_{m,n}/CQ_m$ represents the bandwidth requirement for the first hop from BS to RS_m and $DR_{m,n}/\ CQ_{m,n}$ represents the bandwidth requirement for the second hop from RS_m to $SS_{m,n}$. Fig. 2 shows an example for serving $SS_{1,1}$, in which the minimum bandwidth consumption $B_{1,1} = DR_{1,1}/CQ_1 + DR_{1,1}/CQ_{1,1} = 192/6 + 192/6 = 64$k Hertz.

When a BS (RS) multicasts a video stream, all group-member SSs within the BS's (RS's) coverage receive the stream simultaneously. If the BS (RS) multicasts the video stream using the modulation scheme corresponding to CQ_m ($CQ_{m,n}$), the video stream can only be correctly decoded by SSs that are linked to the BS (RS) with the channel quality higher than or equal to CQ_m ($CQ_{m,n}$). Take Fig. 2 as an example. If the BS multicasts the video stream using 64-QAM (corresponding to $CQ_m = 6$), only RS_1 can correctly decode the video stream. On the other hand, if the BS multicasts using 16-QAM (corresponding to $CQ_m = 4$), RS_1, RS_2, and $SS_{0,2}$ can correctly decode the video stream. This example demonstrates that different modulation schemes favor different numbers of SSs, which is referred to as the *multicast effect* in this paper. Note that a different modulation scheme also consumes a different amount of bandwidth. Therefore, an efficient multicast resource allocation scheme must consider both the multicast effect and bandwidth consumption when determining an appropriate modulation scheme for streaming transmissions.

Based on SSs' data-rate requirements and the applied modulation schemes, the resource allocation schemes make bandwidth allocation decisions once per scheduling frame. The proposed resource allocation scheme uses several *multicast tables* to represent the allocation decisions. The multicast table MT_m represents the allocation decision of RS_m (note that MT_0 is for the BS). For example, MT_1 in Fig. 3 indicated that RS_1 decides to allocate data rates of 128 kbit/s with 16-QAM and of 64 kbit/s with 64-QAM.

MT_1 for RS_1	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	128 kbit/s
64-QAM(6)	64 kbit/s

Fig. 3. An example of multicast table for RS_1.

	$DR_{m,n}$	$B_{m,n}$	$W_{m,n}=DR_{m,n}/B_{m,n}$	Priority
$SS_{0,1}$	64 kbit/s	64/2=32k Hertz	64/32=2	6
$SS_{0,2}$	128 kbit/s	128/4=32k Hertz	128/32=4	1
$SS_{1,1}$	192 kbit/s	192/6 + 192/6 = 64k Hertz	192/64=3	2
$SS_{1,2}$	64 kbit/s	64/6 + 64/4 = 26.67k Hertz	64/26.67=2.4	3
$SS_{1,3}$	128 kbit/s	128/6 + 128/4 = 53.33k Hertz	128/53.33=2.4	4
$SS_{2,1}$	64 kbit/s	64/6 + 64/4 = 26.67k Hertz	64/26.67=2.4	5

Fig. 4. Serving priority for GWA_{NT}.

Note that this decision consumes bandwidth as $128/4 + 64/6 = 42.7\text{k}$ Hertz.

3.3 Proposed Bandwidth Allocation Scheme

This section considers a WiMAX relay network with a limited amount of bandwidth, and proposes an algorithm to maximize the target network performance (e.g., network throughput and number of satisfied users). Because the bandwidth is limited, not all the SSs can be satisfied at the same time. In this case, it is necessary to determine which set of SSs to serve first and determine the corresponding serving priority to maximize network performance. Unfortunately, this type of problem is NP-hard (as Section 4.1 formally proves). This study reduces the well-known NP-hard problem called *0/1 knapsack problem* [30] to the network performance maximization problem. To find the near-optimal solutions of these NP-hard problems in polynomial time, we first develop a greedy weighted algorithm *GWA* that determines the set of SSs and serving priority leading to the suboptimal network performance. Define the weighted value $W_{m,n}$ of each $SS_{m,n}$ as the performance gain per bandwidth unit. Following the decreasing order of $W_{m,n}$, *GWA* sequentially examines the SSs for bandwidth allocation. Note that, if more than one SS has the same $W_{m,n}$ value, the SSs' bandwidth requests can be handled in a random order. To guarantee that the worst case performance is lower bounded, we enhance the proposed *GWA* to be a bounded greedy weighted algorithm, called *BGWA*. *BGWA* maintains the low complexity of *GWA* while improving the worst case behavior of *GWA*.

This section is organized as follows: First, Sections 3.3.1 and 3.3.2 provide two examples targeting different performance objectives to illustrate the basic idea of *GWA*: maximizing network throughput and maximizing the number of satisfied users. Then, Section 3.3.3 elaborates on the *GWA* procedure. Finally, Section 3.3.4 extends *GWA* to the bounded version *BGWA*.

3.3.1 Maximizing Network Throughput under Limited Resources

This section attempts to maximize the network throughput under the limited bandwidth assumption. To approach the goal of maximizing the network throughput, we first apply the greedy weighted algorithm *GWA* using the weighted value $W_{m,n}$ as the throughput gain per bandwidth unit $DR_{m,n}/B_{m,n}$, where $B_{m,n}$ is the bandwidth consumption to serve $SS_{m,n}$ for both the first hop and the second hop. To avoid ambiguity, we use GWA_{NT} to represent the *GWA* with the goal of maximizing network throughput with the weighted value $W_{m,n} = DR_{m,n}/B_{m,n}$.

To elaborate on the basic idea of the proposed GWA_{NT}, consider the example in Fig. 2 and its corresponding multicast tables construction. Fig. 4 illustrates the weighted values for the SSs and the resultant serving priority. Initially, all the DR fields in the multicast tables MT_0, MT_1, and MT_2 are set as zero. Since $SS_{0,2}$ has the highest throughput gain per bandwidth unit, $SS_{0,2}$ is first examined. To satisfy $SS_{0,2}$, the DR field of 16-QAM in MT_0 is modified from 0 to 128 kbit/s (Fig. 5a). Next, GWA_{NT} processes the 192-kbit/s request of $SS_{1,1}$. To support the first-hop transmission to $SS_{1,1}$ (from the BS to RS_1), GWA_{NT} first consults with MT_0. The current MT_0 (i.e., that in Fig. 5a) indicates that 128 kbit/s out of $SS_{1,1}$'s 192-kbit/s request has simultaneously been satisfied during the last bandwidth assignment to $SS_{0,2}$. In this case, only additional 64 kbit/s is required and the DR field of 64-QAM corresponding to $CQ_1 = 6$ is updated as 64 kbit/s (Fig. 5b). Moreover, for the second-hop transmission, the BS must allocate bandwidth to support the 192 kbit/s from RS_1 to $SS_{1,1}$ using 64-QAM. Thus, the DR field of 64-QAM in MT_1 is updated as 192 kbit/s (Fig. 5b). GWA_{NT} further examines $SS_{1,2}$. The MT_0 value in Fig. 5b indicates that RS_1 has been assigned 192 kbit/s (128 kbit/s using 16-QAM and 64 kbit/s using 64-QAM) higher than $DR_{1,2} = 64$ kbit/s. Consequently, no additional bandwidth is required for $SS_{1,2}$'s first-hop transmission. On the other hand, for the second-hop transmission to $SS_{1,2}$, MT_1 shows that although

(a) After serving $SS_{0,2}$

MT_0 for BS	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	0→128 kbit/s
64-QAM(6)	0

MT_1 for RS_1	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	0
64-QAM(6)	0

(b) After serving $SS_{1,1}$

MT_0 for BS	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	128 kbit/s
64-QAM(6)	0→64 kbit/s

MT_1 for RS_1	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	0
64-QAM(6)	0→192 kbit/s

(c) After serving $SS_{1,2}$

MT_0 for BS	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	128 kbit/s
64-QAM(6)	64 kbit/s

MT_1 for RS_1	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	0→64 kbit/s
64-QAM(6)	192 kbit/s→128 kbit/s

Fig. 5. An example for the table-based GWA_{NT}.

	$user_{m,n}$	$B_{m,n}$	$W_{m,n}=user_{m,n}/B_{m,n}$	Priority
$SS_{0,1}$	1	64/2=32k Hertz	1/32=0.03125	4
$SS_{0,2}$	1	128/4=32k Hertz	1/32=0.03125	5
$SS_{1,1}$	1	192/6 + 192/6 = 64k Hertz	1/64=0.015625	6
$SS_{1,2}$	1	64/6 + 64/4 = 26.67k Hertz	1/26.67=0.0375	1
$SS_{1,3}$	2	128/6 + 128/4 = 53.33k Hertz	2/53.33=0.0375	2
$SS_{2,1}$	1	64/6 + 64/4 = 26.67k Hertz	1/26.67=0.0375	3

Fig. 6. Serving priority for GWA_{SU}.

$192\ \text{kbit/s} > DR_{1,2} = 64\ \text{kbit/s}$ has been scheduled at RS_1, this data rate using 64-QAM cannot be received by $SS_{1,2}$ due to its relatively poor channel quality. In this case, the DR field of 16-QAM in MT_1 should be updated as 64 kbit/s with respect to the requirement of $SS_{1,2}$ (Fig. 5c). Note that, to avoid bandwidth wastage, 64 kbit/s can be deducted from the previously assigned 192 kbit/s for transmission from RS_1 to $SS_{1,1}$. This is because using a more robust modulation scheme, $SS_{1,1}$ with better channel quality can also receive the later scheduled 64 kbit/s from RS_1 to $SS_{1,2}$. Therefore, the DR field of 64-QAM in MT_1 is updated from 192 to 128 kbit/s (see MT_1 in Fig. 5c). The reclaimed bandwidth can be used to serve more SSs, improving the efficiency of bandwidth utilization. This process is repeated for the remaining SSs based on their serving priority until the bandwidth has been exhausted or all SSs' requests have been processed. If the current available bandwidth is insufficient to satisfy an SS in a greedy stage (i.e., the decision stage for determining whether the bandwidth is allocated to an SS), GWA_{NT} simply skips the SS and proceeds to serve the next SS whose requirement can be filled unless no such an SS exists. The proposed GWA_{NT} is significantly different from a pure greedy algorithm in that GWA_{NT} can reclaim bandwidth and effectively avoid unnecessary multicast operations by looking up the multicast tables in each greedy stage.

3.3.2 Maximizing Number of Satisfied Users under Limited Resources

This section attempts to maximize the number of satisfied users under the limited bandwidth assumption. User satisfaction is a matter of concern for service providers because it is eventually reflected in operational profits. This study defines *satisfied users* as the users whose data-rate requirements are fully met. To approach the goal of maximizing the number of satisfied users, this study applies the greedy weighted algorithm *GWA* using a weighted value different from that in Section 3.3.1. To maximize the number of satisfied users, we define the weighted value as the ratio of the number of satisfied users to bandwidth consumption. That is, $W_{m,n} = user_{m,n}/B_{m,n}$, where $user_{m,n}$ is the number of concurrently satisfied users while serving the data-rate/bandwidth requirements of $SS_{m,n}$. For instance in Fig. 2, $user_{1,3} = 2$ because when satisfying $SS_{1,3}$, $SS_{1,2}$ with equal channel condition and fewer data-rate requirements can be satisfied simultaneously. On the other hand, $user_{1,1} = 1$ because while serving $SS_{1,1}$, its neighbors $SS_{1,2}$ and $SS_{1,3}$ are associated with poorer channel quality (i.e., $CQ_{1,2} = CQ_{1,3} < CQ_{1,1}$) and thus cannot be satisfied concurrently. Fig. 6 lists $user_{m,n}$, $B_{m,n}$, $W_{m,n}$ and the serving priority for the SSs in Fig. 2. This study uses GWA_{SU} to represent the *GWA*

(a)

Priority	$SS_{m,n}$	Residual Bandwidth	Allocation
1	$SS_{1,2}$	80 – (64/6+64/4) = 53.33k Hertz	Yes
2	$SS_{1,3}$	53.33–(64/6+64/4) = 26.66k Hertz	Yes
3	$SS_{2,1}$	26.66 – [(64/4– 64/6)+64/6] = 10.66k Hertz	Yes
4	$SS_{0,1}$	10.66 – [(64/2– 64/4)+0] < 0k Hertz	No
5	$SS_{0,2}$	10.66 – [(64/4– 64/6)+0] = 5.33k Hertz	Yes
6	$SS_{1,1}$	5.33 – (64/6+64/6) < 0k Hertz	No

(b)

MT_0 for BS	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	128kbit/s
64-QAM(6)	0

MT_1 for RS_1	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	128kbit/s
64-QAM(6)	0

MT_2 for RS_2	DR
BPSK(1)	0
QPSK(2)	0
16-QAM(4)	0
64-QAM(6)	64kbit/s

Fig. 7. An example for the table-based GWA_{SU}.

with the goal of maximizing the number of satisfied users and with the weighted value $W_{m,n} = user_{m,n}/B_{m,n}$.

Consider the network in Fig. 2 as an example to illustrate the execution of GWA_{SU}. Assume that the available bandwidth $B_{Limit} = 80\text{k}$ Hertz. The proposed GWA_{SU} examines the SSs in the prioritized order, $SS_{1,2}$, $SS_{1,3}$, $SS_{2,1}$, $SS_{0,1}$, $SS_{0,2}$ and then $SS_{1,1}$ (Fig. 7a). When $SS_{1,2}$ is examined, the current residual bandwidth B_{Res} is sufficient to support the 64 kbit/s from the BS to $SS_{1,2}$ via RS_1 (i.e., $B_{Res} - (DR_{1,2}/CQ_1 + DR_{1,2}/CQ_{1,2}) \geq 0$). Therefore, the DR fields of 64-QAM in MT_0 and 16-QAM in MT_1 are first modified as 64 kbit/s. Additional bandwidth $(DR_{1,3} - DR_{1,2})/CQ_1 = 10.67\text{k}$ Hertz is required to support the first-hop transmission of $SS_{1,3}$ (from the BS to RS_1). Similarly, the second-hop transmission of $SS_{1,3}$ (from RS_1 to $SS_{1,3}$) requires $(DR_{1,3} - DR_{1,2})/CQ_{1,3} = 16\text{k}$ Hertz. Because B_{Res} is enough to satisfy both requirements of $SS_{1,3}$ (i.e., $B_{Res} - (10.67 + 16)\text{k Hertz} \geq 0$), the DR fields of 64-QAM in MT_0 and 16-QAM in MT_1 are modified as 128 kbit/s. After that, $SS_{2,1}$ is examined. Although MT_0 already scheduled $128\ \text{kbit/s} > DR_{2,1} = 64\ \text{kbit/s}$, the data rate supported by 64-QAM cannot be received by RS_2 due to its poor channel condition. Hence, to satisfy the first hop of $SS_{2,1}$, MT_0 should provide additional 64 kbit/s using 16-QAM. This consumes $DR_{2,1}/CQ_2 = 64/4 = 16\text{k}$ Hertz. In this case, 64 kbit/s can be deducted from the DR field of 64-QAM in MT_0 to avoid redundant transmission and thereby reclaim $64/6 = 10.67\text{k}$ Hertz. To further support the second hop of $SS_{2,1}$, the bandwidth consumption $DR_{2,1}/CQ_{2,1} = 10.67\text{k}$ Hertz is required. Accordingly, the overall bandwidth requirement when processing $SS_{2,1}$ is $(16 - 10.67) + 10.67 = 16\text{k}$ Hertz. Because B_{Res} can satisfy this requirement, the DR fields of 16-QAM in MT_0, 64-QAM in MT_0 and 64-QAM in MT_2 are all modified as 64 kbit/s. Then, GWA_{SU} sequentially examines $SS_{0,1}$, $SS_{0,2}$ and $SS_{1,1}$ in the same way. After all the above steps, the multicast tables are determined (Fig. 7b). GWA_{SU} finally allocates the bandwidth accordingly to satisfy $SS_{1,2}$, $SS_{1,3}$, $SS_{2,1}$, and $SS_{0,2}$.

3.3.3 Procedure of the Proposed Bandwidth Allocation Scheme

Based on the concept of *GWA* mentioned in Sections 3.3.1 and 3.3.2, this section further elaborates on the *GWA* procedure. Note that the GWA_{NT} procedure is the same as

that of GWA_{SU}; the only difference between them is that GWA_{NT} and GWA_{SU} employ different weighted values $W_{m,n}$ for each $SS_{m,n}$. Depending on the concerns of the service provider, using different weighted values in the same procedure achieves different performance objectives, e.g., maximizing network throughput or maximizing the number of satisfied users (see the respective examples in Sections 3.3.1 and 3.3.2).

Algorithm 1 shows the proposed procedure of *GWA* (for both GWA_{NT} and GWA_{SU}). This algorithm defines B_{Limit} as the total bandwidth that can be allocated. Let B_{Res} be a temporary variable that indicates the current residual bandwidth. B_{Res} is initialized as B_{Limit} and all the DR fields of the multicast tables are initialized as zero (lines 1-4). After initialization, the proposed scheme sorts all the SSs in decreasing order based on their weighted values, i.e., $W_{m,n}$ for $SS_{m,n}$ (line 5). Let $hop_{m,n}$ be the $SS_{m,n}$ hop count. Define $B_{m,n}^{i\text{-}hop}$ as $SS_{m,n}$'s ith-hop bandwidth consumption. For an $SS_{m,n}$ taken in the sorting order, *GWA* first sets $B_{m,n}^{i\text{-}hop}, \forall i$, as zero (lines 7 and 8). *GWA* then consults with the current multicast tables to compute $B_{m,n}^{i\text{-}hop}$ in each hop for satisfying $SS_{m,n}$ (lines 9-26). For the first (respectively, second) hop, define *CQ* and *MT* to represent CQ_m and MT_0 (respectively, $CQ_{m,n}$ and MT_m) (lines 10-14). To compute $B_{m,n}^{i\text{-}hop}$, *GWA* employs a temporary variable $DR_{temp}^{i\text{-}hop}$ to record the total data rate associated with the modulation schemes no less reliable than *modulation*[*CQ*] in the current *MT*, where *modulation*[*CQ*] is the modulation scheme corresponding to *CQ* (line 15). If $DR_{m,n} \stackrel{\leq}{=} DR_{temp}^{i\text{-}hop}$, no extra bandwidth allocation is required for $SS_{m,n}$ in the ith hop and thereby $B_{m,n}^{i\text{-}hop} = 0$ (lines 16-17). Otherwise, additional data rate $(DR_{m,n} - DR_{temp}^{i-hop})$ should be supported in *MT*. In this case, the bandwidth consumption $B_{m,n}^{i\text{-}hop}$ becomes $(DR_{m,n} - DR_{temp}^{i-hop})/CQ$ (lines 18-19). However, if the current *MT* already supports any data rates using the other modulation schemes that are less reliable than *modulation*[*CQ*], some bandwidth can be reclaimable (lines 20-24) (e.g., the case that serving $SS_{1,2}$ in Section 3.3.1 or the case that serving $SS_{2,1}$ in Section 3.3.2). To calculate the reclaimable bandwidth, *GWA* checks each DR field corresponding to a modulation scheme less reliable than *modulation*[*CQ*] (line 20). If any existing data rates with less-reliable modulation can also be satisfied by $DR_{m,n}$ using *modulation*[*CQ*], these data rates are redundant (line 21). The total reclaimable bandwidth with respect to redundant data rates is recorded in the temporary variable $B_{reclaim}^{i\text{-}hop}$ (line 22). After the above bandwidth computation of each hop for $SS_{m,n}$, the current residual bandwidth B_{Res} will be examined to determine whether it is sufficient to support $(B^{1-hop}m, n - B_{reclaim}^{1-hop}) + (B_{m,n}^{2-hop} - B_{reclaim}^{2-hop})$ (line 27). If yes, the data-rate requirements of $SS_{m,n}$ are reflected in the corresponding multicast tables (lines 28-42). In the case of $DR_{m,n} > DR_{temp}^{i\text{-}hop}$, *MT* is modified by adding $(DR_{m,n} - DR_{temp}^{i\text{-}hop})$ into its DR field of *CQ* (lines 34-35). *GWA* then deducts the redundant data rates (if any) from the current *MT* (lines 36-40) to reclaim bandwidth for later use (line 43). The procedures of bandwidth computation and table modification repeat for each SS until either B_{Res} is exhausted (lines 44-46) or all SSs have been processed (lines 6-48). Finally, the BS allocates bandwidth according to the generated multicast tables (line 49).

Algorithm 1. The proposed *GWA* for bandwidth allocation in IEEE 802.16j networks

Input: $\{CQ_0, CQ_1, \ldots, CQ_M\}$,
$\{CQ_{0,1}, \ldots, CQ_{m,Nm}\}$,
$\{DR_{0,1}, \ldots, DR_{m,Nm}\}$,
$\{W_{0,1}, \ldots, W_{m,Nm}\}$.

1 $B_{Res} \leftarrow B_{Limit}$
2 **for** $m = 0$ to M
3 $MT_m \leftarrow \emptyset$
4 **end for**
5 sort SSs into monotonously decreasing order by weighted values $W_{m,n}$
6 **for** each $SS_{m,n,}$ taken in the sorting order
7 **for** $i = 1$ to $hop_{m,n}$
8 $B_{m,n}^{i\text{-}hop} \leftarrow 0$
9 **for** $i = 1$ to $hop_{m,n}$
10 **if** $i = 1$
11 let *CQ* and *MT* represent CQ_m and MT_0
12 **else if** $i = 2$
13 let *CQ* and *MT* represent $CQ_{m,n}$ and MT_m
14 **end if**
15 $DR_{temp}^{i\text{-}hop} \leftarrow$ total DR in *MT* using modulation schemes no less reliable than *modulation*[*CQ*]
16 **if** $DR_{m,n} \stackrel{\leq}{=} DR_{temp}^{i\text{-}hop}$
17 $B_{m,n}^{i\text{-}hop} = 0$
18 **else if** $DRm, n > DR_{temp}^{i\text{-}hop}$
19 $B_{m,n}^{i\text{-}hop} = (DR_{m,n} - DR_{temp}^{i\text{-}hop})/CQ$
20 **for** each DR in *MT* using a modulation scheme less reliable than *modulation*[*CQ*] in the order from BPSK to 64-QAM
21 **if** some data rate is redundant due to the support of $DR_{m,n}$ using *modulation*[*CQ*]
22 $B_{reclaim}^{i\text{-}hop} \leftarrow B_{reclaim}^{i\text{-}hop}$ + the reclaimable bandwidth if deducting the redundant data rate from *MT*
23 **end if**
24 **end for**
25 **end if**
26 **end for**
27 **if** $B_{Res} \stackrel{\geq}{=} (B_{m,n}^{1-hop} - B_{reclaim}^{1-hop}) + (B_{m,n}^{2-hop} - B_{reclaim}^{2-hop})$
28 **for** $i = 1$ to $hop_{m,n}$
29 **if** $i = 1$
30 let *CQ* and *MT* represent CQ_m and MT_0
31 **else if** $i = 2$
32 let *CQ* and *MT* represent $CQ_{m,n}$ and MT_m
33 **end if**
34 **if** $DR_{m,n} > DR_{temp}^{i\text{-}hop}$
35 add ($DR_{m,n}$-$DR_{temp}^{i\text{-}hop}$) into the DR field of *CQ* in *MT*
36 **for** each DR in *MT* using a modulation scheme less reliable than *modulation*[*CQ*] in the order from BPSK to 64-QAM
37 **if** some data rate is redundant due to the support of $DR_{m,n}$ using *modulation*[*CQ*]
38 deduct the redundant data rate from *MT*
39 **end if**
40 **end for**

41 **end if**
42 **end for**
43 $B_{Res} \leftarrow B_{Res} - [(B_{m,n}^{1-hop} - B_{reclaim}^{1-hop}) + (B_{m,n}^{2-hop} - B_{reclaim}^{2-hop})]$
44 **if** $B_{Res} = 0$
45 **break**
46 **end if**
47 **end if**
48 **end for**
49 allocate bandwidth according to $\{MT_0, MT_1, \ldots, MT_M\}$

The time complexity of sorting (line 5) in *GWA* is $O(N \log N)$. The complexity of the for-loop (lines 6-48) to compute the bandwidth consumption and to modify the multicast tables is $O(N)$. Thus, the total complexity is $O(N\log N + N) = O(N\log N)$.

3.3.4 Proposed ρ-Approximation Algorithm

This section simply modifies the proposed resource allocation scheme, *GWA*, to improve its performance in the worst case scenario while keeping its high performance in average cases. Although *GWA* can perform well in most cases (e.g., the examples in Sections 3.3.1 and 3.3.2), *GWA* is not a ρ-approximation algorithm, where ρ is an approximation ratio.

Definition 1. *An algorithm A is a ρ-approximation algorithm if and only if the algorithm satisfies the following equation:*

$$P^*(I)/P_A(I) \leq \rho, \text{ for all } I,$$

where I is an instance of the target problem, and $P^*(I)$ and $P_A(I)$ are the profits (e.g., throughput or number of satisfied users) gained by the optimal solution and algorithm A, respectively.

Based on the proposed *GWA*, we propose a ρ-approximation algorithm, *Bounded GWA* (*BGWA*), to guarantee that the performance is lower bounded in the worst case. Algorithm 2 shows the *BGWA* procedure. Line 1 initializes two solution sets, M_1 and M_2, each of which contains all the multicast tables, i.e., MT_0 to MT_M, for the BS and RSs. Line 2 executes *GWA* to generate a set of multicast tables (see details in Section 3.3.3), and records this set into M_1. Line 3 first removes the SSs who can be satisfied using M_1. The algorithm then reexecutes *GWA* for the remainder SSs to generate another set of multicast tables, and records this set in M_2. Line 4 determines the final solution set, M^*, as M_1 or M_2 whichever yields the higher profit (i.e., network performance). Finally, in line 5, the BS allocates bandwidth according to the multicast tables in M^*. Section 4.2 formally proves that *BGWA* is a ρ-approximation algorithm.

Algorithm 2. The ρ-approximation algorithm *BGWA* for bandwidth allocation in IEEE 802.16j networks

Input: $CQ_{RS} = \{CQ_0, CQ_1, \ldots, CQ_M\}$,
$CQ_{SS} = \{CQ_{0,1}, \ldots, CQ_{m,Nm}\}$,
$DR = \{DR_{0,1}, \ldots, D_{Rm,Nm}\}$,
$W = \{W_{0,1}, \ldots, W_{m,Nm}\}$.

1 let M_1 and M_2 be the two solution sets of multicast tables
2 $M_1 \leftarrow$ GWA(CQ_{RS}, CQ_{SS}, DR, W)
3 $M_2 \leftarrow$ GWA(CQ_{RS}', CQ_{SS}', DR', W'), where the inputs (CQ_{RS}', CQ_{SS}', DR', W') are obtained by removing the SSs who can be satisfied using M_1
4 $M^* \leftarrow M_1$ or M_2 whichever results in the higher profits
5 allocate bandwidth according to M^*

BGWA executes *GWA* twice to generate two possible sets of multicast tables, and then simply selects the better set as the solution. The complexity of *GWA* is $O(N\log N)$. Therefore, the total complexity of *BGWA* is $O(2N\log N) = O(N\log N)$. Note that *BGWA* maintains the low complexity of *GWA* while further providing a performance bound to guarantee its effectiveness in the worst case.

4 Theoretical Analysis

4.1 NP-Hardness of the Multicast Bandwidth Allocation Problems

This section proves that both the maximization problem of network throughput and the maximization problem of the number of satisfied users are NP-hard. Specifically, this section proves the above statement by reducing the well-known NP-hard problem called *0/1 knapsack problem* [30] to the problems of 1) maximizing throughput and 2) maximizing the number of satisfied users. The corresponding definition and property are given as follows:

Definition 2. *The 0/1 knapsack problem is a combinational-optimization problem: Given n objects, each with a weight W_i and a profit P_i, determine which objects should be taken so that the total weight is less than or equal to the limit W_{limit} and the total profit is as large as possible.*

Property 1. *The 0/1 knapsack problem is NP-hard [30].*

Theorem 1. *The maximization problem of network throughput and the maximization problem of the number of satisfied users in WiMAX relay networks are NP-hard.*

Proof. Let the maximization problem of network throughput and the maximization problem of the number of satisfied users in the general WiMAX relay networks be problems A and B, respectively. First consider the simple case in Fig. 8, where the channel qualities between the BS and RSs are perfect (i.e., $CQ_1 = \infty$, $CQ_2 = \infty$, and $CQ_3 = \infty$). Let A' and B' be the maximization problem of network throughput and the maximization problem of the number of satisfied users in the above special case. To prove A and B are NP-hard, it is sufficient to show that A' and B' are NP-hard because A' and B' are simpler than A and B. When solving A' and B', the bandwidth consumption for multicast through the relay links (between the BS and RSs) approximates to zero (i.e., $\max(DR_{1,1}, DR_{2,1}, DR_{3,1})/\min(CQ_1, CQ_2, CQ_3) = 192/\infty \approx 0$) and can therefore be neglected. Based on this observation, A' and B' can be regarded as bandwidth allocation problems in one-hop networks and can thus be simply modeled as the 0/1 knapsack problem. Suppose the limited bandwidth is $W_{limit} = 32$k Hertz. The bandwidth consumptions for serving $SS_{1,1}$, $SS_{2,1}$, and $SS_{3,1}$ are $W_1 = DR_{1,1}/CQ_{1,1} = 16$k Hertz, $W_2 = DR_{2,1}/CQ_{2,1} = 10.67$k Hertz, and $W_3 = DR_{3,1}/CQ_{3,1} = 32$k Hertz. While serving $SS_{1,1}$, $SS_{2,1}$, and $SS_{3,1}$, the gains of network throughput or satisfied users can be regarded as the profit values P_1, P_2,

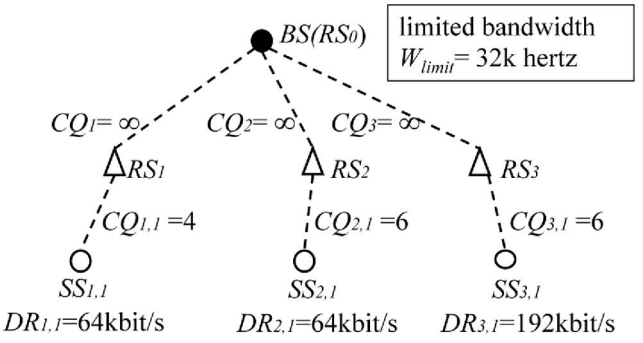

Fig. 8. A special case that can be transformed to the 0/1 knapsack problem.

and P_3, respectively. Note that $P_1 = P_2 = 64$ kbit/s and $P_3 = 192$ kbit/s for A' while $P_1 = P_2 = P_3 = 1$ user for B'. Let S_{opt} be the set of SSs which maximize the total profit $\sum P_i X_i$ while the total bandwidth consumption $\sum W_i X_i$ is less than or equal to W_{limit}, where X_i is an indicator function. If $SS_{i,1}$ is served, $X_i = 1$. Otherwise, $X_i = 0$. Solving A' and B' is the same as determining S_{opt} (i.e., determining which SSs should be served so that the total profit is maximized). From Definition 2 and Property 1, both A' and B' are 0/1 knapsack problem with NP-hard complexity. Since A' and B' in the one-hop special case are already NP-hard, A and B in the two-hop general cases must be NP-hard. □

4.2 Performance Analysis of the Proposed Bandwidth Allocation Scheme

Section 3.3 proposes greedy weighted algorithms, *GWA* and *BGWA*, to solve the multicast bandwidth allocation problem. This section analyzes the worst case performance of the proposed algorithms. Theorem 2 shows that *BGWA* is lower bounded by the approximation ratio of $2 \times D_{BS} \times D_{RS}$, where D_{BS} and D_{RS} are the degrees of the BS and RS, respectively. Furthermore, Theorem 3 shows that we can enhance the lower bound of *BGWA* to be the approximation ratio of $2 \times D_{max}$, where $D_{max} = \max(D_{BS}, D_{RS})$.

First, we note that *GWA* outperforms a unicast bandwidth allocation algorithm when allocating the same amount of bandwidth to the same SS. Consider the example in Fig. 2. Suppose *GWA* and a unicast algorithm allocate the same bandwidth to a certain SS, e.g., $SS_{1,2}$. In this case, *GWA* can multicast video streaming to satisfy $SS_{1,2}$ and $SS_{1,3}$ concurrently, while the unicast algorithm can only satisfy $SS_{1,2}$. Therefore, given the same amount of available bandwidth and the same serving priority, *GWA* consistently satisfies more SSs and thereby yields more profit (e.g., network throughput and the number of satisfied users) than a unicast bandwidth allocation algorithm.

Remark 1. With the same bandwidth budget and priority, *GWA* outperforms a unicast bandwidth allocation algorithm.

Because *GWA* outperforms a unicast bandwidth allocation algorithm, we can regard a unicast algorithm as the lower bound of profit for *GWA*. To derive a real bound for *GWA*, we first concentrate on the performance bound of a unicast bandwidth allocation algorithm. We can model the unicast bandwidth allocation problem as the well-known 0/1 knapsack problem [30] (see Definition 2). First, let the total amount of available bandwidth be the capacity of the knapsack. Second, regard SSs as objects. Third, regard the bandwidth consumption and the performance gain for serving an SS as the weight and the profit for taking an object, respectively. Finally, the 0/1 knapsack problem is how to take a set of objects (SSs) so that the profit (performance gain) can be maximized while the total weight (bandwidth consumption) of the taken objects (served SSs) is less than or equal to the knapsack's capacity (available bandwidth).

Remark 2. The unicast bandwidth allocation problem can be modeled as the 0/1 knapsack problem.

The *nonincreasing first fit* (*NIFF*) [31] algorithm is the most well-known greedy solution to the 0/1 knapsack problem. This algorithm sorts objects in the nonincreasing order of their profit-to-weight ratio. Following the sorting order, the objects are taken into the knapsack under the constraint that the total weight of the taken objects cannot exceed the knapsack's capacity. When applying *NIFF* to the bandwidth allocation problem, *NIFF* serves SSs following the same priority as *GWA*, i.e., according to the SSs' profit-to-weight ratio. Following Remark 1, because *NIFF* is a unicast algorithm, we can regard *NIFF* as the lower bound of profit for *GWA*. However, *NIFF* cannot provide a real bound for *GWA* because *NIFF* is not a ρ-approximation algorithm [32]. Without a real bound, *GWA* is also not a ρ-approximation algorithm.

Fortunately, the authors in [32] developed a ρ-approximation algorithm called *Bounded NIFF* (*BNIFF*) for the 0/1 knapsack problem. *BNIFF* provides a tight bound of profit by executing *NIFF* twice. Because *BNIFF* can guarantee a performance bound in the worst case, this study uses the concept of *BNIFF* to modify *GWA*, i.e., Section 3.3.4 extends *GWA* to a bounded version *BGWA*. We formally prove that the proposed *BGWA* is a ρ-approximation algorithm for the multicast bandwidth allocation problem as follows.

Lemma 1 first shows that the profit gained by *BNIFF* is lower bounded. Note that Lemma 1 has been proved in [32]. Based on Lemma 1, Lemma 2 proves that the profit gained by *BGWA* is also lower bounded. On the other hand, Lemma 3 shows that the optimal solution to the multicast bandwidth allocation problem is upper bounded. Using the lower bound in Lemma 2 and the upper bound in Lemma 3, Theorem 2 proves that *BGWA* is a ρ-approximation algorithm. Let I be an instance of the bandwidth allocation problem. Define $W[i]$ and $P[i]$ as the weight and the profit of object i in a 0/1 knapsack problem.

Lemma 1. $P_{BNIFF}(I) \geq \frac{1}{2} \times \sum_{i=1}^{k} P[i]$, for all I.

Proof. Based on Remark 2, we first model a unicast bandwidth allocation problem as a 0/1 knapsack problem. For a general 0/1 knapsack problem, assume that

$$W[i] \leq W_{limit}, \text{ for all } i, \tag{1}$$

where $W[i]$ is the weight of object i and W_{limit} is the capacity of knapsack. Consider the unicast solution *BNIFF*. To solve the 0/1 knapsack problem, *BNIFF* executes *NIFF* twice. The *BNIFF* procedure is the same

as that of *BGWA* (see Algorithm 2) but replaces the multicast tables and *GWA* in *BGWA* with unicast tables and *NIFF*. *BNIFF* denotes the first and second *NIFF*s as $NIFF_1$ and $NIFF_2$, respectively. For each *NIFF*, objects are taken following the nonincreasing order of their profit-to-weight ratio. We denote k to be the first number such that kth object cannot be taken into the knapsack. From this definition and (1), we can infer that the 1st-$(k-1)$th objects must be taken by $NIFF_1$ while the kth object must be taken by $NIFF_2$. Accordingly, $NIFF_1$ and $NIFF_2$ take at least k objects. Thus,

$$\sum_{\substack{\text{object } i \text{ taken} \\ \text{by } NIFF_1}} P[i] + \sum_{\substack{\text{object } i \text{ taken} \\ \text{by } NIFF_2}} P[i] \geq \sum_{i=1}^{k} P[i], \tag{2}$$

where $P[i]$ is the profit of object i. *BNIFF* adopts $NIFF_1$ or $NIFF_2$ whichever results in higher profit. Therefore,

$$P_{BNIFF}(I) \geq \frac{1}{2} \times \left(\sum_{\substack{\text{object } i \text{ taken} \\ \text{by } NIFF_1}} P[i] + \sum_{\substack{\text{object } i \text{ taken} \\ \text{by } NIFF_2}} P[i] \right), \text{ for all } I. \tag{3}$$

Equations (2) and (3) can then derive a lower bound of profit for *BNIFF*.

$$\begin{aligned} P_{BNIFF}(I) &\geq \frac{1}{2} \times \left(\sum_{\substack{\text{object } i \text{ taken} \\ \text{by } NIFF_1}} P[i] + \sum_{\substack{\text{object } i \text{ taken} \\ \text{by } NIFF_2}} P[i] \right) \\ &\geq \frac{1}{2} \times \sum_{i=1}^{k} P[i], \text{ for all } I. \end{aligned} \tag{4}$$

□

Lemma 2. $P_{BGWA}(I) \geq \frac{1}{2} \times \sum_{i=1}^{k} P[i]$, for all I.

Proof. We use the concept of *BNIFF* to design the proposed *BGWA*. The *BGWA* procedure (see Algorithm 2) is the same as that of *BNIFF* while *NIFF* in *BNIFF* is replaced by *GWA* in *BGWA*. Note that *GWA* and *NIFF* serve SSs (take objects) following the same order, based on the SSs' (objects') profit-to-weight ratio. Remark 1 indicates that *GWA* outperforms *NIFF* because *NIFF* is a unicast solution to the bandwidth allocation problem. Therefore,

$$P_{GWA}(I) \geq P_{NIFF}(I), \text{ for all } I. \tag{5}$$

BGWA executes *GWA* twice and then selects the better *GWA* to allocate bandwidth. Likewise, *BNIFF* executes *NIFF* twice and selects the better *NIFF* to allocate bandwidth. Accordingly, the profits gained by *BGWA* and by *BNIFF* are actually gained by *GWA* and by *NIFF*, respectively. Based on this observation and (5), we can then derive that

$$P_{BGWA}(I) \geq P_{BNIFF}(I), \text{ for all } I. \tag{6}$$

Consequently, (6) and Lemma 1 can derive a lower bound of profit for *BGWA*.

$$P_{BGWA}(I) \geq P_{BNIFF}(I) \geq \frac{1}{2} \times \sum_{i=1}^{k} P[i], \text{ for all } I. \tag{7}$$

□

Lemma 3. $P^*(I) \leq (D_{BS} \times D_{RS}) \times \sum_{i=1}^{k} P[i]$, for all I.

Proof. First, consider unicast solutions to the bandwidth allocation problem. For a bandwidth allocation problem I, denote $P_u{}^*(I)$ as the profit gained by the optimal unicast solution. Applying Remark 2, we can model the unicast bandwidth allocation problem I as a 0/1 knapsack problem. For the 0/1 knapsack problem, it is well known [31] that

$$P_u^*(I) \leq \sum_{i=1}^{k} P[i], \text{ for all } I, \tag{8}$$

where the definitions of k and of $P[i]$ are the same as those in Lemma 1. Next, consider multicast solutions to the bandwidth allocation problem. Denote D_{BS} as the degree of BS and D_{RS} as the maximum degree of RSs. For example, Fig. 2 shows that $D_{BS} = 4$ and $D_{RS} = 3$. Given a certain amount of available bandwidth, the profit gained by multicast solutions is at most $(D_{BS} \times D_{RS})$ times higher than the profit gained by unicast solutions. Thus,

$$P^*(I) \leq (D_{BS} \times D_{RS}) \times P_u^*(I), \text{ for all } I. \tag{9}$$

Consequently, (8) and (9) can derive an upper bound of profit for the multicast bandwidth allocation problem.

$$\begin{aligned} P^*(I) &\leq (D_{BS} \times D_{RS}) \times P_u^*(I) \\ &\leq (D_{BS} \times D_{RS}) \times \sum_{i=1}^{k} P[i], \text{ for all } I. \end{aligned} \tag{10}$$

□

Theorem 2. $\frac{P^*(I)}{P_{BGWA}(I)} \leq 2 \times D_{BS} \times D_{RS}$, for all I. *BGWA is a* $(2 \times D_{BS} \times D_{RS})$*-approximation algorithm.*

Proof. Following Lemma 2 and Lemma 3, we can derive the approximation ratio of the proposed *BGWA* as follows:

$$\begin{aligned} \frac{P^*(I)}{P_{BGWA}(I)} &\leq \frac{P^*(I)}{\frac{1}{2} \times \sum_{i=1}^{k} P[i]} \leq \frac{D_{BS} \times D_{RS} \times \sum_{i=1}^{k} P[i]}{\frac{1}{2} \times \sum_{i=1}^{k} P[i]} \\ &= 2 \times D_{BS} \times D_{RS}, \text{ for all } I. \end{aligned} \tag{11}$$

Definition 1 shows that *BGWA* is a $(2 \times D_{BS} \times D_{RS})$-approximation algorithm for the multicast bandwidth allocation problem in WiMAX relay networks.

Although the worst case performance of *BGWA* is already bounded by the approximation ratio of $2 \times D_{BS} \times D_{RS}$, the worst case performance can be further enhanced. Theorem 3 shows that we can enhance the worst case performance of *BGWA* by reserving some bandwidth in the BS. To prove this, first consider the special cases of the bandwidth allocation problem, where the channel qualities between the BS and RSs are perfect, i.e., $CQ_1 = CQ_2 = \cdots = CQ_M = \infty$. Let I' be an instance of the bandwidth allocation problem in the special case. □

Lemma 4. $P_{BGWA}(I') \geq \frac{1}{2} \times \sum_{i=1}^{k} P[i]$, for all I'.

Proof. I' must be included in I because I' is a special case of the bandwidth allocation problem. That is,

$$\exists I : I' \equiv I, \text{ for all } I'. \tag{12}$$

Since I' is included in I, we can rewrite Lemma 2 as follows:

$$P_{BGWA}(I') \geq \frac{1}{2} \times \sum_{i=1}^{k} P[i], \text{ for all } I'. \quad (13)$$

□

Lemma 5. $P^*(I') \leq D_{max} \times \sum_{i=1}^{k} P[i]$, for all I'.

Proof. Following (12) in Lemma 4, we can rewrite Lemma 3 as

$$P^*(I') \leq (D_{BS} \times D_{RS}) \times \sum_{i=1}^{k} P[i], \text{ for all } I'. \quad (14)$$

The upper bound in (14) can be tighter in the special case. Consider the general network model in Fig. 1. In the special case, assume that $CQ_1 = CQ_2 = \cdots = CQ_M = \infty$. First, consider the SSs subordinated to the BS (i.e., $SS_{0,1} - SS_{0,N0}$). Suppose a unicast algorithm and a multicast algorithm allocate the same bandwidth to serve a certain SS, i.e., $SS_{0,i}$, where $1 \leq i \leq N_0$. In this case, the unicast algorithm can get profit (e.g., network throughput) from only $SS_{0,i}$, while the multicast algorithm may get profit from all the $SS_{0,i}$'s neighbors. Note that each SS subordinated to the BS has at most D_{BS} neighbors. Therefore, in the special case, when allocating bandwidth to a certain SS subordinated to the BS, the profit gained by a multicast algorithm can be at most D_{BS} times higher than that gained by a unicast algorithm. Second, consider the SSs subordinated to an RS. Note that in the special case, because $CQ_1 = CQ_2 = \cdots = CQ_M = \infty$, it is possible to ignore the relay links between the BS and RSs. Thus, concentrate on the access links between the RSs and SSs. Suppose a multicast algorithm allocates a certain amount of bandwidth to serve a certain SS, i.e., $SS_{j,k}$, where $1 \leq j \leq M$ and $1 \leq k \leq N_M$. In this case, the multicast algorithm may get profits (e.g., network throughput) from all the $SS_{j,k}$'s neighbors. Because D_{RS} is the maximum degree of RSs, each SS subordinated to an RS has at most D_{RS} neighbors. Accordingly, in the special case, when allocating bandwidth to a certain SS subordinated to an RS, a multicast algorithm can gain the profit at most D_{RS} times higher than a unicast algorithm. The discussion above indicates that in the special case (i.e., $CQ_1 = CQ_2 = \cdots = CQ_M = \infty$), when allocating bandwidth to a certain SS subordinated to either a BS or an RS, the profit gained by a multicast algorithm is at most $\max(D_{BS}, D_{RS})$ times higher than that gained by a unicast algorithm. Let $D_{max} = \max(D_{BS}, D_{RS})$. We can then derive a tighter bound of profit for the optimal multicast solution to the bandwidth allocation problem under the special case

$$P^*(I') \leq D_{max} \times P_u^*(I'), \text{ for all } I'. \quad (15)$$

Following (8) in Lemma 3, (12) in Lemma 4 and (15), we can finally derive the upper bound of profit for the optimal multicast bandwidth allocation algorithm in the special case as follows:

$$P^*(I') \leq D_{max} \times P_u^*(I') \leq D_{max} \times \sum_{i=1}^{k} P[i], \text{ for all } I'. \quad (16)$$

□

Theorem 3. *BGWA can be improved to be a $(2 \times D_{max})$-approximation algorithm by reserving bandwidth in the BS.*

Proof. Following Lemma 4 and Lemma 5, we can derive the approximation ratio of *BGWA* in the special case

$$\frac{P^*(I')}{P_{BGWA}(I')} \leq \frac{P^*(I')}{\frac{1}{2} \times \sum_{i=1}^{k} P[i]} \leq \frac{D_{max} \times \sum_{i=1}^{k} P[i]}{\frac{1}{2} \times \sum_{i=1}^{k} P[i]} = 2 \times D_{max}, \text{ for all } I'. \quad (17)$$

Definition 1 indicates that *BGWA* is a $(2 \times D_{max})$-approximation algorithm in the special case. Then, we show how to transform a problem instance I into I' by reserving bandwidth in the BS. Consider the general network model in Fig. 1. To eliminate the effect of the relay links between the BS and RSs, we can reserve some bandwidth for these relay links. Let DR_l be the data-rate requirement of a relay link l between the BS and RS_l. DR_l is at most the same as the maximum data-rate requirement of the SSs subordinated to RS_l. That is, $DR_l \leq \max(DR_{l,1}, \ldots, DR_{l,Nl})$, where $1 \leq l \leq M$. Accordingly, the data-rate requirement of the relay links is at most

$$\max(DR_1, \ldots, DR_M) \leq \max[\max(DR_{1,1}, \ldots, DR_{1,N1}), \ldots, \max(DR_{M,1}, \ldots, DR_{M,NM})] = \max(DR_{1,1}, \ldots, DR_{M,NM}).$$

To satisfy the requirement of all the relay links, the BS can multicast a video stream with the data rate of $\max(DR_{1,}\ 1, \ldots, DR_{M,NM})$ using the modulation scheme corresponding to the (relatively) poorest channel quality, i.e., $\min(CQ_1, \ldots, CQ_M)$. Consequently, when we reserve $\max(DR_{1,1}, \ldots, DR_{M,NM})/\min(CQ_1, \ldots, CQ_M)$k Hertz in the BS, we can simply transform the bandwidth allocation problem in Fig. 1 into the special case, where the links between the BS and RSs are neglected. Let $\lambda = \max(DR_{1,1}, \ldots, DR_{M,NM})/\min(CQ_1, \ldots, CQ_M)$. Thus, the amount of bandwidth to reserve in the BS is at most λ k Hertz. That is, we can transform I into I' by reserving at most λ k Hertz bandwidth in the BS

$$\forall I : I \rightarrow I', \text{ by reserving at most } \lambda \text{ k Hertz in the BS.} \quad (18)$$

Equations (17) and (18) indicate that *BGWA* can be improved to be a $(2 \times D_{max})$-approximation algorithm by reserving at most λ k Hertz bandwidth in the BS. Note that, in the general case, $\lambda \ll B_{limit}$. Thus, in general case, it is worth reserving some bandwidth (at most λ k Hertz) in the BS so that the worst case performance of *BGWA* can be bounded by the approximation ratio of $2 \times D_{max}$. □

Finally, note that the simulations in Section 5 show that the results achieved by *BGWA* are generally much closer to the optimal solution than that indicated by the worst case bound.

5 Simulation Results

5.1 Simulation Environment

This section discusses the performance evaluation of the proposed algorithm. The simulations were conducted using C++. We consider a WiMAX relay network with a single BS controlling five RSs. To determine channel quality, this study adopts the well-known model in [34]. This model states that the channel quality can be determined by the

TABLE 1
The Parameters in the Simulation Environment

Number of BSs	1
Number of RSs	5
Number of SSs	10-150
Modulation scheme for relay link	QPSK,16-QAM, 64-QAM
Modulation scheme for access link	BPSK, QPSK, 16-QAM, 64-QAM
Data rate requirements	64 kbit/s, 128 kbit/s, 192 kbit/s, 384 kbit/s, 768 kbit/s, 2048 kbit/s

location distribution of users. Specifically, the channel quality of a link is inversely proportional to the distance between the BS/RS and the RS/SS, i.e., $qi\ \alpha\ di^{-a}$, where q_i is the channel quality of a link i, d_i is the distance and a is an attenuation factor (usually $2 \leq a \leq 4$). Using this model, the simulations in this study first randomize the locations of the RSs and SSs. Then, according to the random locations, the channel quality of each access link (from BS to SS or from RS to SS) is randomly set as 1, 2, 4, or 6 corresponding to its adopted modulation scheme BPSK, QPSK, 16-QAM, or 64-QAM. On the other hand, since RSs are usually deployed at locations with less interference, the channel quality of each relay link (from BS to RS) is randomly set as 2, 4, or 6 corresponding to the three higher rate modulation schemes QPSK, 16-QAM, and 64-QAM. Note that this channel setting is for convenience during the simulations. The proposed algorithm can also operate under any other channel models. In addition, in accordance with the video levels mentioned in Section 3.1, the data-rate requirement of each SS is randomly set as 64, 128, 192, 384, 768, or 2,048 kbit/s. Table 1 lists the parameters used in these simulations.

5.2 Comparison with Optimal Algorithm

This section compares the proposed *BGWA* algorithm with the optimal algorithm, and simulates *BGWA* for two performance objectives. First, this study simulates $BGWA_{NT}$ for maximizing network throughput while setting its weighted value $W_{m,n} = DR_{m,n}/B_{m,n}$ (see the reasons in Section 3.3.1). Second, this study simulates $BGWA_{SU}$ for maximizing the number of satisfied users while setting its weighted value $W_{m,n} = user_{m,n}/B_{m,n}$ (see the reasons in Section 3.3.2). For comparison, this study also simulates the optimal algorithm for the same two performance objectives. The optimal algorithm applies the brute force method to solve the 0/1 knapsack problem for the two objectives: 1) maximizing network throughput and 2) maximizing the number of satisfied users. That is, the optimal algorithm enumerates all possible combinations of taken objects (i.e., served SSs) to find the optimal solution to the respective maximization problem.

Suppose that the number of SSs varies from 10 to 30 and the limited bandwidth is given as 3 megahertz. Fig. 9a shows that the network throughput in $BGWA_{NT}$ is close to that in the optimal algorithm. Fig. 9b demonstrates that $BGWA_{SU}$ also achieves near-optimal performance in terms of the number of satisfied users. These results indicate that

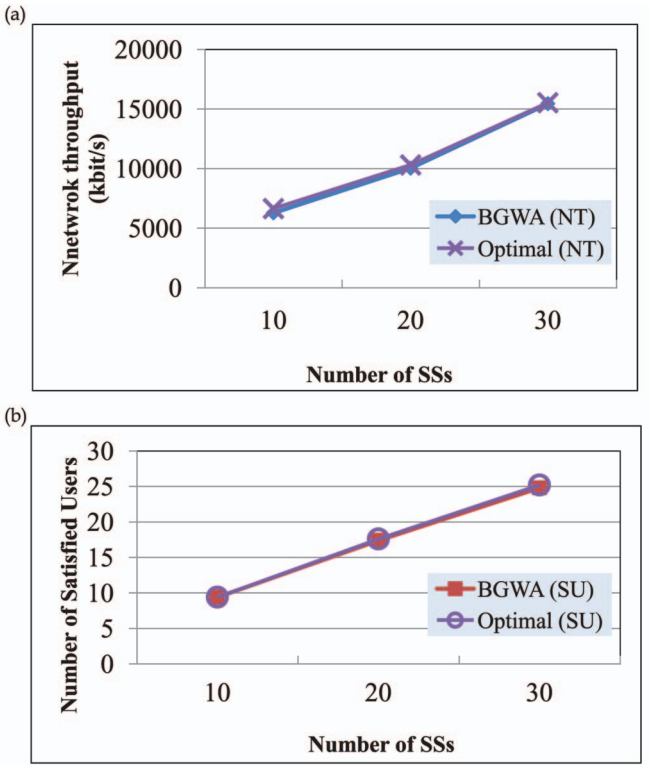

Fig. 9. (a) Network throughput and (b) number of satisfied users for different number of SSs.

the proposed *BGWA* is a good approximation of the optimal solution. Note that the optimal solution is clearly impractical due to its NP-hardness (see the proof in Section 4.1).

5.3 $BGWA_{NT}$ Performance

The following experiments simulate $BGWA_{NT}$ and $BGWA_{SU}$, as mentioned in Section 5.2, while further simulating a naive algorithm for comparison. The naive algorithm greedily allocates the bandwidth to the SSs using a modulation scheme that multicasts the video stream to the maximum number of SSs. The naive algorithm allocates the bandwidth using modulation schemes following the order from the most reliable scheme corresponding to the lowest channel quality (e.g., BPSK) to the least reliable yet fastest scheme corresponding to the highest channel quality (e.g., 64-QAM). This approach ensures that the maximum number of SSs can receive the video stream. Specifically, the naive algorithm first sorts the RSs and SSs into increasing order by channel qualities. Then, following the sorting order, the naive algorithm greedily allocates the bandwidth to the RSs and SSs. Note that the naive algorithm employs no table consulting mechanisms.

Fig. 10a plots the ratio of network throughput to bandwidth consumption as a function of the number of SSs. Because $BGWA_{NT}$ greedily examines the SSs according to their weighted values defined as throughput per bandwidth unit, $BGWA_{NT}$ outperforms $BGWA_{SU}$ and the naive algorithm in terms of the throughput-to-bandwidth ratio. Fig. 10b plots network throughput as a function of the number of SSs. For different numbers of SSs, $BGWA_{NT}$ always achieves higher throughput than $BGWA_{SU}$ and the naive algorithm.

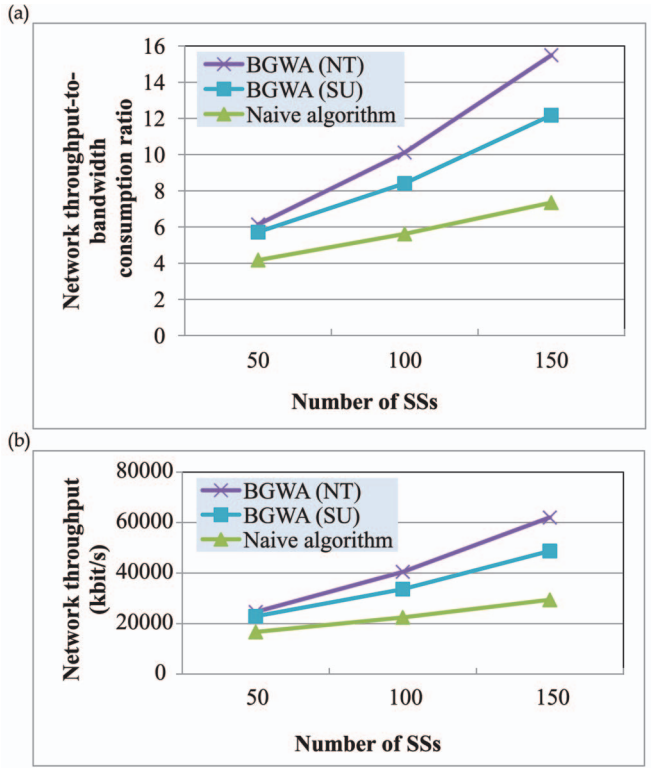

Fig. 10. (a) Network throughput-to-bandwidth consumption ratio and (b) network throughput for different number of SSs.

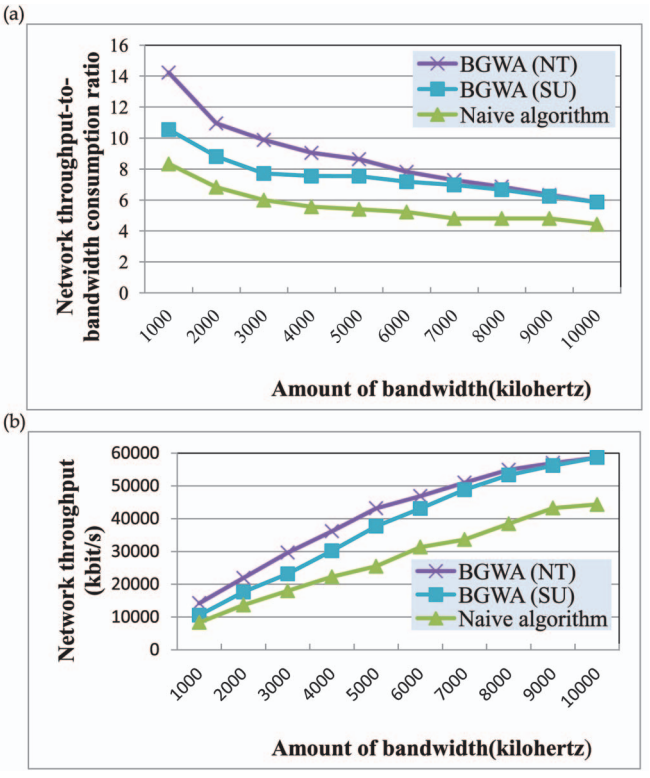

Fig. 11. (a) Network throughput-to-bandwidth consumption ratio and (b) network throughput for different amounts of bandwidth.

Figs. 11a and 11b plot the throughput-to-bandwidth ratio and network throughput as functions of the amount of bandwidth, respectively. These figures show the intuitive results that when the amount of bandwidth increases, the throughput-to-bandwidth ratio decreases while the network throughput increases. In addition, when the amount of bandwidth is small, the curves of $BGWA_{NT}$ are much higher than those of $BGWA_{SU}$ and the naive algorithm. This is because $BGWA_{NT}$ first allocates the bandwidth to the SS with higher ratios of throughput to bandwidth consumption when the bandwidth is insufficient to satisfy all the SSs. This results in higher efficiency of bandwidth utilization. On the other hand, when the bandwidth is large, the curves of $BGWA_{SU}$ approach those of $BGWA_{NT}$. This phenomenon is explained as follows: When the bandwidth is large enough, it is sufficient to satisfy all the SSs. In this case, the throughput performance of $BGWA_{SU}$ is the same as that of $BGWA_{NT}$. Besides, the curves in the two figures indicate that $BGWA_{NT}$ provides higher throughput performance in various scenarios with the bandwidth ranging from 1,000 to 10,000 kilohertz.

5.4 $BGWA_{SU}$ Performance

This section simulates $BGWA_{NT}$, $BGWA_{SU}$ and the naive algorithm described in Section 5.3. The performance of $BGWA_{SU}$ is evaluated as follows: Fig. 12a plots the ratio of satisfied users to bandwidth consumption as a function of the number of SSs. The curves indicate that $BGWA_{SU}$ allocates the bandwidth more efficiently than $BGWA_{NT}$ and the naive algorithm in terms of the number of satisfied users per bandwidth unit. This is because $BGWA_{SU}$ always chooses the SS with the highest user-to-bandwidth ratio for bandwidth allocation in each greedy stage. Fig. 12b plots the number of satisfied users as a function of the number of SSs. As expected, this figure demonstrates that $BGWA_{SU}$ satisfies more users than $BGWA_{NT}$ and the naive algorithm.

Figs. 13a and 13b plot the user-to-bandwidth ratio and the number of satisfied users as functions of the amount of bandwidth. The curves in these two figures show the intuitive results that by increasing the amount of bandwidth, the user-to-bandwidth ratio decreases while the number of satisfied users increases. Furthermore, when the limited bandwidth varies from 1,000 to 10,000 kilohertz, $BGWA_{SU}$ yields more satisfied users and higher user-to-bandwidth ratios than $BGWA_{NT}$ and the naive algorithm. Specifically, the performance gap between $BGWA_{SU}$ and $BGWA_{NT}$ narrows as the amount of bandwidth increases. The explanations for this phenomenon are the same as those for Figs. 11a and 11b in Section 5.3.

6 Conclusions

This study first models the bandwidth allocation problem of scalable video multicast in WiMAX relay networks. We proved that the problems of 1) maximizing network throughput and 2) maximizing the number of satisfied users are both NP-hard. This study provides the polynomial-time suboptimal solution, *BGWA*, to these two NP-hard problems using greedy weighted methods that incorporate table-consulting mechanisms. Instead of enumerating all the possible choices to find the globally optimal solution, the proposed *BGWA* greedily makes the locally optimal choice based on the weighted value. This

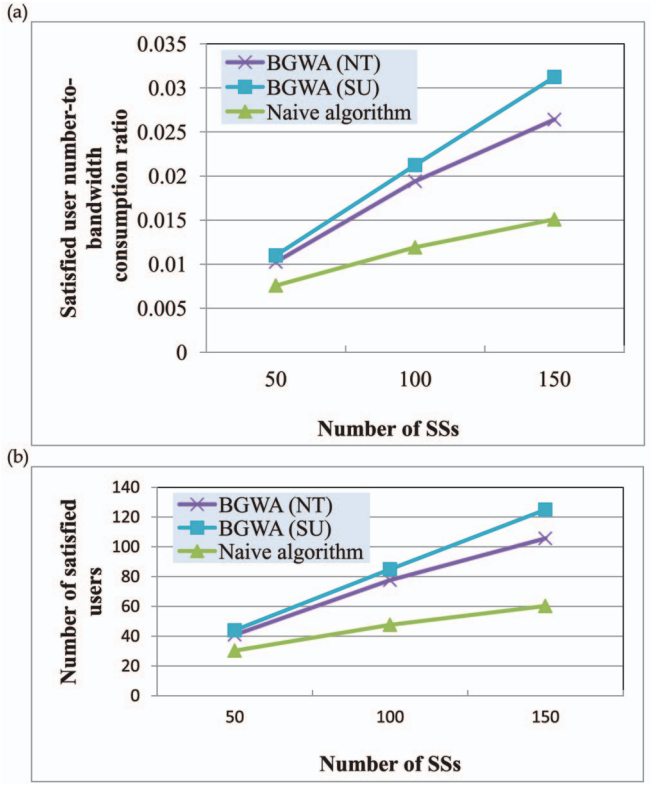

Fig. 12. (a) Ratio of satisfied users to bandwidth consumption and (b) number of satisfied users for different number of SSs.

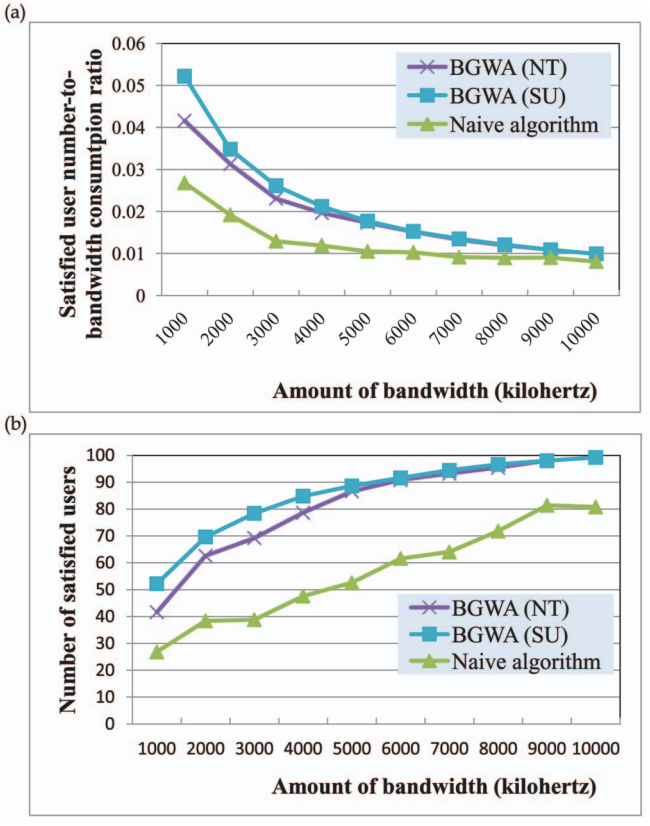

Fig. 13. (a) Number of satisfied users-to-bandwidth consumption ratio and (b) number of satisfied users for different amounts of bandwidth.

approach significantly reduces computational complexity. In addition, by consulting the multicast tables in each greedy stage, the proposed *BGWA* can effectively avoid redundant bandwidth allocation.

To evaluate the performance of the proposed *BGWA*, this study theoretically analyzes the worst case performance of *BGWA*. Theorem 2 shows that *BGWA* is lower bounded by the approximation ratio of $2 \times D_{BS} \times D_{RS}$, where D_{BS} and D_{RS} are the degrees of the BS and RS, respectively. Theorem 3 shows that it is possible to enhance the lower bound of *BGWA* to be the approximation ratio of $2 \times D_{max}$, where $D_{max} = \max(D_{BS}, D_{RS})$. Simulation results for *BGWA* indicate the following. First, the proposed *BGWA* approximates the optimal solution, i.e., the performance of *BGWA* is at least 94 percent of that of the optimal solution, while *BGWA* is much more practical than the optimal (brute-force) algorithm. Second, this study compares the performance of *BGWA* with that of the naive heuristic. In various scenarios with different performance objectives, *BGWA* consistently achieves the highest network performance, i.e., yields the highest network throughput and satisfies the largest number of users, which is consistent with the target objective. These results show that the proposed bandwidth allocation scheme *BGWA* can effectively strike a balance between computational complexity and network performance.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers, whose valuable comments have significantly enhanced the quality of this paper. They would also like to thank Professor Wing-Kai Hon for his help in checking the accuracy of the theoretical analysis in this paper.

REFERENCES

[1] *IEEE 802.16j-2009 Standard, Part 16: Air Interface for Broadband Wireless Access Systems Amendment 1: Multiple Relay Specification*, IEEE, Dec. 2009.

[2] *IEEE 802.16e-2006 Standard, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems - Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, IEEE, Feb. 2006.

[3] C.-W. Huang, P.-H. Wu, S.-J. Lin, and J.-N. Hwang, "Layered Video Resource Allocation in Mobile WiMAX Using Opportunistic Multicasting," *Proc. IEEE Wireless Comm. and Networking Conf. (WCNC)*, pp. 1-6, 2009.

[4] J. She, F. Hou, P.-H. Ho, and L.-L. Xie, "IPTV over WiMAX: Key Success Factors, Challenges, and Solutions [Advances in Mobile Multimedia]," *IEEE Comm. Magazine*, vol. 45, no. 8, pp. 87-93, Aug. 2007.

[5] W.-H. Kuo, T. Liu, and W. Liao, "Utility-Based Resource Allocation for Layer-Encoded IPTV Multicast in IEEE 802.16 (WiMAX) Wireless Networks," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 1754-1759, June 2007.

[6] S.W. Peters and R.W. Heath, "The Future of WiMAX: Multihop Relaying with IEEE 802.16j," *IEEE Comm. Magazine*, vol. 47, no. 1, pp. 104-111, Jan. 2009.

[7] K.-W. Cheng and J.-C. Chen, "Dynamic Pre-Allocation HARQ (DP-HARQ) in IEEE 802.16j Mobile Multihop Relay (MMR)," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 1-6, June 2009.

[8] W.-H. Kuo and J.-F. Lee, "Multicast Recipient Maximization in IEEE 802.16j WiMAX Relay Networks," *IEEE Trans. Vehicular Technology*, vol. 59, no. 1, pp. 335-343, Jan. 2010.

[9] W.-H. Kuo, "Recipient Maximization Routing Scheme for Multicast over IEEE 802.16j Relay Networks," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 1-6, June 2009.

[10] Y.-C. Pan, Y.S. Sun, C. Hsu, and M.C. Chen, "A User-Decided Service Model and Resource Management in a Cooperative WiMAX/HSDPA Network," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 1-6, June 2009.
[11] I. Guvenc, U.C. Kozat, M.-R. Jeong, F. Watanabe, and C.-C. Chong, "Reliable Multicast and Broadcast Services in Relay-Based Emergency Communications," *IEEE Wireless Comm.*, vol. 15, no. 3, pp. 40-47, June 2008.
[12] *Int'l Standard ISO/IEC 14496—10, Information Technology—Coding of Audio-Visual Objects—Part 10: Advanced Video Coding; H.264/AVC*, ISO/IEC, 2004.
[13] D. Marpe, T. Wiegand, and G.J. Sullivan, "The H.264/MPEG4 Advanced Video Coding Standard and Its Applications," *IEEE Comm. Magazine*, vol. 44, no. 8, pp. 134-143, Aug. 2006.
[14] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, July 2003.
[15] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, Sept. 2007.
[16] I. Kofler, R. Kuschnig, and H. Hellwagner, "Improving IPTV Services by H.264/SVC Adaptation and Traffic Control," *Proc. IEEE Int'l Symp. Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1-6, May 2009.
[17] J. Cho and Z.-J. Haas, "On the Throughput Enhancement of the Downstream Channel in Cellular Radio Networks through Multihop Relaying," *IEEE J. Selected Areas in Comm.*, vol. 22, no. 7, pp. 1206-1219, Sept. 2004.
[18] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, 99th ed. Wiley-Interscience, Aug. 1991.
[19] R. Pabst, B.H. Walke, and D.C. Schultz, "Relay-Based Deployment Concepts for Wireless and Mobile Broadband Radio," *IEEE Comm. Magazine*, vol. 42, no. 9, pp. 80-89, Sept. 2004.
[20] G.-M. Su, Z. Han, A. Kwasinski, M. Wu, K.J.R. Liu, and N. Farvardin, "Distortion Management of Real-Time MPEG-4 Video over Downlink Multicode CDMA networks," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 3071-3075, June 2004.
[21] *Int'l Standard ISO/IEC14496-10:2005/Amd.3, Information Technology—Coding of Audio-Visual Objects—Part 10: Advanced Video Coding; Amendment 3 Scalable Video Coding*, ISO/IEC, July 2005.
[22] X. Guo, W. Ma, Z. Guo, X. Shen, and Z. Hou, "Adaptive Resource Reuse Scheduling for Multihop Relay Wireless Network Based on Multicoloring," *IEEE Comm. Letters*, vol. 12, no. 3, pp. 176-178, Mar. 2008.
[23] Y. Shi, W. Zhang, and K.B. Letaief, "Cooperative Multiplexing and Scheduling in Wireless Relay Networks," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 3034-3038, May 2008.
[24] P. Djukic and S. Valaee, "Link Scheduling for Minimum Delay in Spatial Re-Use TDMA," *Proc. IEEE INFOCOM*, pp. 28-36, May 2007.
[25] P. Djukic and S. Valaee, "Delay Aware Link Scheduling for Multi-Hop TDMA Wireless Networks," *IEEE/ACM Trans. Networking*, vol. 17, no. 3, pp. 870-883, June 2009.
[26] T.-W. Kim, T.-Y. Min, and C.-G. Kang, "Opportunistic Packet Scheduling Algorithm for Load Balancing in a Multi-Hop Relay-Enhanced Cellular OFDMA-TDD System," *Proc. Asia-Pacific Conf. Comm. (APCC)*, pp. 1-5, Oct. 2008.
[27] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, and Y.-D. Kim, "Fairness-Aware Radio Resource Management in Downlink OFDMA Cellular Relay Networks," *IEEE Trans. Wireless Comm.*, vol. 9, no. 5, pp. 1628-1639, May 2010.
[28] M.K. Awad and X. Shen, "OFDMA Based Two-Hop Cooperative Relay Network Resources Allocation," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 4414-4418, May 2008.
[29] C.-Y. Hong and A.-C. Pang, "Link Scheduling with QoS Guarantee for Wireless Relay Networks," *Proc. IEEE INFOCOM*, pp. 2806-2810, Apr. 2009.
[30] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, p. 247. W.H. Freeman, 1990.
[31] E. Horowitz and S. Sahni, *Fundamentals of Computer Algorithms*. CS Press, 1984.
[32] U.K. Sarkar, P.P. Chakrabarti, S. Ghose, and S.C.D. Sarkar, "A Simple 0.5-Bounded Greedy Algorithm for the 0/1 Knapsack Problem," *Information Processing Letters*, vol. 42, pp. 173-177, 1992.
[33] S.-M. Huang, C.-W. Huang, P.-H. Wu, J.-N. Hwang, V. Gau, and Y.-C. Chen, "Resource Efficient Opportunistic Multicast Scheduling for IPTV over Mobile WiMAX," *Proc. IEEE Vehicular Technology Conf. (VTC)*, pp. 1-5, May 2010.
[34] T.S. Rappaport, *Wireless Communications: Principles and Practice*. Prenitice-Hall, 1996.
[35] C. Huang, S. Huang, P. Wu, S. Lin, and J. Hwang, "OLM: Opportunistic Layered Multicasting for Scalable IPTV over Mobile WiMAX," *IEEE Trans. Mobile Computing*, vol. 11, no. 3, pp. 453-463, Mar. 2012.

Jang-Ping Sheu received the BS degree in computer science from Tamkang University, Taiwan, Republic of China, in 1981, and the MS and PhD degrees in computer science from National Tsing Hua University, Taiwan, Republic of China, in 1983 and 1987, respectively. He is currently a chair professor of the Department of Computer Science, National Tsing Hua University. He was a chair of the Department of Computer Science and Information Engineering, National Central University from 1997 to 1999. He was a director of Computer Center, National Central University from 2003 to 2006. His current research interests include wireless communications and mobile computing. He was an associate editor of the *IEEE Transactions on Parallel and Distributed Systems*. He is an associate editor of the *International Journal of Ad Hoc and Ubiquitous Computing* and *International Journal of Sensor Networks*. He received the Distinguished Research Awards of the National Science Council of the Republic of China in 1993-1994, 1995-1996, and 1997-1998. He received the Distinguished Engineering Professor Award of the Chinese Institute of Engineers in 2003. He received the K.-T. Li Research Breakthrough Award of the Institute of Information and Computing Machinery in 2007. He received the Y. Z. Hsu Scientific Chair Professor Award in 2009. He is a fellow of the IEEE, a member of the ACM, and Phi Tau Phi Society.

Chien-Chi Kao received the BS degree in computer science and information engineering from National Chung Cheng University, Chiayi, Taiwan, in 2006 and the MS degree in communications engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2008. He is currently working toward the PhD degree with the Department of Computer Science, National Tsing Hua University. In 2006, he joined the Wireless and Mobile Network Laboratory, National Tsing Hua University. His current research interests include wireless communications and mobile computing. He is a student member of the IEEE and an honorary member of the Phi Tau Phi Society.

Shun-Ren Yang received the BS and MSc degrees in computer science and information engineering and the PhD degree from National Chiao Tung University, Hsinchu, Taiwan, R.O.C., in 1998, 1999, and 2004, respectively. From April 1, 2004 to July 31, 2004, he was appointed as a research assistant in the Department of Information Engineering, the Chinese University of Hong Kong. Since August 2004, he has been with the Department of Computer Science and Institute of Communications Engineering, National Tsing Hua University, Taiwan, where he is now an associate professor. His current research interests include design and analysis of mobile telecommunications networks, computer telephony integration, mobile computing, and performance modeling. He is a member of the IEEE.

Lee-Fan Chang received the BS degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2008 and the MS degree in computer science from National Tsing Hua University, Hsinchu, Taiwan, in 2010. His current research interests include wireless communications and mobile computing.