# Energy-Efficient Scalable Video Multicasting for Overlapping Groups in a Mobile WiMAX Network

Zi-Tsan Chou, *Member, IEEE*, and Yu-Hsiang Lin

***Abstract*—Scalable video coding, together with adaptive modulation and coding, is a promising technique for providing real-time video multicast services on mobile stations (MSs) in heterogeneous channel conditions. On the other hand, since MSs are often powered by batteries, reducing the energy consumption of mobile devices is an important concern. In this paper, we study the problem of scalable video multicast scheduling in a mobile worldwide interoperability for microwave access (WiMAX) network with the objective of maximizing the multicast energy throughput. This problem has never been seriously studied when the assumption that different multicast groups may partially overlap is made. We prove that this problem is NP-complete. Our proposed solution, which is called energy-efficient multicast scheduling with adaptive modulation and coding (EEMS-AMC), consists of three modules, namely, online admission control, base-layer data scheduling, and enhancement-layer data scheduling. The admission control rules not only help the base station to admit the maximum number of video streams but ensure that the video data can be delivered to mobile users within the timeliness requirement as well. Moreover, EEMS-AMC adopts the greedy strategy to schedule base-layer data such that even in the worst-case scenario, the average duty cycle of admitted MSs can be within a bounded factor from the theoretical minimum. Finally, EEMS-AMC employs the metric "expected multicast throughput" to derive the proper modulation and coding scheme for each enhancement-layer data and employs the metric "marginal multicast energy throughput" to efficiently derive near-optimal enhancement-layer data scheduling. Simulation results exhibit that EEMS-AMC achieves relatively good performances in terms of multicast energy throughput and normalized total utility.**

***Index Terms*—Adaptive modulation and coding (AMC), IEEE 802.16m, multicast scheduling, NP-complete, power saving, scalable video coding (SVC), worldwide interoperability for microwave access (WiMAX).**

## I. Introduction

THE international standard for wireless metropolitan area networks, i.e., IEEE 802.16 [7], which is now called *fixed worldwide interoperability for microwave access (WiMAX)* [21], provides several quality-of-service (QoS) mechanisms to support high-speed wireless access for fixed subscriber stations. IEEE 802.16e [8], which is also known as *mobile WiMAX* [21], enhances the original standard by additionally addressing the issues of mobility management and power conservation for mobile stations (MSs). The QoS mechanisms specified in 802.16/802.16e include automatic repeat request, various bandwidth request mechanisms (e.g., unsolicited bandwidth grants and bandwidth stealing), and uplink scheduling services for different connection classes (e.g., unsolicited grant service, real-time polling service, extended real-time polling service, nonreal-time polling service, and best effort). For a more detailed description about QoS provision in WiMAX, see [7], [8], [18], [22], [25], and [26]. Clearly, these QoS mechanisms are mainly designed for *unicast* traffic. One of the major applications driving the success of WiMAX is *video streaming* [6], which is based on the ability to simultaneously *multicast* the same video contents from the base station to a group of users, thus reducing bandwidth consumption. To achieve this goal, 802.16e defines the multicast and broadcast services (MBS), and the next-generation WiMAX standard, i.e., 802.16m [9], which is also known as *WiMAX-Advanced* [21], further defines the enhanced-MBS (E-MBS).

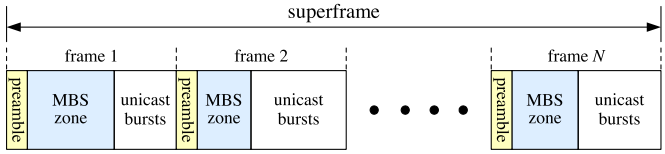


Fig. 1. Simplified superframe structure in the IEEE 802.16e/m TDD system.

### A. Network Model and Problem Statement

Our considered network architecture is similar to the works in [3], [6], [19], and [24]. Specifically, in a WiMAX cell, the video server forwards the video stream to the base station, which then helps broadcast that video data to the multicast members. We assume that the base station and the video server are interconnected by wired networks, which are not the bottleneck; accordingly, we focus only on the transmissions between the base station and the multicast members. Referring to Fig. 1, in the 802.16e time-division duplexing (TDD) system, time is divided into fixed-sized frames. Each frame starts with a preamble, followed by the MBS zone, and then a sequence of unicast bursts. The *MBS zone* is set aside for multicast-only data [8]. In 802.16m, a *superframe* further consists of $N$ frames. Prior to the start of each superframe, the base station broadcasts the E-MBS MAP message, which specifies the schedule of

Manuscript received May 31, 2014; revised December 11, 2014, March 2, 2015, and May 21, 2015; accepted August 6, 2015. Date of publication August 21, 2015; date of current version August 11, 2016. This work was supported by the Ministry of Science and Technology (formerly National Science Council) of Taiwan under Grant NCS 100-2221-E-110-071 and Grant NCS 101-2221-E-110-090. The review of this paper was coordinated by Prof. M. Dianati.

Z.-T. Chou is with the Department of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung 804, Taiwan (e-mail: ztchou@mail.ee.nsysu.edu.tw).

Y.-H. Lin is with the Institute of Information Industry, Taipei 106, Taiwan (e-mail: allenlin0715@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2015.2471302

0018-9545 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

multicast data in the upcoming superframe. On the other hand, since MSs are generally battery powered, reducing the power consumption of mobile devices is an important concern. 802.16e has defined different sleep-mode operations to facilitate energy conservation for MSs. When the sleep mode is enabled, an MS has two states: the awake state and the sleep state. If an MS has requested to receive multicast data, and its request was granted (by the base station), then upon examining the E-MBS MAP, that MS can know which frames (in a superframe) it should wake up to receive its requested data and which frames it could switch to the sleep state [3], [9]. In the sleep state, an MS consumes very low power [5], [8].

The current video codec technologies [16] enable the base station to adjust the video bit rate *on the fly* according to the available bandwidth so that the best possible streaming quality can be achieved in time. Specifically, the *scalable video coding* (SVC) scheme [16] divides a video stream into one *base* layer and one or multiple *enhancement* layers. The base layer that contains the data with the most important features of the video can deliver acceptable quality. The enhancement layers include additional data that can be combined with the base layer to refine the video quality. Definitely, the more video layers an MS has received, the better video quality the mobile user may perceive. On the other hand, 802.16e/m supports a variety of *modulation and coding schemes* (MCSs) and allows these MCSs to be changed on a packet-by-packet basis, depending on the channel conditions. Hence, we can apply different MCSs on different layers of the scalable video sequence such that the users in bad channel conditions receive only base-layer data via a lower physical layer (PHY) rate, whereas the users in good channel conditions can receive additional enhancement-layer data via a higher PHY rate. Such a concept is called *opportunistic* multicast [6], [13].

In this paper, the problems we want to study can be formally stated as follows. Let the multicast group $\mathrm{MS}(s_i)$ be the set of MSs that all operate in the sleep mode and request to see the same video $s_i$ stored in the video server. We assume that the intersection of $\mathrm{MS}(s_i)$ and $\mathrm{MS}(s_j)$ may be *nonempty*, where $s_i \neq s_j$. Moreover, each video is encoded into one base layer and one enhancement layer. Clearly, the maximum MBS zone size of each frame is limited, and the video data have the timeliness requirement. The base station should first determine whether a requested video[1] can be admitted. If so, we must ensure that the base-layer data of that video can be scheduled, and the delay requirement of that video data can be met. Once videos have been admitted, the base station schedules the data of admitted video streams on a superframe-by-superframe basis. Then, the base station wants to determine which video data should be allocated in which frame in a superframe via which MCS such that the *multicast energy throughput* in a WiMAX cell can be maximized. We define the multicast energy throughput as the ratio of the normalized multicast throughput to the average duty cycle of admitted MSs, where the *normalized multicast throughput* [3] is defined as the ratio of the amount of admitted video data *received* by admitted MSs to the amount of admitted video data *requested* by admitted MSs, and the *duty cycle* of an admitted station is the minimum fraction of time during which that MS must stay awake. We call this problem the maximum multicast energy throughput scheduling (MMES) problem.

[1]Note that the admission control function residing in the base station is in charge of deciding whether a *call* should be accepted or rejected. However, in this paper, we focus only on the call that requests to see a video. Thus, for writing convenience and without ambiguity, the phrases "a call is admitted" and "a requested video is admitted" have the same meaning. Note that an MS is called an *admitted* station if at least one of its requested videos is admitted.

### B. Our Contributions

The tradeoff between energy conservation and multicast throughput can be explained as follows: Since the enhancement layer is optional, the base station can schedule only base-layer data of admitted videos. This way, MSs can save more power. However, in such circumstances, all mobile users are unable to enjoy high video quality. In the literature, several video multicast scheduling algorithms for a WiMAX network have been proposed. While some of them try to minimize the total power consumption [24] or maximize the multicast throughput [3], we try to strike the right balance between multicast throughput and energy conservation. While some [11] of them try to maximize the total utility of all users, we believe that the metric "multicast energy throughput" is more objective than "total utility" since different users may have different utility functions.

Specifically, the contributions of this paper are as follows. We formally prove that the MMES problem is NP-complete (see Section III-B). To deal with the MMES problem, we propose a cross-layer solution, which is called energy-efficient multicast scheduling with adaptive modulation and coding (EEMS-AMC), which consists of three components: *online admission control*, *base-layer data scheduling*, and *enhancement-layer data scheduling*. By means of admission control and the restriction of superframe length, EEMS-AMC guarantees that the base-layer data of admitted video streams can be delivered to mobile users in a timely way. Then, EEMS-AMC employs the greedy strategy to schedule the base-layer data such that, in a superframe, the total number of frames in which admitted stations need to wake up to receive their base-layer data is always within a bounded factor from the optimal value. Finally, EEMS-AMC employs the metric "expected multicast throughput" to derive the proper MCS for each enhancement-layer data and adopts the metric "marginal multicast energy throughput" to derive near-optimal enhancement-layer data scheduling. Extensive simulation results show that EEMS-AMC outperforms existing related schemes [2], [3] in terms of multicast energy throughput and normalized total utility.

The rest of this paper is organized as follows. The related work is presented and discussed in Section II. In Section III, we describe EEMS-AMC in detail. Extensive simulations are conducted in Section IV. Section V concludes this paper.

## II. Related Work

IEEE 802.16e defines three types of sleep-mode operations, which correspond to three *power-saving classes* (PSCs), to meet the different characteristics of traffic flow [8]. The PSC of Type I is designed for nonreal-time unicast flow. The PSC of Type II is designed for real-time unicast flow. The PSC of

Type III is designed for management operations, which can be considered to be a supplementary PSC to support the multicast of nonreal-time control packets [5], such as NBR-ADV (neighbor advertisement). Obviously, none of these three PCSs are designed for real-time multicast flow. In the literature, based on the concept of opportunistic multicast, several energy-efficient video scheduling algorithms [11], [13], [19], [24] have been proposed. In [13] and [24], general cellular networks are targeted, whereas those in [11] and [19] are targeted specifically for WiMAX networks. Liu *et al.* in [13] proposed an energy-efficient opportunistic multicast scheduling, in which the base station broadcasts the *scheduled-user set* at the start of each slot. The scheduled-user set is initially empty, and the base station iteratively adds an MS into it according to the priority of an MS until the amount of data that will be received by the scheduled users reaches the predefined threshold. Upon examining the scheduled-user set, the MSs that are not in it can enter the sleep state, thus saving power. In [24], energy-efficient video multicast scheduling with the objective of minimizing the total awake time of MSs is proposed, and it is proven that this problem is NP-hard; then, approximation algorithms are proposed to solve it. However, in [24], Yu *et al.* assume that, in a frame, an MS can immediately switch to the sleep state after receiving its required data, whereas we assume that an MS needs to stay awake during the *whole* frame if it receives data in that frame [3], [8]. In addition, [13] and [24] focus the multicast scheduling on a *single* slot or frame, whereas we focus the multicast scheduling on a *superframe*. In [11], Kao *et al.* proposed two-phase interleaving multicast scheduling algorithms for MSs executing sleep-mode operations of Type-II PSC. The goal in [11] is to derive the optimal parameters of Type-II PSC for each multicast group. In [19], Sharangi *et al.* proposed video multicast scheduling algorithms, whose goal is mainly to maximize the video quality. They reduce the power consumption of MSs by trying to put the base-layer data and the enhancement-layer data of the same video in the same frame. However, such an energy-saving scheme offers no help when the base station admits a lot of video streams and only base-layer data can be scheduled.

It is important that we notice that the algorithms mentioned previously [11], [13], [19], [24] are not applicable to our MMES problem since they tried to optimize energy efficiency *merely* for *nonoverlapping* multicast groups. When different multicast groups partially overlap, reducing the power consumption of one multicast group may raise the power consumption of another multicast group.[2] In fact, the MMES problem remains NP-complete even if the base station schedules only base-layer data of the same size and if the WiMAX system provides only one MCS (see Section III-B and D). In other words, the hardness of the MMES problem stems from one of the key factors that multicast groups may partially overlap. To the best of our knowledge, this paper is the first to seriously address the energy-efficient video multicasting for *overlapping* multicast groups. Although [3] and our previous paper [2] have studied the *best-effort data* multicasting for overlapping multicast groups, the hardness of scalable video multicasting is higher than that of best-effort data multicasting. The reasons are as follows. The base station should first schedule base-layer data since they are mandatory. Assume that multicast groups do not overlap, all base-layer data have been scheduled by a best-effort data multicasting algorithm, and some frames in a superframe still have enough room to accommodate some enhancement-layer data. Since the WiMAX system provides several MCSs and different enhancement-layer data may have different sizes, how to schedule these enhancement-layer data via appropriate MCSs in the residual MBS zones such that a WiMAX cell has maximum multicast energy throughput is similar to the multiple-knapsack problem[3] [4], which is NP-complete, and, thus, suffers higher computational complexity.

## III. Energy-Efficient Multicast Scheduling with Adaptive Modulation and Coding

To deal with the MMES problem, we propose a simple yet novel solution, named EEMS-AMC, which consists of three components: online admission control, base-layer data scheduling, and enhancement-layer data scheduling. Importantly, EEMS-AMC requires that 1) the duty cycle of each MS can be *different* and that 2) each video data can be sent via different MCS in a frame but can be sent only *once* in a superframe.

### A. Online Admission Control

In 802.16, when an MS wants to see to a video stream, it should first send the dynamic service addition request (DSA-REQ) packet to the base station [7]. If that video stream has yet to be admitted, the base station connects to the video server to fetch the information about the data generation rates of that video stream, including the base-layer data rate and the enhancement-layer data rate, which are assumed to be constant [6]. Then, the base station performs the admission control and replies the dynamic service addition response (DSA-RSP) packet telling whether it admits that video stream or not. The following theorem presents the online admission control rules of EEMS-AMC.

*Theorem 1:* Assume that a superframe consists of $N$ frames. Let $F$ be the frame length, $Z$ be the maximum MBS zone size of each frame, $R_{\min}$ be the minimum PHY rate, and $1_{\text{bit\_time}}$ be the time to broadcast one bit using rate $R_{\min}$. Assume that the base station has admitted $n-1$ video streams, which are denoted by $s_1, \ldots, s_{n-1}$, and an MS requests to see a new video

[2]For example, assume that there are four multicast groups. The first group, which is denoted $G_1$, contains three MSs $\{\text{MS}_1, \text{MS}_2, \text{MS}_3\}$, which want to receive multicast data $d_1$. The second group $G_2 = \{\text{MS}_1, \text{MS}_4, \text{MS}_5\}$ wants to receive data $d_2$, the third group $G_3 = \{\text{MS}_4, \text{MS}_5, \text{MS}_6\}$ wants to receive data $d_3$, and the fourth group $G_4 = \{\text{MS}_7, \text{MS}_8\}$ wants to receive data $d_4$. Assume that the MBS zone in a frame can hold only two data. If we hope that the members in $G_1$ can save the energy most, the base station should put $d_1$ and $d_2$ in a single frame and $d_3$ and $d_4$ in another frame. In this case, the total number of frames where MSs need to wake up is 10. However, if the base station puts $d_1$ and $d_4$ in the first frame and $d_2$ and $d_3$ in the second frame, eight MSs, in total, need to wake up only nine frames.

[3]In the multiple-knapsack problem, we are given a set of $n$ items and a set of $N$ knapsacks. Each item $s_i$ has a profit $p(s_i)$ and a weight $w(s_i)$, and each knapsack has a capacity $z_i$. The problem is how to select $N$ disjoint subsets of items and assign them to different knapsacks such that the profit sum of the selected items is maximized under the constraint that the total weight of each subset cannot exceed the capacity of its corresponding knapsack.

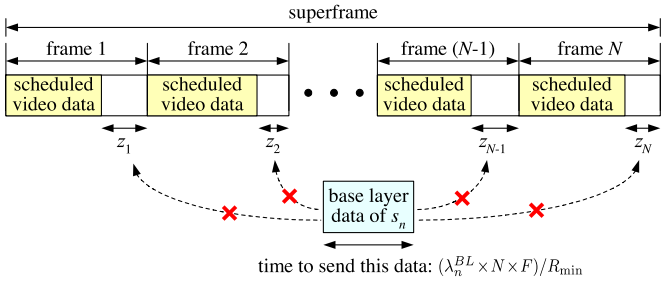


Fig. 2. Scenario where $z_i < (\lambda_n^{BL} \times N \times F)/R_{\min}$ for all $1 \le i \le N$.

stream $s_n$. Let $\lambda_i^{BL} > 0$ be the base-layer data rate of video stream $s_i$, where $1 \le i \le n$. If the following inequalities are all satisfied, the base station admits video stream $s_n$ and ensures that the base-layer data of all admitted videos can be successfully held in a superframe regardless of how these data are scheduled. Otherwise, the base station rejects the request of that MS. Thus

$$\frac{N \times F \times \lambda_n^{BL}}{R_{\min}} \le Z \tag{1}$$

$$\frac{N \times F \times \sum_{i=1}^{n} \lambda_i^{BL}}{R_{\min}} \le Z \times N \tag{2}$$

$$Z \times N - \frac{N \times F \times \sum_{i=1}^{n-1} \lambda_i^{BL}}{R_{\min}} \ge N \times \left( \frac{N \times F \times \max_{1 \le i \le n} \left\{ \lambda_i^{BL} \right\}}{R_{\min}} - 1_{\text{bit\_time}} \right). \tag{3}$$

*Proof:* Since an MS may move, the channel quality between itself and the base station may vary over time [23]. Thus, in the proof, we consider only the worst case where the base station needs to use the most robust MCS (and, hence, the minimum PHY rate) to broadcast all video data to mobile users. Clearly, to watch a scalable video, the base layer is mandatory, whereas the enhancement layer is optional. Thus, if video streams $s_1, \ldots, s_n$ are all admitted, the base station must ensure that it can use the minimum PHY rate to send the base-layer data of these streams within a superframe. Since the superframe length is $N \times F$ and the base-layer data rate of video stream $s_i$ is $\lambda_i^{BL}$, the amount of base-layer data of $s_i$, which is denoted by $\text{size}(s_i^{BL})$, that must be sent in a superframe is $N \times F \times \lambda_i^{BL}$. Inequality (1) is used to ensure that the base-layer data of $s_n$ can be held in a frame, and (2) is used to ensure that the base-layer data of $s_1, \ldots, s_n$ can be held in a superframe.

On the other hand, although EEMS-AMC requires that each video data can be broadcast only once in a superframe, it does not enforce which frame in a superframe the video data must be placed. Hence, referring to Fig. 2, it may happen that after the base-layer data of video streams $s_1, s_2, \ldots, s_{n-1}$ have been scheduled, the base-layer data of $s_n$ cannot be held in any frame even if $z_i > 0$ and $\sum_{i=1}^{N} z_i \ge (\lambda_n^{BL} \times N \times F)/R_{\min}$, where $z_i$ denotes the residual MBS zone size in frame $i$. Note that in Fig. 2, we show only the "MBS zone part" in each frame. To avoid the occurrence of such an unfortunate scenario, inequality (3) must hold.

In what follows, we prove (3) by contradiction. Assume that there exists a scheduling algorithm, called $\mathcal{X}$, which is unable to schedule the base-layer data of some $k$ video streams, which are denoted by $s_{i_1}, s_{i_2}, \ldots, s_{i_k}$, when video stream $s_n$ is admitted, where $\mathcal{S}' = \{s_{i_1}, \ldots, s_{i_k}\} \subseteq \{s_1, \ldots, s_n\}$. Let $\Gamma = Z \times N - (N \times F \times \sum_{i=1}^{n} \lambda_i^{BL})/R_{\min}$. This assumption implies that after the base station performs $\mathcal{X}$, the total residual MBS zone size in a superframe $T_{\text{available}}$ is $\Gamma + \sum_{j=1}^{k} \text{time}(s_{i_j}^{BL})$, where $\text{time}(s_{i_j}^{BL}) = (N \times F \times \lambda_{i_j}^{BL})/R_{\min}$. Furthermore, since the base-layer data of every video stream in $\mathcal{S}'$ cannot be put in any frame in a superframe, this implies that $T_{\text{available}} = \Gamma + \sum_{j=1}^{k} \text{time}(s_{i_j}^{BL}) \le N \times [\text{time}(s_{i_j}^{BL}) - 1_{\text{bit\_time}}]$ for all $1 \le j \le k$. Hence, we have

$$T_{\text{available}} \le N \times \left[ \min_{1 \le j \le k} \left\{ \text{time}\left( s_{i_j}^{BL} \right) \right\} - 1_{\text{bit\_time}} \right]. \tag{4}$$

On the other hand, a video stream will not be admitted if (3) is not satisfied. Since $s_n$ can be admitted, this implies that

$$\begin{aligned} \Gamma &= Z \times N - \frac{N \times F \times \sum_{i=1}^{n} \lambda_i^{BL}}{R_{\min}} \\ &\ge N \times \left( \frac{N \times F \times \max_{1 \le i \le n} \left\{ \lambda_i^{BL} \right\}}{R_{\min}} - 1_{\text{bit\_time}} \right) - \frac{N \times F \times \lambda_n^{BL}}{R_{\min}} \\ &\ge N \times \left( \frac{N \times F \times \max_{1 \le i \le n} \left\{ \lambda_i^{BL} \right\}}{R_{\min}} - 1_{\text{bit\_time}} \right) - \frac{N \times F \times \max_{1 \le i \le n} \left\{ \lambda_i^{BL} \right\}}{R_{\min}} \\ &= (N-1) \times \frac{N \times F \times \max_{1 \le i \le n} \left\{ \lambda_i^{BL} \right\}}{R_{\min}} - N \times 1_{\text{bit\_time}}. \end{aligned} \tag{5}$$

Accordingly, we have

$$\begin{aligned} T_{\text{available}} &= \Gamma + \sum_{j=1}^{k} \text{time}\left( s_{i_j}^{BL} \right) \\ &\ge (N-1) \times \frac{N \times F \times \max_{1 \le i \le n} \left\{ \lambda_i^{BL} \right\}}{R_{\min}} - N \times 1_{\text{bit\_time}} + \sum_{j=1}^{k} \text{time}\left( s_{i_j}^{BL} \right) \\ &> (N-1) \times \min_{1 \le j \le k} \left\{ \text{time}\left( s_{i_j}^{BL} \right) \right\} - N \times 1_{\text{bit\_time}} + \min_{1 \le j \le k} \left\{ \text{time}\left( s_{i_j}^{BL} \right) \right\} \\ &= N \times \left[ \min_{1 \le j \le k} \left\{ \text{time}\left( s_{i_j}^{BL} \right) \right\} - 1_{\text{bit\_time}} \right]. \end{aligned} \tag{6}$$

Clearly, (4) and (6) contradict each other. This implies that the scheduling algorithm $\mathcal{X}$ does not exist. ■

Now, we show how to determine the superframe length such that every newly generated base-layer data of an admitted video stream can be delivered to mobile users within the timeliness requirement. Assume that the tolerable delay of video data is $T_{\text{delay}}$, that is, the video data generated from the video server and cannot be sent within $T_{\text{delay}}$ is considered useless and will be dropped by the base station.

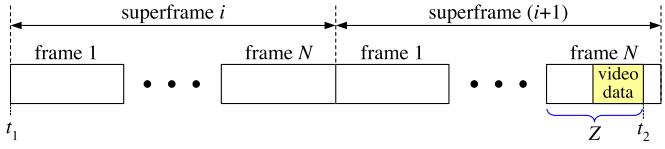


Fig. 3. Relationship between the superframe length and the tolerable delay of video data.

Referring to Fig. 3, assume that the base-layer data $s^{BL}$ of an admitted video stream $s$ is generated at time $t_1+\ell$, where $0<\ell<F$. Unfortunately, prior to this time, the base station has already announced the schedule of multicast data for the $i$th superframe. Since $s$ is admitted, by Theorem 1, we can guarantee that even in the worst case, the base-layer data $s^{BL}$ can be sent before time $t_2$. Thus, by limiting the superframe length $N\times F\le T_{\text{delay}}/2$, we have delay $=t_2-(t_1+\ell)\le 2\times N\times F-(F-Z)<T_{\text{delay}}$.

### B. Problem Hardness

Before presenting our scheduling algorithms to the MMES problem mentioned in Section I-A, we first show its NP-completeness by a reduction from the minimum graph equipartition problem [4], which is known to be NP-complete.

*Theorem 2:* MMES problem is NP-complete.

*Proof:* The minimum graph equipartition problem is defined as follows: Given an integer $N\ge 3$ and a graph $G=(V,E)$ with $|V|=k\times N$, is there a partition of vertex set $V$ into $N$ equal-sized disjoint subsets $V_1,\ldots,V_N$ such that the sum of edges connecting different subsets, which is denoted by $\omega$, is, at most, $W$?

Given an instance $\mathcal{I}$ of the minimum graph equipartition problem, we show how to construct an instance $\mathcal{I}'$ of the MMES problem in polynomial time such that the instance $\mathcal{I}$ has a solution $\{V_i\}_{1\le i\le N}$ that satisfies $|V_i|=k$ and $\omega=W$ if and only if in the instance $\mathcal{I}'$, there exists a scheduling of $|V|$ admitted scalable videos whose multicast energy throughput can be at least $N|E|/(W+|E|)$, where $|E|$ is the number of MSs in the instance $\mathcal{I}'$. The construction procedure is as follows. Let $\lambda_i^{BL}=\lambda^{BL}$ and $\lambda_i^{EL}=\lambda^{EL}$ for all $1\le i\le|V|$. We have $\text{size}(s_i^{BL})=\text{size}(s^{BL})=\lambda^{BL}\times N\times F$ and $\text{size}(s_i^{EL})=\text{size}(s^{EL})=\lambda^{EL}\times N\times F$. Assume that the channel quality of every mobile user is poor such that the base station must use the minimum PHY rate $R_{\min}$ to broadcast all video data. Thus, we have $\text{time}(s_i^{BL})=\text{time}(s^{BL})=\text{size}(s^{BL})/R_{\min}$ and $\text{time}(s_i^{EL})=\text{time}(s^{EL})=\text{size}(s^{EL})/R_{\min}$. Let $Z=k\times[\text{time}(s^{BL})+\text{time}(s^{EL})]$ be the maximum MBS zone size of each frame. By such a configuration, we ensure that all base-layer and enhancement-layer data can be scheduled in a superframe. Let $F=Z+\Delta$ be the frame length, where $\Delta>0$ is the minimum size to hold unicast bursts. Clearly, given $T_{\text{delay}}$, we can configure the values of $\Delta$, $\lambda^{BL}$, and $\lambda^{EL}$ such that $N\times F\le T_{\text{delay}}/2$. Under such circumstances, the delay requirement of every video stream can be met.

Now, we associate a video stream $s_i$ with each vertex $u_i\in V$ and an MS $\text{MS}_i$ with each edge $e_i\in E$. Moreover, given an edge $e_i=(u_{i_1},u_{i_2})$, we assume that $\text{MS}_i$ requests to see two videos $s_{i_1}$ and $s_{i_2}$. Since we have ensured that every MS can receive its requested base-layer and enhancement-layer data, maximizing the multicast energy throughput is, thus, equivalent to minimizing the total number $\mathcal{W}$ of frames in which MSs must stay awake in a superframe. Since the base layer is mandatory, the best way to schedule the data of video streams is to put the base-layer data and the enhancement-layer data of a video in the same frame, thus minimizing the value of $\mathcal{W}$. In a superframe, each MS $\text{MS}_i$ needs to wake up either *once*, if the video data of both $s_{i_1}$ and $s_{i_2}$ are scheduled in the same frame, or *twice*, if the video data of $s_{i_1}$ and $s_{i_2}$ are scheduled in different frames. Let $W$ be the number of MSs that wake up twice in a superframe. We have $\mathcal{W}=W+|E|$. Let $s(F_i)$ be the video streams whose data are placed in the $i$th frame and $s^{-1}(F_i)$ be the set of vertices that associate the video streams in $s(F_i)$. Thus, in the instance $\mathcal{I}'$, if we can schedule $|V|$ admitted videos such that the multicast energy throughput is $N|E|/(W+|E|)$, then the instance $\mathcal{I}$ has the solution $V_i=s^{-1}(F_i)$ that satisfies $|V_i|=k$ and $\omega=W$. On the other hand, let $s(V_i)$ be the video streams that associate the vertices in $V_i$. If the instance $\mathcal{I}$ has a solution $\{V_i\}_{1\le i\le N}$ that satisfies $|V_i|=k$ and $\omega=W$, then in the instance $\mathcal{I}'$, the multicast energy throughput can be at least $N|E|/(W+|E|)$, for example, by letting $s(F_i)=s(V_i)$. ■

### C. MCS for Base-Layer Data

Since the base layer is mandatory, the base station must use the MCS that is robust enough to send the base-layer data of a video such that all multicast members can correctly receive it. We assume that upon receiving a ranging request (RNG-REQ) packet from the base station, an MS has to reply the ranging response (RNG-RSP) packet [7], recording its measured downlink channel quality in terms of signal-to-noise ratio (SNR). Given the value of SNR, the approximate bit error rate (BER) for quadrature phase-shift keying (QPSK) or quadrature amplitude modulation (QAM) found in [15] is given by

$$\text{BER}\approx 4\left(1-\frac{1}{\sqrt{K}}\right)Q\left(\sqrt{\frac{3\times\log_2 K\times\text{SNR}}{c(K-1)}}\right) \quad (7)$$

where $c$ is the code rate, $K$ is the signal constellation size, and $Q(x)=(1/\sqrt{2\pi})\int_x^\infty\exp(-y^2/2)dy$. According to [1], for $0<x<\infty$, the value of $Q(x)$ can be accurately approximated by $Q(x)\approx 1/((1-\rho_1)x+\rho_1\sqrt{x^2+\rho_2})(1/\sqrt{2\pi})e^{-x^2/2}$, where $\rho_1=1/\pi$, and $\rho_2=2\pi$. Assume that the WiMAX system provides $r$ different MCSs, which are denoted by $\text{MCS}_1,\ldots,\text{MCS}_r$, respectively. For convenience, we say that $\text{MCS}_i>\text{MCS}_j$ if the PHY rate of $\text{MCS}_i$ is higher than that of $\text{MCS}_j$. If the base station adopts $\text{MCS}_k$ to send data $d$ to MS $\text{MS}_i$, the probability [or formally called the *packet success probability* (PSP)] that $\text{MS}_i$ can correctly receive $d$ is $\text{PSP}(\text{MS}_i, d|\text{MCS}_k)=[1-\text{BER}(\text{MCS}_k)]^{\text{size}(d)}$, where $\text{BER}(\text{MCS}_k)$ is the BER of $\text{MCS}_k$. Let $\theta_1$ be the tolerable packet loss ratio for the base-layer data. Assume that $h$ MSs, which are denoted by $\text{MS}_{i_1},\ldots,\text{MS}_{i_h}$, request to watch the video stream $s$. Clearly, the maximum MCS that can be used to send the base-layer data of $s$ to $\text{MS}_{i_j}$ is $\text{MCS}_{i_j}^*(s^{BL})=\max_{1\le k\le r}\{\text{MCS}_k|\text{PSP}(\text{MS}_{i_j}, s^{BL}|\text{MCS}_k)\ge 1-\theta_1\}$. However, from (7), we can know that under the same SNR, $\text{BER}(\text{MCS}_i)>\text{BER}(\text{MCS}_j)$ if

```
01  index = 1;
02  while (S^BL ≠ ∅) {    // By Theorem 1, all admitted base layer data can be scheduled.
03    residualMBSzone = Z;
04    s^BL(F_index) = ∅;
05    while (residualMBSzone ≥ min{time(s_i^BL) | s_i^BL ∈ S^BL}) {
06      s_*^BL = arg_{s_i^BL ∈ S^BL} min {|MS(s_i^BL) ⋃ MS(F_index)| | time(s_i^BL) ≤ residualMBSzone};
07      s^BL(F_index) = s^BL(F_index) ⋃ s_*^BL;
08      residualMBSzone = residualMBSzone − time(s_*^BL);
09      S^BL = S^BL \ s_*^BL;
10    }  // end of while
11    index++;
12  }  // end of while
13  return the scheduling result {s^BL(F_i) | 1 ≤ i ≤ N};
```

Fig. 4. Base-layer data scheduling algorithm of EEMS-AMC.

$\text{MCS}_i > \text{MCS}_j$. Since $\text{MS}_{i_1}, \ldots, \text{MS}_{i_h}$ all request to watch the video $s$, the base station should adopt $\text{MCS}^*(s^{BL})$ to broadcast $s^{BL}$, where $\text{MCS}^*(s^{BL}) = \min\{\text{MCS}^*_{i_1}(s^{BL}), \ldots, \text{MCS}^*_{i_h}(s^{BL})\}$.

### D. Base-Layer Data Scheduling

The enhancement layer is optional. When only the base-layer data of admitted videos can be scheduled, the MMES problem still remains NP-complete[4] and under such circumstances, maximizing the multicast energy throughput is equivalent to minimizing the total number $\mathcal{W}$ of frames in which MSs must wake up in a superframe. Let $m$ be the total number of MSs. We observe that

$$\mathcal{W} = \sum_{i=1}^{m} (\text{number of frames in which } \text{MS}_i \text{ must wake up to receive its requested base-layer data}) \quad (8)$$

is equal to

$$\mathcal{W} = \sum_{j=1}^{N} (\text{number of mobile stations that must wake up to receive their base-layer data in frame } j). \quad (9)$$

Hence, the basic idea behind our base-layer data scheduling is that the smaller the number of MSs that must wake up to receive their base-layer data in each frame, the smaller the value of $\mathcal{W}$. To attain this goal, for each frame $F_i$ in a superframe, we iteratively find a yet-to-be-scheduled base-layer data, which is small enough to be placed in the residual MBS zone of frame $F_i$ such that the increased number of MSs that must wake up in frame $F_i$ is minimum.

Let $\mathcal{S}^{BL} = \{s_1^{BL}, \ldots, s_n^{BL}\}$ be the set of base-layer data of video streams $\{s_1, \ldots, s_n\}$ and $\text{MS}(s_i^{BL})$ be the set of MSs requesting to receive the base-layer data of $s_i$. Let $F_i$ denote the $i$th frame in a superframe. Let $s^{BL}(F_i)$ be the set of base-layer data that the base station plans to place in $F_i$ and $\text{MS}(F_i)$ be the set of MSs that must wake up in $F_i$ to receive their requested base-layer data. Fig. 4 presents the base-layer data scheduling algorithm of EEMS-AMC, whose time complexity is $O(n^2)$. Above all, we have the following result.

*Theorem 3:* Under the worst case scenario (i.e., no enhancement-layer data can be scheduled and all base-layer data have to be sent via the minimum PHY rate), the multicast energy throughput of EEMS-AMC is at least $1/\alpha$ times the optimal value, where $\alpha = (Z \times R_{\min})/(\min_{1\le i\le n}\{\lambda_i^{BL}\} \times N \times F)$, and $R_{\min}$ is the minimum PHY rate.

*Proof:* According to the definition of multicast energy throughput, proving this theorem is equivalent to proving that, under the worst case scenario, the value of $\mathcal{W}$ obtained by the base-layer data scheduling of EEMS-AMC is no more than $\alpha$ times the optimal value of $\mathcal{W}$, which is denoted by $\mathcal{W}_{\text{opt}}$. Assume that the optimal schedule of the set of base-layer data $\mathcal{S}^{BL} = \{s_1^{BL}, \ldots, s_n^{BL}\}$ is to place the set of base-layer data $s^{BL}(F_j) = \{s_{j,1}^{BL}, \ldots, s_{j,k_j}^{BL}\}$ in frame $j$, where $1 \le j \le N$, $\sum_{j=1}^{N} k_j = n$, $s^{BL}(F_j) \subseteq \mathcal{S}^{BL}$, and $\bigcup_{j=1}^{N} s^{BL}(F_j) = \mathcal{S}^{BL}$. Obviously, the value of $\mathcal{W}_{\text{opt}}$ satisfies the following inequality:

$$\begin{aligned}\mathcal{W}_{\text{opt}} &\ge \max_{1\le i\le k_1} \left|\text{MS}\left(s_{1,i}^{BL}\right)\right| + \cdots + \max_{1\le i\le k_N} \left|\text{MS}\left(s_{N,i}^{BL}\right)\right| \\ &= \mathcal{W}_{\text{lower\_bound}}. \end{aligned} \quad (10)$$

On the other hand, assume that the worst schedule of the set of base-layer data $\mathcal{S}^{BL}$ is to put the set of base-layer data $\widetilde{s}^{BL}(F_j) = \{\widetilde{s}_{j,1}^{BL}, \ldots, \widetilde{s}_{j,k'_j}^{BL}\}$ in frame $j$, where $1 \le j \le N$, $\sum_{j=1}^{N} k'_j = n$, $\widetilde{s}^{BL}(F_j) \subseteq \mathcal{S}^{BL}$, and $\bigcup_{j=1}^{N} \widetilde{s}^{BL}(F_j) = \mathcal{S}^{BL}$. Let $\phi = \max\{k_1, \ldots, k_N\}$. Since $\bigcup_{j=1}^{N} \widetilde{s}^{BL}(F_j) = \mathcal{S}^{BL} = \bigcup_{j=1}^{N} s^{BL}(F_j)$, the worst value of $\mathcal{W}$, which is denoted by $\mathcal{W}_{\text{worst\_case}}$, satisfies the following inequality:

$$\begin{aligned}\mathcal{W}_{\text{worst\_case}} &\le \sum_{i=1}^{k'_1} \left|\text{MS}\left(\widetilde{s}_{1,i}^{BL}\right)\right| + \cdots + \sum_{k=1}^{k'_N} \left|\text{MS}\left(\widetilde{s}_{N,i}^{BL}\right)\right| \\ &= \sum_{i=1}^{n} \left|\text{MS}\left(s_i^{BL}\right)\right| \\ &= \mathcal{W}_{\text{upper\_bound}} \\ &= \sum_{i=1}^{k_1} \left|\text{MS}\left(s_{1,i}^{BL}\right)\right| + \cdots + \sum_{i=1}^{k_N} \left|\text{MS}\left(s_{N,i}^{BL}\right)\right| \end{aligned}$$

[4]The proof is as follows. Let MMES-B denote the MMES problem where only the base-layer data of admitted videos can be sent. Then, in the proof of Theorem 2, we let $Z = k \times \text{time}(s^{BL})$. Under such circumstances, it can be shown that given an instance $\mathcal{I}$ of the minimum graph equipartition problem, we can construct, in polynomial time, an instance $\mathcal{I}'$ of the MMES-B problem such that the instance $\mathcal{I}$ has a solution $\{V_i\}_{1\le i\le N}$ that satisfies $|V_i| = k$ and $\omega = W$ if and only if in the instance $\mathcal{I}'$, there exists a scheduling of $|V|$ admitted videos whose multicast energy throughput can be at least $[N|E| \times \text{size}(s^{BL})]/[(\text{size}(s^{BL}) + \text{size}(s^{EL}))(W + |E|)]$.

$$\leq k_1 \times \max_{1\leq i\leq k_1} \left|\text{MS}\left(s_{1,i}^{BL}\right)\right| + \cdots + k_N \times \max_{1\leq i\leq k_N} \left|\text{MS}\left(s_{N,i}^{BL}\right)\right| \leq \phi \times \left[\max_{1\leq i\leq k_1} \left|\text{MS}\left(s_{1,i}^{BL}\right)\right| + \cdots + \max_{1\leq i\leq k_N} \left|\text{MS}\left(s_{N,i}^{BL}\right)\right|\right]. \quad (11)$$

Since $\mathcal{W}_{\text{lower\_bound}} \leq \mathcal{W}_{\text{opt}} \leq \mathcal{W} \leq \mathcal{W}_{\text{worst\_case}} \leq \mathcal{W}_{\text{upper\_bound}}$, by (10) and (11), we have

$$\frac{\mathcal{W}}{\mathcal{W}_{\text{opt}}} \leq \frac{\mathcal{W}_{\text{upper\_bound}}}{\mathcal{W}_{\text{lower\_bound}}} \leq \phi. \quad (12)$$

Since assuming that the base station broadcasts only base-layer data via the minimum PHY rate, we have $\phi \leq \alpha$ and $\mathcal{W} \leq \alpha \times \mathcal{W}_{\text{opt}}$. ■

### E. MCS for Enhancement-Layer Data

Since the enhancement layer is optional, the base station can theoretically use the most efficient MCS to send enhancement-layer data, therefore consuming the least bandwidth. However, from (7), we can deduce that the more efficient the MCS used to send the enhancement-layer data, the lower the expected percentage of requesting MSs that can correctly receive it. Hence, to strike the right balance between bandwidth efficiency and multicast reliability, EEMS-AMC employs the metric "expected multicast throughput" to determine the proper MCS for each enhancement-layer data.

Assume that MSs $\text{MS}_{i_1}, \text{MS}_{i_2}, \ldots, \text{MS}_{i_h}$ request to receive the video stream $s$. Let $\text{time}(s^{EL}|\text{MCS}_k)$ be the time for the base station to send the enhancement-layer data of $s$ using $\text{MCS}_k$. Let $\theta_2$ be the tolerable packet loss ratio for the enhancement-layer data. Let $\text{PSP}(\text{MS}_{i_j}, s^{EL}|\text{MCS}_k)$ be the probability that MS $\text{MS}_{i_j}$ can correctly receive $s^{EL}$ using $\text{MCS}_k$. The *expected multicast throughput* of broadcasting $s^{EL}$, which is denoted by $emt(s^{EL}|\text{MCS}_k)$, using $\text{MCS}_k$ is defined by

$$emt(s^{EL}|\text{MCS}_k) = \frac{\text{size}(s^{EL})}{\text{time}(s^{EL}|\text{MCS}_k)} \times \frac{1}{h} \times \sum_{j=1}^{h} \left[\delta\left(\text{MS}_{i_j}, s^{EL}|\text{MCS}_k\right) \times \text{PSP}\left(\text{MS}_{i_j}, s^{EL}|\text{MCS}_k\right)\right] \quad (13)$$

where

$$\delta\left(\text{MS}_{i_j}, s^{EL}|\text{MCS}_k\right) = \begin{cases} 1, & \text{if PSP}\left(\text{MS}_{i_j}, s^{EL}|\text{MCS}_k\right) \geq 1-\theta_2 \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathcal{M} = \{\text{MCS}_1, \ldots, \text{MCS}_r\}$. To maximize the expected multicast throughput of broadcasting $s^{EL}$, the base station in EEMS-AMC thus adopts $\text{MCS}^*(s^{EL})$ to send $s^{EL}$, where

$$\text{MCS}^*(s^{EL}) = \arg_{\text{MCS}_k \in \mathcal{M}} \max \left\{\text{emt}(s^{EL}|\text{MCS}_k)\right\}. \quad (14)$$

### F. Enhancement-Layer Data Scheduling

After scheduling all base-layer data, the base station will proceed to schedule enhancement-layer data only when some frames in a superframe still have enough room to accommodate at least one enhancement-layer data. If such residual MBS zones in a superframe exist, we may be eager to first schedule the enhancement-layer data with maximum expected multicast throughput. However, power consumption is also an important concern with MSs. In other words, the enhancement-layer data scheduling algorithm must also ensure that the number of frames in which MSs need to additionally wake up to receive enhancement-layer data can be minimized. Thus, to strike the right balance between multicast throughput and energy conservation, EEMS-AMC employs the metric "marginal multicast energy throughput" to derive the enhancement-layer data scheduling. Specifically, let $F_i$ denote the $i$th frame in a superframe and $\text{residualMBSzone}(F_k)$ be the residual MBS zone size in frame $k$. If $\text{time}(s_i^{EL})$ is no more than $\text{residualMBSzone}(F_k)$, the *marginal multicast energy throughput*, $\text{mmet}(s_i^{EL}|F_k)$, under the condition that the base station adopts $\text{MCS}^*(s_i^{EL})$ to broadcast $s_i^{EL}$ in frame $k$ is defined by

$$\text{mmet}\left(s_i^{EL}|F_k\right) = \frac{emt\left(s_i^{EL}|\text{MCS}^*\left(s_i^{EL}\right)\right)}{\text{duty\_cycle}\left(s_i^{EL}, F_k\right)} \quad (15)$$

where $\text{duty\_cycle}(s_i^{EL}, F_k)$ denotes the average duty cycle of admitted MSs if the base station places $s_i^{EL}$ in frame $k$. Note that we let the value of $\text{mmet}(s_i^{EL}|F_k)$ be 0 if $\text{time}(s_i^{EL}) > \text{residualMBSzone}(F_k)$. Moreover, assume that an MS $\text{MS}_h$ requested to watch only the video $s_i$, and prior to the start of a superframe, the base station announces that it will broadcast $s_i^{BL}$ using $\text{MCS}^*(s_i^{BL})$ in frame $x$ and $s_i^{EL}$ using $\text{MCS}^*(s_i^{EL})$ in frame $y$; if $x \neq y$ and $\text{MS}_h$ evaluates that $\text{PSP}(\text{MS}_h, s_i^{EL}|\text{MCS}^*(s_i^{EL})) < 1-\theta_2$, then $\text{MS}_h$ will not wake up to receive $s_i^{EL}$ in frame $y$.

Let $\mathcal{F} = \{F_1, \ldots, F_N\}$ and $\mathcal{S}^{EL} = \{s_1^{EL}, \ldots, s_n^{EL}\}$. The gist of our enhancement-layer data scheduling is that, given $\mathcal{S}^{EL}$, we iteratively find an ordered pair $(s_i^{EL}, F_k)$, where $F_k \in \mathcal{F}$, and $s_i^{EL}$ is an unscheduled data in $\mathcal{S}^{EL}$, such that if the base station places $s_i^{EL}$ in frame $k$, the resulting scheduling will have the maximum marginal multicast energy throughput. Let $s^{EL}(F_i)$ be the set of enhancement-layer data that the base station plans to place in frame $i$. Fig. 5 shows the enhancement-layer data scheduling algorithm of EEMS-AMC, which will be executed only after the base station finishes the base-layer data scheduling and whose time complexity is $O(mn^2N)$, where $m$ is the total number of MSs.

### G. Example

Assume that the MBS zone size of each frame is 5 ms and a superframe consists of four frames. Assume that there are seven MSs in a WiMAX cell; moreover, MSs $\text{MS}_2$, $\text{MS}_4$, and $\text{MS}_7$ can receive data using 16-QAM 1/2, whereas MSs $\text{MS}_1$, $\text{MS}_3$, $\text{MS}_5$, and $\text{MS}_6$ can receive data only using QPSK 3/4. The PHY rates of QPSK 3/4 and 16-QAM 1/2 are 11.86 and 15.82 Mb/s, respectively [22]. Assume that these seven MSs totally request to view four video streams, which are denoted

```
01  while ( S^EL ≠ ∅ and min{time(s_i^EL) | s_i^EL ∈ S^EL} ≤ max_{1≤k≤N}{residualMBSzone(F_k)} ) {
02    (s_*^EL, F_*) = arg_{(s_i^EL, F_k)∈(S^EL, F)} max { mmet(s_i^EL | F_k) };
03    s^EL(F_*) = s^EL(F_*) ⋃ s_*^EL;
04    residualMBSzone(F_*) = residualMBSzone(F_*) − time(s_*^EL);
05    S^EL = S^EL \ s_*^EL;
06  }  // end of while
07  return the scheduling result {s^EL(F_i) | 1 ≤ i ≤ N};
```

Fig. 5. Enhancement-layer data scheduling algorithm of EEMS-AMC.

TABLE I
MSs AND THEIR REQUESTED VIDEO STREAMS

| Admitted video stream | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
|---|---|---|---|---|
| $size(s_i^{BL})$ | 23.72 | 23.72 | 23.72 | 23.72 |
| $size(s_i^{EL})$ | 29.65 | 29.65 | 29.65 | 23.72 |
| Stations that requested to see the video | $MS_1$ $MS_5$ | $MS_1$ $MS_2$ | $MS_3$ $MS_6$ | $MS_4$ $MS_7$ |

by $s_1$, $s_2$, $s_3$, and $s_4$. The relations between MSs and their requested video streams are shown in Table I, where the unit of the video packet size is Kbit. Fig. 6 shows the scheduling result of EEMS-AMC. In this example, the multicast energy throughput of EEMS-AMC is $((23.72 \times 10 + 29.65 \times 6)/(23.72 \times 10 + 29.65 \times 6))/(12/(4 \times 7)) = 2.333$.

## IV. PERFORMANCE EVALUATION

### A. Compared Scheduling Schemes

So far, only two *best-effort data* multicast scheduling algorithms, namely, genetic-algorithm-based multicast scheduling (GAMS) [2] and scheduling over multiple broadcast channels with AMC (SMBC-AMC) [3], have been proposed for *overlapping* multicast groups. Hence, we slightly modify GAMS and SMBC-AMC such that they can be also used to schedule scalable video streams. Then, we evaluate the performances of EEMS-AMC and compare our results with those of GAMS and SMBC-AMC.

*1) GAMS:* Genetic algorithms, which mimic the process of natural evolution, are considered as an efficient and effective heuristic scheme to deal with the NP-complete problem [14]. Our previously proposed GAMS [2] is a genetic algorithm and can be modified as follows. It starts from *randomly* generating an initial population of *chromosomes*, which are the feasible solutions to the MMES problem, and these chromosomes evolve over several generations. Specifically, a chromosome $ch_u$ can be defined as a vector $ch_u = \{\pi(s_1^{BL}), \ldots, \pi(s_n^{BL}), \langle \pi(s_1^{EL}), e_1 \rangle, \ldots, \langle \pi(s_n^{EL}), e_n \rangle\}$, where $1 \le u \le U$, $U$ is the predefined population size, and each vector element of $ch_u$ is called *gene*. In a gene $\langle \pi(s_i^{EL}), e_i \rangle$, $\pi(s_i^{EL})$ represents the scheduling order of the enhancement-layer data of admitted video $s_i$, and $e_i$ is a binary variable that indicates whether $s_i^{EL}$ will be scheduled. For example, given a chromosome $\{3, 6, 1, 2, 4, 5, \langle 4, 0 \rangle, \langle 3, 1 \rangle, \langle 2, 0 \rangle, \langle 6, 0 \rangle, \langle 5, 1 \rangle, \langle 1, 1 \rangle\}$, the order that the base station schedules the base-layer data is $s_3^{BL} \to s_4^{BL} \to s_1^{BL} \to s_5^{BL} \to s_6^{BL} \to s_2^{BL}$. More specifically, the base station first puts $s_3^{BL}$ in the first frame and then schedules $s_4^{BL}$; if the MBS zone of the first frame is full after scheduling $s_3^{BL}$, the base station will put $s_4^{BL}$ in the second frame. If there are some unexhausted MBS zones after scheduling all base-layer data, the base station will try to schedule the enhancement-layer data according to the order of $s_6^{EL} \to s_2^{EL} \to s_5^{EL}$. Note that if $e_i = 1$ and the residual MBS zones still have enough room to accommodate $s_i^{EL}$, the base station will adopt $\text{MCS}^*(s_i^{EL})$ shown in (14) to send it.

In each generation, we first calculate the *fitness* of each chromosome $ch_u$, denoted by $fit(ch_u)$, which is defined as the multicast energy throughput of the scheduling $ch_u$. Then, we select a pair of chromosomes for cloning or for mating and crossover. To produce superior offspring a chromosome $ch_u$ is selected as a parent with probability $\gamma(ch_u) = fit(ch_u)/\sum_{u=1}^{U} fit(ch_u)$. The chosen pair of chromosomes has a probability, 1-*crossover_rate*, to be cloned to the next generation, and has a probability, *crossover_rate*, to be mated to produce two offsprings. When a pair of chromosomes, e.g., $ch_u$ and $ch_v$, must be mated, we exchange some genes of either their base-layer parts or their enhancement-layer parts. Fig. 7 shows the crossover procedure that exchanges the genes of the enhancement-layer parts of two chromosomes. Note that the crossover procedure, which exchanges the genes of the base-layer parts of two chromosomes, can be realized via the similar way. First, we generate a random binary string of length $n$, which is called a crossover mask. As shown in Fig. 7(a), if the $i$th bit of the crossover mask is equal to 1, we then swap the $i$th genes of $ch_u$ and $ch_v$. By doing so, we can produce two offsprings $os'_u$ and $os'_v$. However, we notice that $os'_u$ and $os'_v$ shown in Fig. 7(a) are illegal since, for example, 1, 2, and 3 appear twice in $os'_u$. We fix this problem by noting that we made the swaps $4 \leftrightarrow 2$, $6 \leftrightarrow 1$, and $5 \leftrightarrow 3$, and then repeating these swaps on the genes outside the crossover points, giving us the results $os_u$ and $os_v$, as shown in Fig. 7(b). To reduce the chance of falling into local optima, we occasionally mutate the chromosomes of offsprings. Specifically, a new-born offspring is mutated with probability, *mutation_rate*. Once an offspring $os_u$ has to be mutated, we first swap two randomly chosen genes, i.e., $\pi(s_i)$ and $\pi(s_j)$, both of which must be either in the base-layer part or in the enhancement-layer part. If the chosen genes are in the enhancement-layer part, we then randomly decide whether to flip the values of $e_i$ and $e_j$. The above-introduced breeding process is repeated until the size of the new population is equal to $U$, and then, the current population is replaced with the new population. Since we demand that the base station must obtain the multicast scheduling in the duration of a superframe, the apparent advantage of GAMS is that we can control when to terminate the algorithm by stopping the evolution at any time needed and use the best individual in the current generation. The drawback of GAMS is that it cannot provide any performance guarantees.

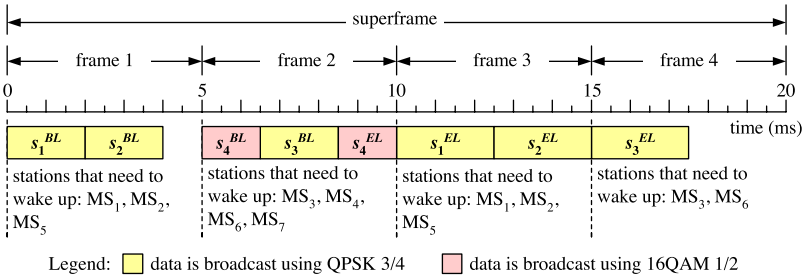


Fig. 6. Example of how EEMS-AMC schedules scalable video streams. Note that we show only the "MBS zone part" in each frame.

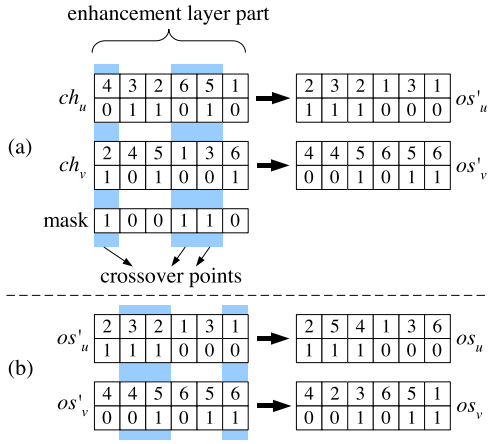


Fig. 7. Example of the crossover procedure, where we show the genes of only the enhancement-layer part in each chromosome.

*2) SMBC-AMC:* The SMBC-AMC, which is originally designed for scheduling best-effort data, works as follows. Assume that the WiMAX system provides $r$ different MCSs. SMBC-AMC requires that all MSs must have the same duty cycle $1/C$, where $C/r$ is an integer. To achieve this goal, as shown in Fig. 8, a superframe is divided into $C$ *logical broadcast channels* such that each logical broadcast channel contains $N/C$ frames. Each MS wakes up only in *one* of the logical broadcast channels. Moreover, SMBC-AMC requires that the base station must use the *same* MCS $\text{MCS}_i$ to send data during the $j$th logical broadcast channel, where $1 \le i \le r$, and $(i-1) \times (C/r) + 1 \le j \le i(C/r)$. The scheduling algorithm of SMBC-AMC consists of two phases: It first decides which logical broadcast channel an MS should listen to, and then, it decides which multicast data should be sent on each logical broadcast channel. Let $\text{MS}(\text{MCS}_i)$ be the set of MSs that can correctly receive data using $\text{MCS}_i$. Let $\text{LBC}(\text{MCS}_i)$ be the set of logical broadcast channels in which the base station must send data using $\text{MCS}_i$. In the first phase, the base station first divides the MSs in $\text{MS}(\text{MCS}_i)$ into $C/r$ groups such that members in the same group requested to receive the *most similar* data. Then, the base station assigns the resulting groups to $\text{LBC}(\text{MCS}_i)$. In the second phase, the base station first calculates the *profit* of each data $d_j$, which is defined as the ratio of the number of MSs that requested to receive $d_j$ to the size of $d_j$. Let $\text{MS}(\text{LBC}_i)$ be the set of MSs that must wake up in the $i$th logical broadcast channel, which is denoted by $\text{LBC}_i$. Then, the base station iteratively puts the highest-profit data requested by $\text{MS}(\text{LBC}_i)$ in $\text{LBC}_i$ until the MBS zones in $\text{LBC}_i$ are full or all data have been scheduled, whichever comes first.

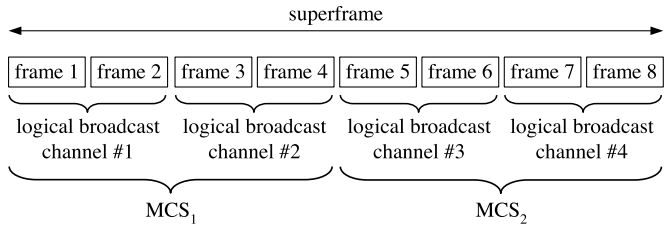


Fig. 8. Concept of logical broadcast channels. In this example, $N = 8$, $C = 4$, and $r = 2$.

Now, we modify SMBC-AMC to schedule scalable video streams as follows. First, we use the SMBC-AMC algorithm to schedule base-layer data. If there are some unexhausted MBS zones in a superframe, we then perform the same algorithm to schedule the enhancement-layer data whose corresponding base-layer data have been already scheduled. Fig. 9 shows the result of applying SMBC-AMC to the example provided in Section III-G. In Fig. 9, the multicast energy throughput of SMBC-AMC is $((23.72\times10+29.65\times3)/(23.72\times10+29.65\times6))/(14/(4\times7)) = 1.571$, which is only 67.3% times that of EEMS-AMC under the same scenario. By carefully observing Fig. 9, we notice that SMBC-AMC may suffer from two problems: 1) *Redundant multicast problem*: If MSs requesting the same data are scheduled in different logical broadcast channels, then these data will be sent several times, hence wasting the scarce bandwidth. For example, in Fig. 9, the base-layer data $s_2^{BL}$ is broadcast twice in a superframe. 2) *Capacity waste problem*: We know that the advantage of SMBC-AMC is that the duty cycle of every MS can be planned in advance. However, as such, some video data may be unable to be scheduled in a superframe, even if there exists at least one frame whose remaining MBS zone is large enough to accommodate one of these video data. For example, the base station cannot adopt QPSK 3/4 to send $s_3^{EL}$ in frame 4 although the residual MBS zone size of frame 4 is still large enough. This is because if the base station did so, the duty cycles of $\text{MS}_3$ and $\text{MS}_6$ would be larger than 1/2. On the other hand, if $\text{MS}_3$ and $\text{MS}_6$ were allowed to wake up in frame 4, and the base station adopted QPSK 3/4 to broadcast $s_3^{EL}$ in frame 4, the multicast energy throughput could rise to $((23.72 \times 10 + 29.65 \times 5)/(23.72 \times 10 + 29.65 \times 6))/(16/(4 \times 7)) = 1.625$.

## B. Simulation Model

We follow an event-driven approach [12] to build the simulator. In our simulations, we adopt the COST-231 Hata model [20], which is suitable for urban environments, as our

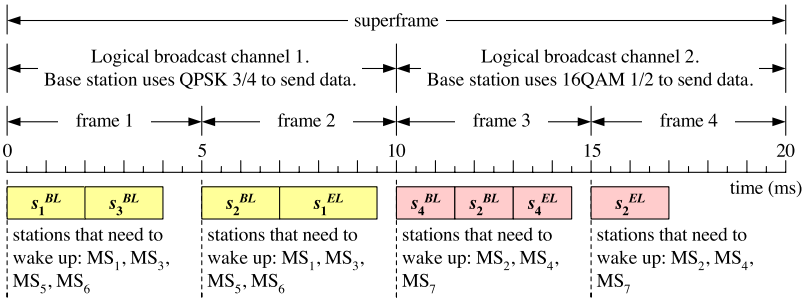


Fig. 9. Example of how SMBC-AMC schedules scalable video streams. Note that we show only the "MBS zone part" in each frame.

TABLE II
SYSTEM PARAMETERS

| Parameter | Value |
|---|---|
| Operating frequency $(f)$ | 2.5 GHz |
| Channel bandwidth $(B)$ | 10 MHz |
| Base station antenna height $(h_b)$ | 32 m |
| Mobile station antenna height $(h_m)$ | 1.5 m |
| Base station antenna gain $(G_b)$ | 15 dBi |
| Mobile station antenna gain $(G_m)$ | $-1$ dBi |
| Noise figure $(N_f)$ | 7 dB |
| Correction factor $(c_m)$ | 3 dB |
| Superframe length | 16 frames |
| Frame length | 5 ms |
| Base station transmit power $(P_b)$ | 43 dB |
| Base station's radius | 1 Km |
| Number of mobile stations | 70 |
| Number of videos in the server | 48 |
| MCSs and corresponding PHY rates | QPSK 1/2 (7.91 Mbps)<br>QPSK 3/4 (11.86 Mbps)<br>16QAM 1/2 (15.82 Mbps)<br>64QAM 2/3 (31.64 Mbps) |

radio propagation model. Specifically, the SNR value of a mobile receiver is formulated as $\text{SNR[dB]} = P_b + G_b + G_m - P_{\text{noise}} - PL(d)$, where $P_b$ is the transmit power of the base station; $G_b$ and $G_m$ are, respectively, the antenna gains of the base station and the MS; $P_{\text{noise}}$ is the receiver noise power; $PL(d)$ is the path loss; and $d$ is the distance between the base station and the MS. $P_{\text{noise}}$ and $PL(d)$ are defined as follows [20]: $P_{\text{noise}} = -174 + 10\log_{10}(B)v + N_f$. $PL(d) = 46.3 + 33.9\log_{10}(f) - 13.82\log_{10}(h_b) - a(h_m) + [44.9 - 6.55\log_{10}(h_b)]\log_{10}(d) + c_m$, where $a(h_m) = 3.2 \times [\log_{10}(11.75h_m)]^2 - 4.97$. Table II shows our used system parameters, which follow the specifications adopted in [20] and [22]. To fairly compare EEMS-AMC with GAMS and SMBC-AMC, we assume that the WiMAX system offers only four different MCSs since SMBC-AMC requires that $C/r$ must be an integer.

We assume that there are four classes of scalable videos in the video server. Table III summarizes the data rate of each layer of each video class, which is based on the experimental results in [17]. We assume that the tolerable delay of video data is 160 ms, and the tolerable packet loss ratio for each layer is 3%. Note that our simulation results show that, with the aid of AMC schemes, the packet loss due to transmission errors can be negligible. In our simulations, the probability that a video is requested by a mobile user (called the *video request probability*) varies from 0.5% to 3.5%. To fairly compare EEMS-AMC with

TABLE III
CLASSES OF VIDEO STREAMS

| Class | Data generation rate (Kbps) | |
|---|---|---|
| | Base layer | Enhancement layer |
| Foreman | 48 | 192 |
| City | 64 | 256 |
| Bus | 96 | 384 |
| Football | 192 | 768 |

GAMS and SMBC-AMC, we assume that all schemes adopt our proposed admission control presented in Section III-A. Note that in GAMS, we assume that both the number of initial chromosomes and the total number of generations are 100; moreover, we set $crossover\_rate = 0.7$ and $mutation\_rate = 0.1$.

Some previous studies on scalable video multicasting [6] used the metric "utility" to represent the satisfaction degree of the video quality for a mobile user. Hence, we also adopt the *normalized total utility* [6] to judge the goodness of a multicast scheduling algorithm. Some notations used to facilitate the definition of normalized total utility are provided below. Let $\gamma(\text{MS}_i, s_j) = 1$ if MS $\text{MS}_i$ requests to see video $s_j$ and $s_j$ is admitted, and $\gamma(\text{MS}_i, s_j) = 0$ otherwise. Let $\xi(\text{MS}_i, s_j^{BL \cup \overline{EL}} | \text{SF}_k)$ be equal to 1 if $\text{MS}_i$ correctly received only $s_j^{BL}$ in the $k$th superframe, and to 0 otherwise. Let $\xi(\text{MS}_i, s_j^{BL \cup EL} | \text{SF}_k)$ be equal to 1 if $\text{MS}_i$ correctly received $s_j^{BL}$ and $s_j^{EL}$ in the $k$th superframe, and to 0 otherwise. Assume that $m$ MSs request to view $n$ admitted videos $\{s_1, \ldots, s_n\}$. The average percentage of users that received *only* base-layer data per average video (denoted by $P_{\text{only}BL}$) and the average percentage of users that received *both* base-layer and enhancement-layer data per average video (denoted by $P_{\text{both}}$) are, respectively, defined as follows:

$$P_{\text{only}BL} = \frac{\sum_{k=1}^{N_{\text{SF}}} \sum_{i=1}^{m} \sum_{j=1}^{n} \xi\left(\text{MS}_i, s_j^{BL \cup \overline{EL}} | \text{SF}_k\right)}{N_{\text{SF}} \times \sum_{i=1}^{m} \sum_{j=1}^{n} \gamma(\text{MS}_i, s_j)} \tag{16}$$

$$P_{\text{both}} = \frac{\sum_{k=1}^{N_{\text{SF}}} \sum_{i=1}^{m} \sum_{j=1}^{n} \xi\left(\text{MS}_i, s_j^{BL \cup EL} | \text{SF}_k\right)}{N_{\text{SF}} \times \sum_{i=1}^{m} \sum_{j=1}^{n} \gamma(\text{MS}_i, s_j)}. \tag{17}$$

Note that $N_{\text{SF}}$ is the total number of superframes during the entire simulation time. The normalized total utility is defined as $u_{BL} \times P_{\text{only}BL} + (u_{BL} + u_{EL}) \times P_{\text{both}}$, where $u_{BL}$ and

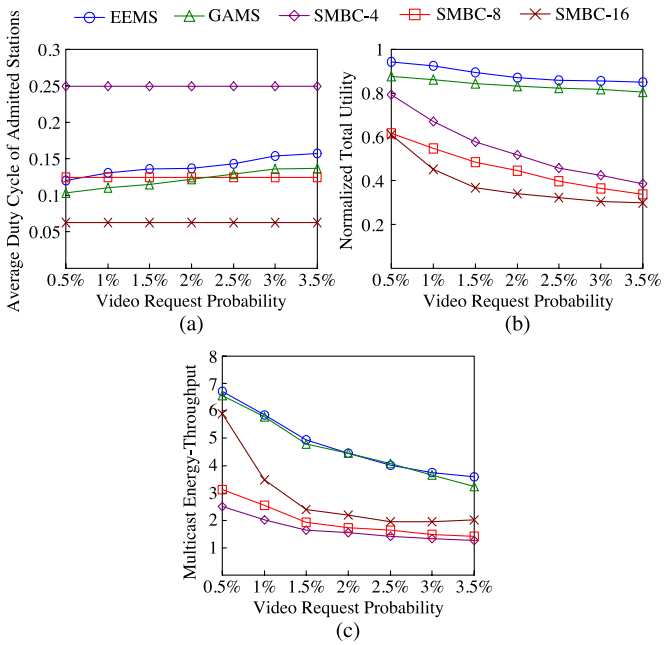


Fig. 10. Performance comparisons among EEMS-AMC, GAMS, and SMBC-AMC under different video request probabilities.

$u_{EL}$ are two positive real numbers that represent the utilities of base-layer data and enhancement-layer data, respectively. Note that we require $u_{BL} + u_{EL} = 1$. In our simulations, we set $u_{BL} = 0.7$, $u_{EL} = 0.3$ [6], and $N_{\mathrm{SF}} = 3.6 \times 10^4$. In the following figures, we use EEMS to denote EEMS-AMC and use SMBC-$k$ to denote SMBC-AMC with $k$ logical broadcast channels.

### C. Effect of Video Request Probability

In the experiments of this section, we vary the video request probability of each mobile user from 0.5% to 3.5%; moreover, we assume that all users are static and that the MBS zone size of each frame is fixed at 4 ms. In Fig. 10(a), we can see that the average duty cycles of EEMS-AMC and GAMS increase with the increase in video request probability. This is because when the video request probability increases, each station will request to see more videos, and thus, more videos may be admitted. This implies that, on average, each station needs to wake up more often to receive the requested video data. We notice that the duty cycle of EEMS-AMC is higher than that of GAMS. This is because EEMS-AMC prefers to schedule the enhancement-layer data with larger expected multicast throughput and, thus, can schedule more enhancement-layer data than GAMS. On the other hand, it can be expected that the average duty cycle of SMBC-AMC is fixed at the reciprocal of the number of local broadcast channels.

Fig. 10(b) exhibits that a larger video request probability has a negative effect on the normalized total utility of all schemes. The reasons are as follows. Since all schemes must schedule base-layer data first and the total MBS zone size in a superframe is fixed, the residual MBS zone that can be used to hold enhancement-layer data hence becomes less as the video request probability increases. We also see that EEMS-AMC has higher normalized total utility than GAMS. This reflects the fact that EEMS-AMC can schedule more enhancement-layer data than GAMS. On the other hand, SMBC-4 has higher normalized total utility than SMBC-16. This is because, in one logical broadcast channel, the total MBS zone size in SMBC-4 is four times that in SMBC-16. Nevertheless, regardless of the number of logical broadcast channels, SMBC-AMC has much lower normalized total utility than EEMS-AMC and GAMS. The reasons are as follows. In a logical broadcast channel, the base station in SMBC-AMC may spend longer time to send the enhancement-layer data than it does in EEMS-AMC and GAMS since it must use the same MCS to send both base-layer data and enhancement-layer data. To make matters worse, SMBC-AMC may also suffer from the redundant multicast problem; that is, the same video data may be sent several times in different logical broadcast channels.

Fig. 10(c) shows that a larger video request probability also has a negative effect on the multicast energy throughput of all schemes. In particular, the multicast energy throughput of SMBC-4 is lower than that of SMBC-16. This is because SMBC-16 is one quarter the average duty cycle of SMBC-4. On the other hand, we see that EEMS-AMC and GAMS have about the same multicast energy throughput. This is because GAMS has a lower duty cycle than EEMS-AMC. Recall that in Section I-B, we have mentioned that the goal of our proposed scheme is not to minimize the duty cycle, which can be accomplished by scheduling only base-layer data of admitted videos. Our goal is to design a scheduling scheme that has high multicast energy throughput. When two scheduling schemes have about the same multicast energy throughput, we prefer that which has higher normalized total utility since normalized total utility represents the average perceived video quality per mobile user [6]. The simulation results shown in Fig. 10(b) and (c) together imply that EEMS-AMC is superior to GAMS.

### D. Effect of MBS Zone Size

In the experiments of this section, we vary the MBS zone size of each frame from 1 to 4 ms; moreover, we assume that all users are static and that the video request probability of each mobile user is fixed at 2%. In Fig. 11(a), we can find that the average duty cycles of EEMS-AMC and GAMS decrease with the decreasing of the MBS zone size in each frame. The reasons are as follows. Although the video request probability is fixed, as the MBS zone size of each frame decreases, less videos requested by mobile users can be admitted; therefore, on average, each MS wakes up more infrequently to receive its requested data. Expectably, the duty cycle of SMBC-AMC is fixed at the reciprocal of the number of local broadcast channels.

Fig. 11(b) shows that the normalized total utility of EEMS-AMC and GAMS can increase as the MBS zone size of each frame increases. However, it seems surprising that a larger MBS zone size has a negative effect on the normalized total utility of SMBC-AMC. The reasons are as follows. In EEMS-AMC and GAMS, the base-layer data of all admitted videos can be completely scheduled. Thus, the normalized total utility of EEMS-AMC and GAMS can be no less than 0.7. After scheduling the base-layer data, the residual MBS zone can be used to schedule the enhancement-layer data. When the MBS zone size

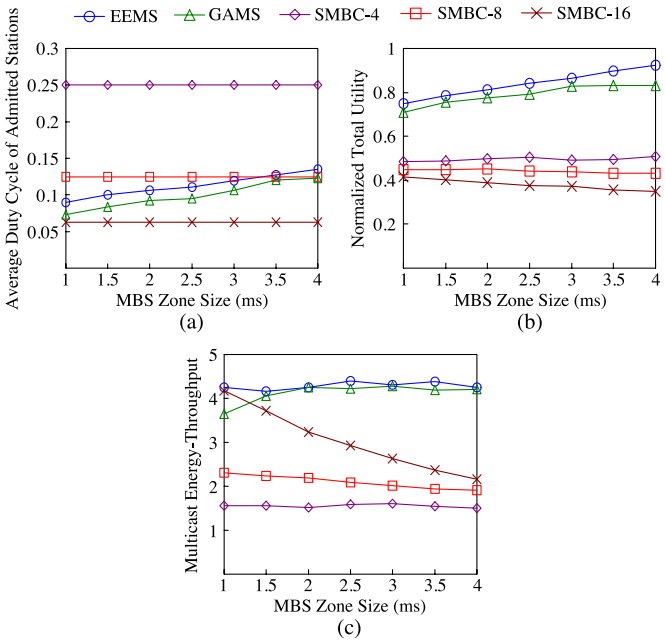


Fig. 11. Performance comparisons among EEMS-AMC, GAMS, and SMBC-AMC under different MBS zone sizes of each frame.

increases, more enhancement-layer data can be scheduled, the normalized total utility of EEMS-AMC and GAMS can hence increase. However, SMBC-AMC suffers from not only the redundant multicast problem but the capacity waste problem as well. In particular, these two problems occur more often when the MBS zone size of each frame increases. On the other hand, since in SMBC-AMC, the total MBS zone size of each logical broadcast channel decreases as the number of logical broadcast channels increases, SMBC-16 has the lowest normalized total utility. Fortunately, the normalized total utility of SMBC-AMC can decline *gradually* as the MBS zone size of each frame increases since the utility of enhancement-layer data is relatively small.

In Fig. 11(c), we can find that the multicast energy throughput of EEMS-AMC and GAMS does not significantly change as the MBS zone size of each frame increases from 1.5 to 4 ms. This is because as the MBS zone size of each frame increases, MSs can receive more requested video data at the expense of a higher duty cycle. On the other hand, Fig. 11(c) shows that the multicast energy throughput of SMBC-AMC decreases as the MBS zone size of each frame increases. The reasons are similar to those described in the previous paragraph. However, we notice that as the MBS zone size increases, the multicast energy throughput of SMBC-16 declines much more steeply than the normalized total utility of SMBC-16. This is because the enhancement-layer data size is generally much larger than the base-layer data size.

### E. Effect of User Distributions

Here, we evaluate the performances of EEMS-AMC, GAMS, and SMBC-AMC under different user distributions. We assume that the video request probability of each mobile user is fixed at 2% and that the MBS zone size of each frame is fixed at 4 ms. In addition, we consider four types of static user distributions: *uniform*, *near*, *middle*, and *far*. The near user distribution assumes that most users are close to the base station, the middle user distribution assumes that the majority of users are in the middle of the base station coverage area, and the far user distribution assumes that most users are far from the base station. Table IV details the percentages of population under different ranges of distances between mobile users and the base station.

TABLE IV
USER DISTRIBUTIONS

| Distance between mobile users and the base station | Percentage of Mobile Users | | |
|---|---|---|---|
| | **Near** | **Middle** | **Far** |
| 0 m ∼ 250 m | 50% | 15% | 5% |
| 250 m ∼ 500 m | 30% | 35% | 15% |
| 500 m ∼ 750 m | 15% | 35% | 30% |
| 750 m ∼ 1000 m | 5% | 15% | 50% |

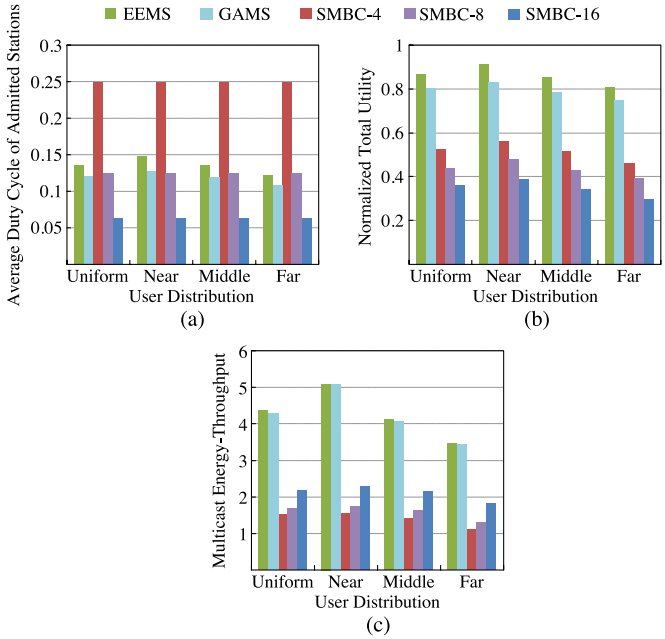


Fig. 12. Performance comparisons among EEMS-AMC, GAMS, and SMBC-AMC under different user distributions.

In Fig. 12(b) and (c), we find that in all schemes, the normalized total utility and the multicast energy throughput under the near user distribution are, respectively, higher than those under the other user distributions. The reasons are as follows. In the near user distribution, most users are close to the base station; under such circumstances, the base station can adopt the more efficient MCS (thus higher PHY rate) to send the majority of base-layer data. This implies that the residual MBS zone size after scheduling the base-layer data becomes larger. Therefore, the base station can schedule more enhancement-layer data.

Fig. 12(a) shows that, as expected, the average duty cycle of SMBC-AMC is equal to the reciprocal of the number of local broadcast channels. Interestingly, we observe that the average duty cycles of EEMS-AMC and GAMS under the near user distribution are, respectively, *slightly* higher than those under the other user distributions. The reasons are as follows. In GAMS, only the feasible schedules that have relatively good performance can survive the evolutionary process. On the other

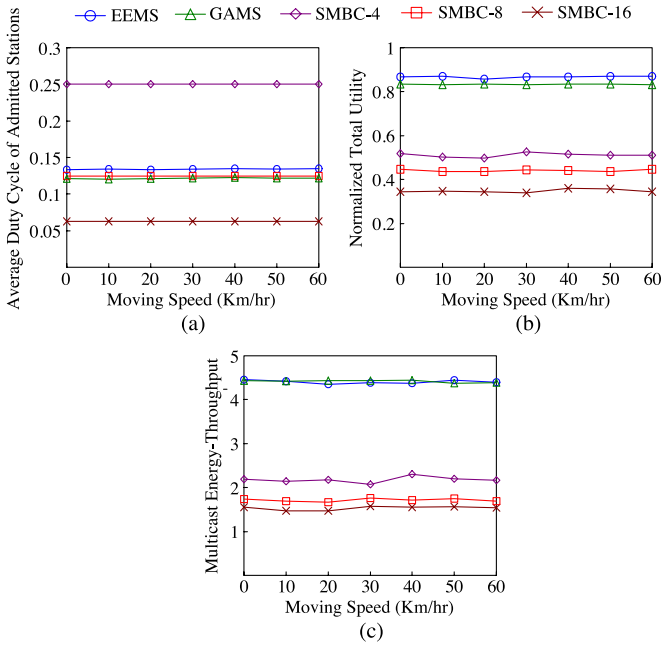


Fig. 13. Performance comparisons among EEMS-AMC, GAMS, and SMBC-AMC under varying moving speeds.

hand, since EEMS-AMC adopts the metric of marginal multicast energy throughput to schedule enhancement-layer data, the increased duty cycle for each admitted station in the near user distribution to receive the extra enhancement-layer data that will not be sent in the other user distributions is expected to be small.

### F. Effect of Mobility

Here, we evaluate the performances of EEMS-AMC, GAMS, and SMBC-AMC under various degrees of user mobility. We assume that the video request probability of each mobile user is fixed at 2% and that the MBS zone size of each frame is fixed at 4 ms; moreover, we consider the random waypoint model [10], in which all mobile users alternate between pausing and then moving to a randomly chosen location (in the coverage of the base station) at a fixed speed. The pause time is fixed at 30 s.

According to [23], the channel quality between the base station and an MS mainly depends on their distance. When mobile users move toward the base station, the base station can adopt the more efficient MCS (thus higher PHY rate) to send data to them. When mobile users move away from the base station, the base station may adopt the more robust MCS (thus lower PHY rate) to send data to them. Since mobile users move sometimes closer to and sometimes farther from the base station, Fig. 13 shows that the performances of all schemes do not significantly change when the moving speed of stations varies.

## V. Conclusion

As mobile devices have grown in popularity, functionality, and capability, people have become more enthusiastic about browsing multimedia programs, such as video streaming or mobile TV anytime and anywhere. Since mobile devices are often battery powered, the investigation of power-saving technologies becomes important. Therefore, in this paper, we studied the problem of scalable video multicasting with the objective of maximizing the multicast energy throughput in a mobile WiMAX network. Unfortunately, we have proven that this problem is NP-complete, and its hardness arises from one of the key factors that different multicast groups may partially overlap. Then, we proposed near-optimal polynomial-time solutions, named EEMS-AMC, to support the multicasting of scalable video streams in a cross-layer fashion. Importantly, by means of online admission control, EEMS-AMC guarantees that all base-layer data of admitted video streams not only can be held in a superframe, regardless of how these data are scheduled, but can also be delivered to mobile users within the timeliness requirement. Furthermore, EEMS-AMC employs different greedy strategies that take multicast throughput and power conservation jointly into account to schedule base-layer data and enhancement-layer data. Theoretical analysis shows that under the worst case scenario, the multicast energy throughput of EEMS-AMC can be always within a bounded factor from the optimal value. Simulation results show that in general scenarios, EEMS-AMC delivers relatively good performances in terms of normalized total utility and multicast energy throughput under uniform, nonuniform, and dynamic user distributions.

## References

[1] P. O. Börjesson and C.-E. Sundberg, "Simple approximations of the error function $Q(x)$ for communications applications," *IEEE Trans. Commun.*, vol. COMM-27, no. 3, pp. 639–643, Mar. 1979.

[2] Z.-T. Chou and Y.-J. Hou, "Genetic algorithm-based energy efficient multicast scheduling for WiMAX relay networks," *Proc. IEEE ICNC*, Feb. 2014, pp. 1061–1065.

[3] R. Cohen, L. Katzir, and R. Rizzi, "On the trade-off between energy and multicast efficiency in 802.16e-like mobile networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 3, pp. 346–357, Mar. 2008.

[4] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA, USA: Freeman, 1979.

[5] C.-H. Hsu, K.-T. Feng, and C.-J. Chang, "Statistical control approach for sleep-mode operations in IEEE 802.16m systems," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4453–4466, Nov. 2010.

[6] C.-W. Huang, S.-M. Huang, P.-H. Wu, S.-J. Lin, and J.-N. Hwang, "OLM: Opportunistic layered multicasting for scalable IPTV over mobile WiMAX," *IEEE Trans. Mobile Comput.*, vol. 11, no. 3, pp. 453–463, Mar. 2012.

[7] *IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems*, IEEE Std. 802.16, Dec. 2004.

[8] *IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, IEEE Std. 802.16e, Dec. 2005.

[9] *IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems. Amendment 3: Advanced Air Interface*, IEEE Std. 802.16m, Mar. 2011.

[10] D. B. Johnson, D. A. Maltz, and J. Broch, "DSR: The dynamic source routing protocol for multihop wireless ad hoc networks," in *Ad Hoc Network*. Reading, MA, USA: Addison-Wesley Inc., 2001, ch. 5, pp. 139–172.

[11] C.-C. Kao, S.-R. Yang, and H.-C. Chen, "A sleep-mode interleaving algorithm for layered-video multicast services in IEEE 802.16e networks," *Elsevier Comput. Netw.*, vol. 56, no. 16, pp. 3639–3654, Nov. 2012.

[12] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 2000.

[13] J. Liu, Y. Zhang, and J. Song, "Energy saving multicast mechanism for scalable video service using opportunistic scheduling," *IEEE Trans. Broadcast.*, vol. 60, no. 3, pp. 464–473, Sep. 2014.

[14] M. Negnevitsky, *Artifical Intelligence: A Guide to Intelligent System*, 2nd ed. Reading, MA, USA: Addison-Wesley, 2005.
[15] Q. Pang, V. C. M. Leung, and S.-C. Liew, "An enhanced autorate algorithm for wireless local area networks employing loss differentiation," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 521–531, Jan. 2008.
[16] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.
[17] C. A. Segall and G. J. Sullivan, "Spatial scalability within the H.264/AVC scalable video coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1121–1135, Sep. 2007.
[18] Y. A. Şekercioğu, M. Ivanovich, and A. Yeğin, "A survey of MAC based QoS implementations for WiMAX networks," *Comput. Netw.*, vol. 53, no. 14, pp. 2517–2536, Sep. 2009.
[19] S. Sharangi, R. Krishnamurti, and M. Hefeeda, "Energy-efficient multicasting of scalable video streams over WiMAX networks," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 102–115, Feb. 2011.
[20] P.-J. Wu, T.-M. Tsai, Y.-C. Kao, J.-N. Hwang, and C.-N. Lee, "An NS2 simulation module for multicast and OFDMA of IEEE 802.16e mobile WiMAX," in *Proc. Asia-Pacific Radio Sci. Conf.*, Taipei, Taiwan, Sep. 2013, pp. 1–5. [Online]. Available: http://tinyurl.com/ne6zmty
[21] WiMAX, 2015. [Online]. Available: http://en.wikipedia.org/wiki/WiMAX
[22] Y. Xiao, *WiMAX/MobileFi: Advanced Research and Technology*. New York, NY, USA: Auerbach, 2007.
[23] D.-Y. Yang, T.-J. Lee, K. Jang, J.-B. Chang, and S. Choi, "Performance enhancement of multirate IEEE 802.11 WLANs with geographically scattered stations," *IEEE Trans. Mobile Comput.*, vol. 5, no. 7, pp. 906–919, Jul. 2006.
[24] Y.-J. Yu, P.-C. Hsiu, and A.-C. Pang, "Energy-efficient video multicast in 4G wireless systems," *IEEE Trans. Mobile Comput.*, vol. 11, no. 10, pp. 1508–1522, Oct. 2012.
[25] S. H. Alsamhi and N. S. Rajput, "An intelligent HAP for broadband wireless communications: Developments, QoS, and applications," *Int. J. Electron. Elect. Eng.*, vol. 3, no. 2, pp. 134–143, Apr. 2015.
[26] S. H. Alsamhi and N. S. Rajput, "An intelligent hand-off algorithm to enhance quality of service in high altitude platforms using neural network," *Wireless Pers. Commun.*, vol. 82, no. 4, pp. 2059–2073, Jun. 2015.

**Zi-Tsan Chou** (S'98–M'10–M'13) received the B.S. degree in applied mathematics from National Chengchi University, Taipei, Taiwan, in 1993; the M.B.A. degree in information management from National Taiwan University of Science and Technology, Taipei, in 1995; and the Ph.D. degree in computer science and information engineering from National Taiwan University, Taipei, in 2003.

He is currently an Associate Professor with the Department of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan. His industrial experience includes affiliations with the NEC Research Institute, Princeton, NJ, USA, and the Institute for Information Industry, Taiwan. He is the holder of eight U.S. patents and 11 Taiwan patents in the field of wireless communications. His research interests include medium access control, power management, and quality-of-service control for wireless networks.

Dr. Chou is a member of the IEEE Communications Society.

**Yu-Hsiang Lin** received the B.S. degree in computer science and information engineering from Tamkang University, New Taipei City, Taiwan, in 2002 and the M.S. and Ph.D. degrees in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2004 and 2015, respectively.

He is currently a Senior Engineer with the Institute of Information Industry, Taipei, Taiwan. His current interests include application development and performance evaluation in Long-Term Evolution-advanced, small cells, and software-defined networking.