Reducing Handover Delay by Location Management in Mobile WiMAX Multicast and Broadcast Services

Ji Hoon Lee, *Student Member, IEEE*, Sangheon Pack, *Member, IEEE*, Taekyoung Kwon, *Associate Member, IEEE*, and Yanghee Choi, *Senior Member, IEEE*

***Abstract*—Mobile Worldwide Interoperability for Microwave Access (WiMAX) includes a multimedia multicast/broadcast service (MBS), but delay-sensitive applications such as video and audio streaming require the combination of efficient handling of wireless-link bandwidth and reduced handover delays, which remains a challenge. To reduce the handover delay in the MBS, the IEEE 802.16e standard introduces an MBS zone, which is a group of base stations that are broadcasting the same multicast packets. However, this raises the MBS traffic load on Mobile WiMAX networks, particularly the wireless part. This paper presents an MBS architecture that is based on location-management areas (LMAs), which can increase the sizes of MBS zones to reduce the average handover delay without too much bandwidth waste. An analytical model is developed to quantify service-disruption time, bandwidth usage, and blocking probability for different sizes of MBS zones and LMAs while considering user mobility, user distribution, and MBS session popularity. Using this model, we also propose how to determine the best sizes of MBS zones and LMAs, along with performance guarantees. Analytical and simulation results demonstrate that our LMA-based MBS scheme can achieve a bandwidth-efficient multicast delivery while retaining an acceptable service-disruption time.**

***Index Terms*—Broadband wireless, broadcast, multicast, multicast/broadcast service (MBS), multimedia streaming, Worldwide Interoperability for Microwave Access (WiMAX).**

I. Introduction

THE RAPID and widespread deployment of broadband wireless networks has raised the expectation of real-time multimedia services in mobile environments. However, supporting bandwidth-intensive multimedia applications requires efficient handling of network resources. When many users want to simultaneously receive the same multimedia content (e.g., news, live sport, or movies), even high-bandwidth wireless-link resources would fall short if a separate point-to-point channel is required for each user. The need for resource- and cost-efficient delivery of multimedia content to many users in parallel has motivated almost all the relevant standardization groups, including the Third Generation Partnership Project (3GPP) and the 3GPP2, to support efficient networkwide multicast and broadcast services [1], [2].

A richer bandwidth fixed wireless network has been recently specified as IEEE 802.16 [3], and this is also extended in IEEE 802.16e [4] to support mobility, sleep/idle mode, and multicast/broadcast service (MBS). As these specifications deal only with the air interface, another industrial forum, i.e., Worldwide Interoperability for Microwave Access (WiMAX), was launched to promote 802.16-based network deployment and to facilitate service environments. Mobile WiMAX [5] provides a synthesis of mobile and fixed broadband access environments through a flexible network architecture and is now being extended to include an MBS architecture and its component protocols.

One crucial issue to be addressed in the MBS architecture is the provision of seamless multimedia streaming to mobile receivers [6]. That is, a mobile station (MS) should be able to receive a multimedia stream without noticeable disruption while it is moving across cells, even though a handover process is required when an MS moves from one cell of a cellular network to another. During the handover process, the path to the MS is transferred from the serving base station (BS) to the target BS, and the time required for this process is referred to as the handover delay. Mobile WiMAX (or IEEE 802.16e) normally performs hard handovers, in which all connections to the serving BS are broken before new connections are made to the target BS. As a result, packets that are being sent through the serving BS during the handover may not be delivered to the MS. In the case of unicast traffic, this so-called "service disruption" can be overcome by packets that are being stored and forwarded from the serving BS to the target BS for a fast handover and retransmission [7], [8]. However, MBSs cannot rely on such techniques since packets are destined for multiple receivers. Therefore, the handover delay in the MBS should be minimized.

To reduce the handover delay in the MBS, the IEEE 802.16e standard introduces an MBS zone [4], which is a group of adjacent BSs that are transmitting the same content, such as a video stream or an audio stream. A handover between BSs in the same MBS zone involves a reduced delay because packets with the same content will be received from the target BS immediately after the completion of a link-level handover.

Manuscript received October 28, 2009; revised March 3, 2010, June 20, 2010, and September 28, 2010; accepted November 18, 2010. Date of publication December 6, 2010; date of current version February 18, 2011. This work was supported in part by the Information Technology Research and Development program of the Ministry of Knowledge Economy/Korea Evaluation Institute of Industrial Technology under Grant 10035245 [Study on Architecture of Future Internet to Support Mobile Environment and Network Diversity] and in part by the World Class University program through the National Research Foundation under Grant R33-2008-000-10044-0. This paper was presented in part at the 2008 IEEE International Conference on Communication System Software and Middleware, Bangalore, India, January 2008. The review of this paper was coordinated by Prof. V. W. S. Wong.

J. H. Lee, T. Kwon, and Y. Choi are with the School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea.

S. Pack is with the School of Electrical Engineering, Korea University, Seoul 136-701, Korea.

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.2010.2096831

However, a handover that crosses a boundary between different MBS zones requires not only link-level handover signaling but MBS-related signaling as well, which takes a much longer time.

Obviously, larger MBS zones will yield better quality of service for a given level of mobility, but the handover-delay cutback comes at the cost of the amplified traffic. This wastes the link capacity of the air interface,[1] because every BS in the same MBS zone broadcasts the same MBS packets, regardless of the presence of a user in its coverage; moreover, requesting a new MBS session can be blocked due to lack of available bandwidth. Seeing that the service disruption and the wireless-bandwidth usage are necessarily conflicting in the MBS-zone planning, we were motivated to study how to determine the best size of the MBS zones.

In this paper, we propose an MBS architecture based on location-management areas (LMAs), each of which is a set of geographically adjacent BSs within an MBS zone. Then, the multicast and broadcast packets only need to be transmitted to the LMAs that have MBS users that are reducing the requirement for the wireless-link bandwidth. Using LMAs allows large MBS zones to be used, so that the number of inter-MBS-zone handovers can be reduced; in this way, we can reduce the average handover delay without too much bandwidth waste. We analyze the performance of our LMA-based MBS scheme (which is denoted by LMS) by means of an analytical model for different sizes of MBS zones and LMAs. We go on to propose how the sizes of the MBS zones and the LMAs can be determined while retaining an acceptable service-disruption time. To the best of our knowledge, this is the first comprehensive analytical study for MBS-zone planning; our model of the service-disruption time, the bandwidth usage, and the blocking probability in terms of user mobility, user distribution, and session popularity is novel, as is our approach to determining the sizes of the MBS zones and LMAs.

The rest of this paper is organized as follows: In Section II, we summarize the related works in the literature and highlight the major differences between the existing works and our paper. Section III presents our LMA-based MBS-zone-planning scheme and explains how it affects handover delay, and a performance analysis follows in Section IV. Numerical and simulation results are presented in Sections V and VI, respectively, and Section VII concludes this paper.

II. Related Work

The universal mobile telecommunications system (UMTS) multicast architecture [9] employs standard Internet Protocol (IP) multicast protocols. A multicast mechanism for UMTS has been proposed [10], which establishes multicast tunnels throughout the UMTS network that allow multicast packets to be transferred on shared links toward multiple destinations. The tradeoffs between the broadcast, multiple-unicast, and multicast approaches for one-to-many packet-delivery services in the UMTS have been investigated [11]. However, this work lacks an analysis of user-mobility handling; it is assumed that mobility is handled by standard UMTS mobility mechanisms, which are similar to conventional unicast-packet forwarding. As an alternative [12], routing lists can be introduced into the nodes of the UMTS to support resource-efficient multicast transmissions that are combined with a reassessment of the handover types and the mobility-management mechanism in the UMTS. However, multicast-service continuity still cannot be assured unless handover-delay issues are taken into account.

There has been some research on the support of real-time services such as voice over IP and video streaming (IP television) over WiMAX [13], [14]. The hard handovers that are mandated by IEEE 802.16e make seamless mobility with imperceptible interruption of service difficult to achieve in Mobile WiMAX. A fast handover scheme has been proposed [15], along with a new transport connection-identifier mapping strategy for real-time applications to reduce handover delay and the probability of packet loss. This approach could be an option for unicast services (e.g., video on demand), but it is not suitable for multicast and broadcast services. The efficient delivery of video broadcasts over WiMAX has been studied [16], particularly the issue of synchronous transmission over multiple BSs. The effectiveness of data delivery in intra-MBS-zone operations is shown to be improved by macrodiversity [17]; hence, seamless handover is also feasible. However, this work lacks an analysis of the effects of the various MBS-zone sizes, and the inter-MBS-zone scenario is not considered at all. Deploying large MBS zones may be impractical in the original MBS scheme (OMS) because too much bandwidth is wasted. Moreover, macrodiversity requires that the whole MBS zone should be a single-frequency network, which limits the size of the MBS zones even more tightly.

To the best of our knowledge, our paper is the first comprehensive analytical study of MBS-zone planning, which has the aim of reducing both bandwidth usage and service disruption. Nevertheless, some previous studies have addressed the issues of network planning for wireless multicast and broadcast services. An efficient multicast mechanism for heterogeneous wireless networks has been proposed [18], which reduces the total bandwidth requirement of the IP-multicast tree by adaptively selecting the cell and the wireless technology for each MS. Although this is not suitable for multicast services in a homogeneous wireless network, it does allow more MSs to cluster together and leads to the use of fewer cells, thus saving bandwidth. A network operator might use this approach for network planning in small fixed or nomadic wireless networks, but it is impractical for large mobile wireless networks in which the network frequently needs to recompute its low-cost multicast tree due to mobility. In another multicasting mechanism for UMTS [19], multicast packets are distributed to location areas (LAs), which are groups of cells. When an MS moves between LAs, the change is reported to a location server so that the position of the MS is tracked for paging purposes. This scheme's primary concern is with the delivery of short messages to multiple users in these LAs to minimize paging cost, but the solution that is proposed includes the location tracking of multicast receivers, which is relevant to our paper.

[1] An MBS zone will also waste the link bandwidth of the wired-backhaul network, but we focus on the wireless link.

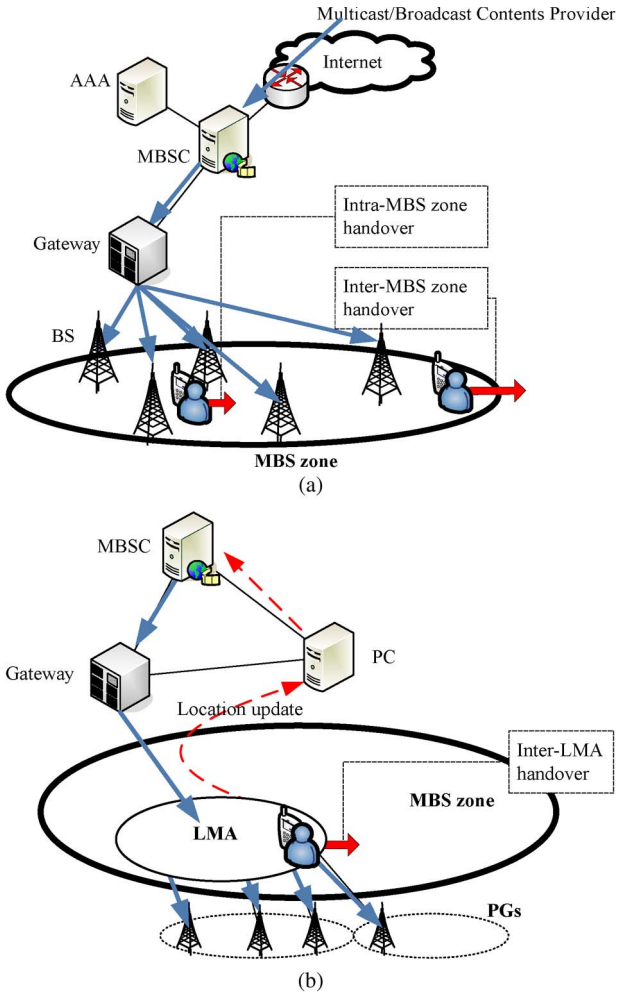

Fig. 1. Mobile WiMAX MBS. (a) Original MBS. (b) LMA-based MBS.

III. Location-Management Area-Based Multicast-Broadcast Service

A WiMAX network typically consists of MSs, BSs, and gateways with an authentication, authorization, and accounting server [5]. To support the MBS, several BSs (which are normally adjacent to each other) construct an MBS zone between them, as shown in Fig. 1(a), which is managed by the multicast- and broadcast-service controller (MBSC). Once the MBS-zone sizes are determined (see Section III-C), the MBS zones can be deployed by a network operator, and all BSs in an MBS zone have a shared multicast connection for the same multicast transmission.[2] Therefore, an MS does not need to create a new connection during handovers between BSs in the same MBS zone, which reduces the handover delay. Accordingly, increasing the sizes of the MBS zones will reduce service disruptions for a given level of mobility, while the wireless-link bandwidth will be wasted in the sense that every BS in the same MBS zone broadcasts the same MBS packets, regardless of the presence of a user. This also results in increasing the probability of blocking a new MBS session.

[2]A BS transmits its MBS-zone identifier(s) using the IEEE 802.16 downlink channel descriptor [4].

A. LMA

To efficiently balance the tradeoff between the bandwidth usage and the service disruption in the MBS, we define an LMA as a set of geographically adjacent BSs, which is used to track the location of MBS users. That is, the network is always aware of the current LMA of each MBS user. Then, we introduce an LMA-based MBS that partitions an MBS zone into multiple LMAs and then selectively transmits packets to the LMAs in which MBS users currently reside. Using LMAs decouples the requirement for the wireless-link bandwidth from the size of an MBS zone. This allows for large MBS zones to be used so that service disruptions will be reduced, as shown in Fig. 1(b).

Our scheme relies on the network that is keeping track of the location of every ("MBS-enabled") MS at the granularity of an LMA. In IEEE 802.16e, a paging group (PG) [4], which is analogous to the LA in cellular networks, is used to track the locations of MSs. A paging controller (PC) manages information that tracks which MSs are currently located in each PG. An LMA might correspond to one or more PGs, but it is possible that the coverage of an LMA is determined independently of those of PGs.

The location of an MS in the normal mode is tracked from one BS to another by the PC. Whenever an IEEE 802.16 media-access-control (MAC)-layer handover is done, the target BS reports this event to the PC, which therefore knows the current BS and PG of each MS. In the idle mode,[3] however, the location of an MS is handled at a coarser level. An MS, even in the idle mode, can acquire the PG information of the current and the neighboring cells as long as it receives MBS packets from a BS. Therefore, each time an idle MS crosses a PG boundary, it can inform the PC of its new PG. To flexibly determine the coverage of an LMA, we suggest that every BS should broadcast an LMA information, as well as the PG information.[4] This enables an MS in the idle mode to update its location upon every inter-LMA movement. Since the MBSC controls which LMAs will transmit MBS packets, it needs to receive the locations of the MBS users from the PC.

B. Handover Delays

The handover of an MBS session involves delays due to the link-level messages that are exchanged during the IEEE 802.16e handover and due to the MBS signaling messages. The former delay occurs whenever an MS with an ongoing MBS session switches to a new BS, irrespective of its MBS zone or LMA. However, the latter delay only occurs when an MS moves from one MBS zone to another. First, MBS handovers can therefore be classified into intra-*MBS-zone handovers* and inter-*MBS-zone handovers*.

Fig. 2(a) shows the sources of the inter-MBS-zone handover delay and describes how MBS signaling messages are

[3]This mode is defined in IEEE 802.16e to conserve power and network resources. An MS in the idle mode performs no handover when it crosses a cell boundary, but it can receive MBS packets [4].

[4]An LMA information element can be easily added as an optional field that is encoded by a type–length–value element of the IEEE 802.16 MAC protocol. It does not break the backward compatibility.

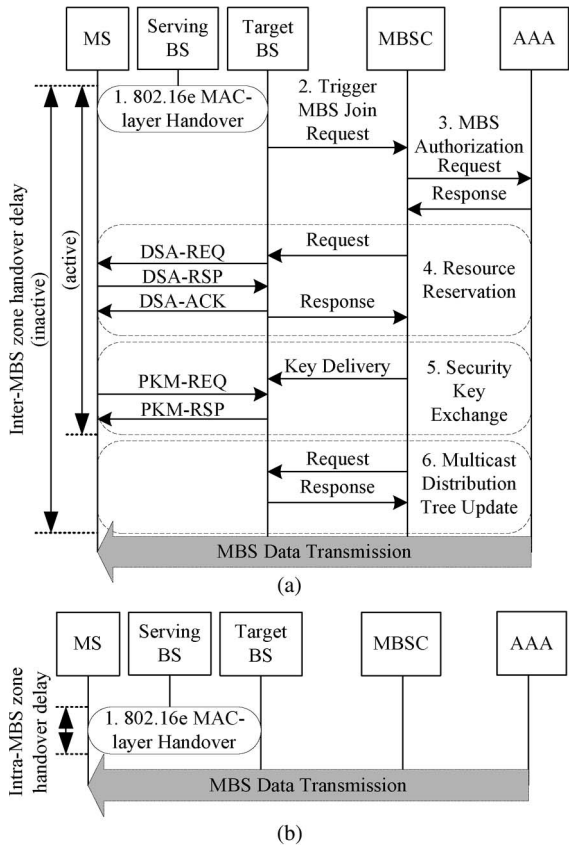

Fig. 2. MBS-zone handovers. (a) Inter-MBS-zone handover. (b) Intra-MBS-zone handover.

exchanged after the MAC-layer handover (step 1). Whenever an MS moves from one MBS zone to another, a new connection is needed, which is triggered by an MBS join request (step 2). This is followed by the MBS authorization procedure (step 3), resource reservation (step 4), security key exchange (step 5), and multicast-distribution tree update (step 6). However, if the MS is only moving from one BS to another within the same MBS zone, no additional processing is required, except for the IEEE 802.16e MAC-layer handover, as shown in Fig. 2(b). The multicast-distribution-tree-update procedure (step 6) may not be necessary for an inter-MBS-zone handover, depending whether there are users of the same MBS session in the target MBS zone. Note that the MBS zones that currently contain users of the current MBS session are called *active* MBS zones, whereas those without such users are called *inactive* MBS zones. If the target MBS zone is already active, the multicast-distribution tree does not have to be updated.

In an LMA-based MBS, intra-MBS-zone handovers are subclassified into intra-LMA handovers, which result from a change of the BS within the same LMA, and inter-LMA handovers between BSs in different LMAs. Fig. 3 shows the signaling messages required for inter-LMA handovers. LMAs containing current users of a particular MBS session are called *active* LMAs with a remainder called *inactive* LMAs, with respect to that session. In Fig. 3(a), the PC maintains up-to-date location information about an MBS user by the location update or as a result of a MAC-layer handover to a target BS (steps 1 and 2). The PC informs the MBSC of the new location

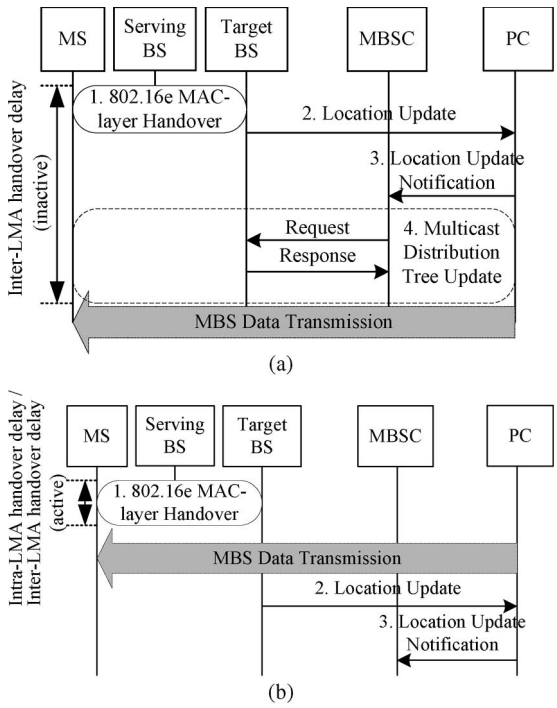

Fig. 3. Inter-LMA handovers. (a) To an inactive LMA. (b) To an active LMA.

of an MS so that the MBSC can decide whether the MS has crossed an LMA boundary or not (step 3). If the MS has moved into an inactive LMA, multicast-distribution tree update (step 4) must be done. If the LMA is active [see Fig. 3(b)], the time required for steps 2 and 3 does not contribute to the handover delay, which only consists of the time required to finish step 1. The delay involved in an intra-LMA handover is equivalent to that resulting from an intra-MBS-zone handover.

C. LMA-Based MBS-Zone Planning

Seeing that the service disruption and the wireless-bandwidth usage are necessarily conflicting when deciding the sizes of the MBS zones, we now formally define the MBS-zone-planning problem. Let N_Z be the number of cells in an MBS zone and N_L be the number of cells in an LMA. We present a method that determines the values of N_Z and N_L, which minimize the bandwidth usage while keeping the average handover delay below a specified value.

Let Y be the random variable that is expressing the length of a handover delay and D_{th} be a threshold value of the previously mentioned handover delay in which an MBS user experiences a noticeable session disruption. The probability that receiving session i is disrupted, such that Y exceeds D_{th}, can be written as $\Pr[Y > D_{\text{th}}]_i$. How to compute the probability will be elaborated in Section V-B.

The problem of the MBS-zone planning for session i can now be considered as a search problem for values of N_Z and N_L, which satisfy $\Pr[Y > D_{\text{th}}]_i \leq \delta$, where δ is the tolerable disruption ratio of session i.[5] We assume that a service

[5]Different sessions may have different values of N_Z and N_L, and different MBS zones can be overlapped.

Assumption: $N_{Z,max} \geq N_{Z,min} \geq N_{L,min} \geq 1$;
$N_L = N_{L,min}$;
while $N_L \leq N_{Z,max}$ **do**
  $N_Z = max\{N_{Z,min}, N_L\}$;
  while $N_Z \leq N_{Z,max}$ **do**
    if $Pr[Y > \gamma]_i \leq \delta$ **then**
      return (N_Z, N_L);
    end if
    increase N_Z;
  end while
  increase N_L;
end while

Fig. 4. Algorithm to determine N_Z and N_L.

provider enforces the MBS zone and the LMA-size constraints by introducing the following min/max bounds: $N_{Z,\min}$ is the minimum number of cells in an MBS zone; $N_{Z,\max}$ is the maximum number of cells in an MBS zone; and $N_{L,\min}$ is the minimum number of cells in an LMA. Then, we present an algorithm in Fig. 4, which determines N_Z and N_L. It takes the initial results of the MBS-zone planning that is produced by minimizing the bandwidth usage and tries to reduce the disruption probability by increasing the number of cells in the MBS zones and the LMAs.[6] This process is repeated until the disruption probability is below a specified value δ. Implementing a dynamic MBS system may be possible by periodically running the algorithm at the MBSC. However, since it will incur a large overhead due to the synchronous transmission and scheduling problems [16] in a dynamic set of BSs, we will only consider a static MBS-zone system; once MBS zones and LMAs are planned, the configuration will not be dynamically changed.

IV. Performance Analysis

We will now analyze the service-disruption time and the bandwidth usage in the LMS while considering user distribution and mobility, and MBS session popularity. We make the following assumptions and use the notations that are summarized in Table I:

1) The MBS-session duration time follows an exponential distribution with mean $1/\lambda_s$.
2) The total number of MBS sessions is S, and all sessions are ranked by popularity. Let β_i be the conditional probability that the ith most popular session is requested $(i = 1, 2, \ldots, S)$, given that a request arrives. β_i is drawn from a cutoff Zipf-like distribution [20] and is given by

$$\beta_i = \frac{\Omega}{i^\alpha}, \qquad \text{where } \Omega = \left(\sum_{i=1}^{S} \frac{1}{i^\alpha} \right)^{-1}; \quad 0 < \alpha \leq 1. \quad (1)$$

3) The spatial distribution of MBS users follows a 2-D Poisson distribution [21] with net rate ρ^*, which is defined as the average number of users per unit area, i.e., $\rho^* = \lambda^*/\mu^*$, where λ^* is the users' arrival rate, and μ^* is the number of users that are leaving per second. Therefore, the probability that x users appear in area A is $(\rho^* A)^x e^{-\rho^* A}/x!$. From (1), the average number of users of the ith most popular session per unit area is $\rho_i = \beta_i \rho^*$.
4) The sizes of cells are identical in the network. For each session, the sizes of MBS zones (and LMAs) are independent and identically distributed (i.i.d.). Let Z, L, and C be random variables that are representing the numbers of MBS-zone crossings (i.e., inter-MBS-zone handovers) per session, LMA crossings (i.e., inter-LMA handovers) per session, and cell crossings (i.e., the total number of handovers) per session, respectively.
5) The residence times in an MBS zone, an LMA, and a cell follow Gamma distributions with mean values $1/\lambda_z$ (variance V_z), $1/\lambda_l$ (variance V_l), and $1/\lambda_c$ (variance V_c), respectively.[7] The Gamma distribution is widely employed to model the MS movement in many studies [22]–[24]. For each session, the residence times in the MBS zone, the LMA, and the cell are i.i.d..

TABLE I
Notation

λ_c	cell crossing rate
$\lambda_z(\lambda_l)$	MBS zone (LMA) crossing rate
λ_s	MBS session service rate
S	total number of MBS sessions
m	number of sessions that can be transmitted simultaneously over a wireless link
α	Zipf-like distribution exponent
ρ^*	average number of users per unit area
ρ_i	average number of users per unit area of session i
$A_z(A_l)$	area of an MBS zone (an LMA)
$Z_{h,i}$	number of MBS zone handovers of session i ($h = 1$: inter-MBS zone handover moving to inactive zones, $h = 2$: inter-MBS zone handover moving to active zones, $h = 3$: intra-MBS zone handover)
$L_{h,i}$	number of LMA handovers of session i ($h = 1$: inter-LMA handover moving to inactive LMAs, $h = 2$: inter-LMA handover moving to active LMAs, $h = 3$: intra-LMA handover)
D_{Zh}	delay of an MBS zone handover ($h = 1$: inter-MBS zone handover moving to inactive zones, $h = 2$: inter-MBS zone handover moving to active zones, $h = 3$: intra-MBS zone handover)
D_{Lh}	delay of an LMA handover ($h = 1$: inter-LMA handover moving to inactive LMAs, $h = 2$: inter-LMA handover moving to active LMAs, $h = 3$: intra-LMA handover)

A. Disruption Time

The service-disruption time for an MBS user is defined as the sum of all handover delays during the service time of an MBS session. For MBS zones, there are three types of handover: 1) the inter-MBS-zone handovers in which an MS moves to an inactive MBS zone; 2) the inter-MBS-zone handovers in which an MS moves to an active MBS zone; and 3) the intra-MBS-zone handovers. Let Z_1 and Z_2 be the numbers of inter-MBS-zone handovers to inactive and active MBS zones, respectively, and let Z_3 be the number of intra-MBS-zone handovers. Then,

[6] If there is no size or shape constraint for planning MBS zones and LMAs, "increase N_Z/N_L" means an increment by one.

[7] Since the values of $1/\lambda_z$ and $1/\lambda_l$ may differ for different sessions, $1/\lambda_{z,i}$ and $1/\lambda_{l,i}$ will be exact expressions. For the sake of simplicity, however, we skip subscript i.

$E[Z] = E[Z_1] + E[Z_2]$, and $E[C] = E[Z] + E[Z_3]$. The average service-disruption time for the ith most popular session can be expressed as

$$\text{Average Disruption Time } (i) = E[Z_{1,i}] \cdot D_{Z1} + E[Z_{2,i}] \cdot D_{Z2} + E[Z_{3,i}] \cdot D_{Z3} \quad (2)$$

where D_{Z1}, D_{Z2}, and D_{Z3} are the unit delays for an inter-MBS-zone handover to an inactive MBS zone, an inter-MBS-zone handover to an active MBS zone, and an intra-MBS-zone handover, respectively.

Since the LMS partitions an MBS zone into LMAs, the intra-MBS-zone handovers can be subclassified into three types of LMA handover: 1) the inter-LMA handovers in which an MS moves to an inactive LMA; 2) the inter-LMA handovers in which an MS moves to an active LMA; and 3) the intra-LMA handovers. Let L_1 and L_2 be the numbers of inter-LMA handovers to inactive LMAs and to active LMAs, respectively, and let L_3 be the number of intra-LMA handovers. Then, we have $E[Z_3] = E[L_1] + E[L_2] + E[L_3]$. From (2), the average service-disruption time for the ith most popular session in the LMS, i.e., $T_{\text{LMS},i}$, can be expressed as

$$T_{\text{LMS},i} = E[Z_{1,i}] \cdot D_{Z1} + E[Z_{2,i}] \cdot D_{Z2} + E[L_{1,i}] \cdot D_{L1} + E[L_{2,i}] \cdot D_{L2} + E[L_{3,i}] \cdot D_{L3} \quad (3)$$

where D_{L1}, D_{L2}, and D_{L3} are the unit delays for an inter-LMA handover to an inactive LMA, an inter-LMA handover to an active LMA, and an intra-LMA handover, respectively. Formulas for determining D_{Z1}, D_{Z2}, D_{L1}, D_{L2}, and D_{L3} are given in [25].

Let $p(x, i, A)$ denote the probability that there are x users that are subscribing the ith most popular session in area A, so that $p(x, i, A) = (\rho_i A)^x e^{-\rho_i A}/x!$. The probability that there is no user that is subscribing to the ith most popular session in an MBS zone with area A_z is given by $p(0, i, A_z) = e^{-\rho_i A_z}$. Let $\Pr(Z_1 = j | Z = n)$ be the conditional probability that there are n inter-MBS-zone handovers, among which j handovers are to inactive the MBS zones. It follows a Bernoulli distribution and can be expressed as $\Pr(Z_1 = j | Z = n) = \binom{n}{j}[p(0, i, A_z)]^j [1 - p(0, i, A_z)]^{n-j}$.

Probability $\Pr(Z = n)$ can be obtained by using the results in [24], i.e.,

$$\Pr(Z = n) = \begin{cases} 1 - \frac{\lambda_z(1 - f_z^*(\lambda_s))}{\lambda_s}, & n = 0 \\ \frac{\lambda_z}{\lambda_s}\left[1 - f_z^*(\lambda_s)\right]^2 \left[f_z^*(\lambda_s)\right]^{n-1}, & n > 0 \end{cases} \quad (4)$$

where $f_z^*(s) = [\lambda_z \gamma/(s + \lambda_z \gamma)]^\gamma$ is the Laplace–Stieltjes transform of a Gamma random variable with parameter $\gamma = 1/(V_z \lambda_z^2)$. The average number of MBS-zone crossings can be computed as $E[Z] = \sum_{n=0}^{\infty} n \Pr(Z = n) = \lambda_z/\lambda_s$.

Then, the average number of inter-MBS-zone handovers to inactive MBS zones for session i (i.e., $E[Z_{1,i}]$) can be expressed as

$$E[Z_{1,i}] = \sum_{n=0}^{\infty} \sum_{j=0}^{n} j \cdot \Pr(Z_1 = j | Z = n) \cdot \Pr(Z = n)$$

$$= \sum_{n=1}^{\infty} \sum_{j=1}^{n} \frac{j\lambda_z}{\lambda_s} \Pr(Z_1 = j | Z = n) \times \left[1 - f_z^*(\lambda_s)\right]^2 \left[f_z^*(\lambda_s)\right]^{n-1}$$

$$= \frac{\lambda_z}{\lambda_s} p(0, i, A_z). \quad (5)$$

From $\Pr(Z_2 = j | Z = n) = \binom{n}{j}[1 - p(0, i, A_z)]^j [p(0, i, A_z)]^{n-j}$, the average number of inter-MBS-zone handovers to active MBS zones (i.e., $E[Z_{2,i}]$) can be expressed as

$$E[Z_{2,i}] = \sum_{n=0}^{\infty} \sum_{j=0}^{n} j \cdot \Pr(Z_2 = j | Z = n) \cdot \Pr(Z = n)$$

$$= \frac{\lambda_z}{\lambda_s} \left(1 - p(0, i, A_z)\right). \quad (6)$$

Recall that L is the random variable that is representing the number of LMA crossings, and then, $E[L] = E[Z] + E[L_1] + E[L_2]$. By a similar derivation from (4), the LMA-crossing probability $\Pr(L = n)$ can also be expressed as

$$\Pr(L = n) = \begin{cases} 1 - \frac{\lambda_l(1 - f_l^*(\lambda_s))}{\lambda_s}, & n = 0 \\ \frac{\lambda_l}{\lambda_s}\left[1 - f_l^*(\lambda_s)\right]^2 \left[f_l^*(\lambda_s)\right]^{n-1}, & n > 0 \end{cases}$$

where $f_l^*(s) = [\lambda_l \gamma/(s + \lambda_l \gamma)]^\gamma$ and $\gamma = 1/(V_l \lambda_l^2)$. Then, the average number of LMA crossings can be computed as $E[L] = \sum_{n=0}^{\infty} n \Pr(L = n) = \lambda_l/\lambda_s$. Since $E[L_1] + E[L_2] = E[L] - E[Z]$, the average number of inter-LMA handovers that do not involve changing MBS zones is given by $E[L] - E[Z] = (\lambda_l - \lambda_z)/\lambda_s$. The probability that there is no user for session i in an LMA with area A_l is $p(0, i, A_l) = e^{-\rho_i A_l}$. Therefore, $E[L_{1,i}]$ and $E[L_{2,i}]$ can be derived as

$$E[L_{1,i}] = \frac{\lambda_l - \lambda_z}{\lambda_s} \cdot p(0, i, A_l) \quad (7)$$

$$E[L_{2,i}] = \frac{\lambda_l - \lambda_z}{\lambda_s} \left(1 - p(0, i, A_l)\right). \quad (8)$$

We can also derive the average number of cell crossings $E[C] = \sum_{n=0}^{\infty} n \Pr(C = n) = \sum_{n=1}^{\infty} (n\lambda_c/\lambda_s)[1 - f_c^*(\lambda_s)]^2 [f_c^*(\lambda_s)]^{n-1} = \lambda_c/\lambda_s$. Since $E[L_3] = E[C] - E[L]$, the average number of intra-LMA handovers can be written as

$$E[L_{3,i}] = \frac{\lambda_c - \lambda_l}{\lambda_s}. \quad (9)$$

From (3) and (5)–(9), we can write

$$T_{\text{LMS},i} = \frac{\lambda_z}{\lambda_s}\left[e^{-\rho_i A_z} \cdot D_{Z1} + (1 - e^{-\rho_i A_z}) \cdot D_{Z2}\right] + \frac{(\lambda_l - \lambda_z)}{\lambda_s}\left[e^{-\rho_i A_l} \cdot D_{L1} + (1 - e^{-\rho_i A_l}) \cdot D_{L2}\right] + \frac{(\lambda_c - \lambda_l)}{\lambda_s} \cdot D_{L3}. \quad (10)$$

TABLE II
TRANSMISSION DELAY PROFILE

Profile	Transmission delay				
	BS ↔ MS	BS ↔ MBSC	BS ↔ PC	MBSC ↔ AAA	MBSC ↔ PC
P-FAIR	1	1	1	1	1
P-AIR	10	1	1	1	1
P-ASN	1	10	10	1	1
P-CSN	1	1	1	10	10
P-REAL	28	21	21	1	1

B. Bandwidth Usage

The bandwidth usage for a particular session is defined as the ratio of the number of cells that are transmitting multicast packets of the session to the total number of cells in the network. We define $U_{\mathrm{LMS},i}$ as the bandwidth usage in LMS, which can be expressed as

$$U_{\mathrm{LMS},i} = (1 - p(0, i, A_l)) = 1 - e^{-\rho_i A_l}. \tag{11}$$

C. Blocking Probability

When an MS that is requesting a particular session i cannot receive its content, we say that session i is blocked. The blocking probability $B_{\mathrm{LMS},i}$ is defined to be the probability that an attempt to request session i fails due to the lack of capacity in a cell. We assume that the blocking only occurs on a wireless link with the finite capacity of the cell, that is, the maximum number of sessions that can be simultaneously transmitted over a wireless link is denoted by m. However, requesting an already-ongoing session is always accepted. Therefore, $B_{\mathrm{LMS},i}$ can be written as

$$B_{\mathrm{LMS},i} = \pi_{0,i} \cdot B^m_{\mathrm{LMS},i} \tag{12}$$

where $\pi_{0,i}$ is the steady-state probability for session i to be not on air (or be in the inactive state), and $B^m_{\mathrm{LMS},i}$ is the probability that the wireless link is consumed by m sessions other than the requested session i [26].

To calculate $\pi_{0,i}$, we consider a Markov chain that alternates between two states, which are *on* and *off*. Note that $p(0, i, A_l)$ means the ratio of OFF-state cells in the network for session i, where A_l is the area of an LMA. If we call OFF-state 0 and ON-state 1, the transition probability matrix is

$$\left(\begin{array}{c|c} B^m_{\mathrm{LMS},i}\,(1 - p(0, i, A_l)) & (1 - B^m_{\mathrm{LMS},i}) \\ \quad + p(0, i, A_l) & \quad \times (1 - p(0, i, A_l)) \\ \hline p(0, i, A_l) & 1 - p(0, i, A_l) \end{array}\right).$$

Therefore, the steady-state probability to be in the inactive state can be expressed as

$$\pi_{0,i} = \frac{p(0, i, A_l)}{1 - B^m_{\mathrm{LMS},i}\,(1 - p(0, i, A_l))}. \tag{13}$$

By combining (12) and (13), we obtain $B_{\mathrm{LMS},i}$, which is the blocking probability of session i in LMS, i.e.,

$$B_{\mathrm{LMS},i} = \frac{B^m_{\mathrm{LMS},i} \cdot e^{-\rho_i A_l}}{1 - B^m_{\mathrm{LMS},i}(1 - e^{-\rho_i A_l})}. \tag{14}$$

Since our MBS system has S available sessions and m admitted sessions, $B^m_{\mathrm{LMS},i}$ can be modeled by an $M/M/m/m/S$ system, which is referred to as the Engset system [27]. Formulas for calculating $B^m_{\mathrm{LMS},i}$ are given in the Appendix.

V. NUMERICAL RESULTS

We will now evaluate the performance of the LMS in terms of service-disruption time, wireless-link bandwidth usage, and blocking probability. To analyze the disruption time, we need to quantify each kind of handover delay by identifying the transmission and processing delays that are caused by signaling messages at each network node [25]. Table II shows five transmission-delay profiles that are used in this analysis with associated parameter values. In the P-FAIR profile, all transmission delays are equal; the P-AIR profile has an increased delay in the wireless link, and P-ASN and P-CSN have increased delays in the wired parts. Finally, P-REAL that contains values of the results from our measurement experiments [25] is more realistic than others. The time required to perform the IEEE 802.16e MAC-layer handover process (step 1 in Figs. 2 and 3) is highly dependent on the physical parameters and the extent to which the handover process is optimized. Since this delay is included in all types of handover, we will ignore it in our analysis of the disruption time. Additionally, all processing delays are assumed to be 1.

For the sake of simplicity, it is assumed that the MBS zones, the LMAs, and the cells are circular or square shaped and that there are N_Z cells in an MBS zone and N_L cells in an LMA. We also assume that each cell has an area of 1 km^2. We can then use a fluid-flow mobility model to express the cell boundary-crossing rate as $\lambda_c = (16v)/(\pi l)$, where v is the average velocity of the MSs and l is the length of the perimeter of a cell. This allows us to approximate λ_z and λ_l by $\lambda_c/\sqrt{N_Z}$ and $\lambda_c/\sqrt{N_L}$, respectively, [28]. The average duration of a session $1/\lambda_s$ is set to 60 min, and the total number of MBS sessions S is set to 100. Compared with the OMS without applying LMAs, we will use two notations to identify each zone-planning scheme:

1) OMS(N_Z): the original method of zone planning with N_Z-cell MBS zones;
2) LMS(N_Z, N_L): LMA-based zone planning with N_Z-cell MBS zones that are partitioned into N_L-cell LMAs.

Note that OMS(k) can be modeled by LMS(k, k) for any value of k. Additionally, we will use notations $T_{\mathrm{OMS},i}$, $U_{\mathrm{OMS},i}$, and $B_{\mathrm{OMS},i}$ for OMS(k), which can be derived from $T_{\mathrm{LMS},i}$, $U_{\mathrm{LMS},i}$, and $B_{\mathrm{LMS},i}$ for LMS(k, k), respectively.

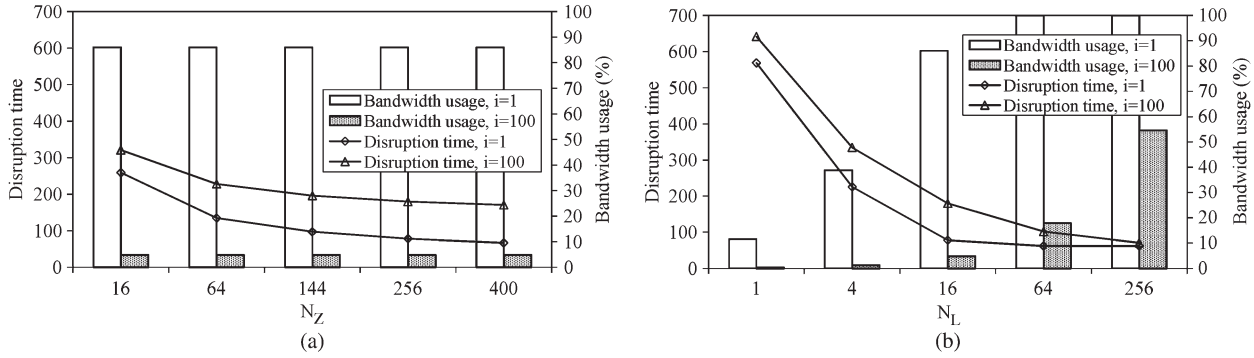

Fig. 5. Disruption time ($T_{\mathrm{LMS},1}$ and $T_{\mathrm{LMS},100}$) and bandwidth usage ($U_{\mathrm{LMS},1}$ and $U_{\mathrm{LMS},100}$) as a function of N_Z or N_L ($\alpha = 0.8$, $\rho^* = 1$ user/cell, $v = 60$ km/h, and using P-FAIR). (a) LMS(N_Z, 16): Effect of N_Z. (b) LMS(256, N_L): Effect of N_L.

A. Effect of N_Z and N_L

Fig. 5 shows the disruption time for the LMS ($T_{\mathrm{LMS},1}$ and $T_{\mathrm{LMS},100}$) and its bandwidth usage ($U_{\mathrm{LMS},1}$ and $U_{\mathrm{LMS},100}$). In Fig. 5(a), the bandwidth usage is shown to be independent of the value of N_Z, which is because MBS packets are selectively transmitted to active LMAs. However, the disruption time decreases as N_Z increases. As N_Z rises from 16 to 400, $T_{\mathrm{LMS},1}$ and $T_{\mathrm{LMS},100}$ can be reduced by 74% and 47%, respectively. Fig. 5(b) shows the disruption time and the bandwidth usage against N_L with $N_Z = 256$. Recall that N_L is a single unit of transmission for the LMS, whereas N_Z is the unit of transmission for the OMS. Therefore, the $U_{\mathrm{LMS},1}$ and $U_{\mathrm{LMS},100}$ curves for the LMS (256, k) are exactly the same as the $U_{\mathrm{OMS},1}$ and $U_{\mathrm{OMS},100}$ curves for OMS(k) for any k. Since $U_{\mathrm{LMS},i}$ is independent of the MBS zone size, we have

$$U_{\mathrm{LMS},i} \text{ of LMS}(n,k) = U_{\mathrm{OMS},i} \text{ of OMS}(k) \le U_{\mathrm{OMS},i} \text{ of OMS}(n), \qquad n \ge k.$$

Unlike the bandwidth usage, the disruption time decreases as N_L increases because the number of MSs that are crossing LMAs during a session is reduced as the LMAs become larger. When N_L is 1, as shown in Fig. 5(b) [i.e., LMS(256, 1)], $T_{\mathrm{LMS},1}$ and $T_{\mathrm{LMS},100}$ have their highest values of 568.7 and 641.9, respectively. However, LMS(256, k) significantly outperforms OMS(k) for a small k in terms of disruption time, whereas both of them use the same amount of bandwidth.[8] When k is large, LMS(256, k) and OMS(k) have a similar performance. In addition, the impact of the session popularity (i.e., the distribution of users) on the disruption time becomes less significant since more cells need to transmit the packets of an unpopular session.

With the same mobility, the number of inter-MBS-zone handovers in OMS(k) and the number of inter-LMA handovers in LMS(256, k) are expected to be equal, and the inter-LMA handover delay is much shorter than the inter-MBS-zone handover delay. Consequently, LMS(n,k) shows less disruption time than OMS(k) for all values of n and k ($n > k$). However, the disruption time for LMS(n,k) is worse than that for OMS(n) since it may include additional inter-LMA handover delays, as well as the same level of inter-MBS-zone handover delays that occur with OMS(n). For example, LMS(256, 16) causes 26% more delay than OMS(256) when $i = 1$. Based on this observation, we can bound the range of $T_{\mathrm{LMS},i}$ for LMS(n,k) as

$$T_{\mathrm{OMS},i} \text{ of OMS}(n) \le T_{\mathrm{LMS},i} \text{ of LMS}(n,k) \le T_{\mathrm{OMS},i} \text{ of OMS}(k), \qquad n \ge k.$$

The blocking probability in the LMS is independent of the value of N_Z such as the bandwidth usage since N_L is a single unit of transmission for the LMS. For example, $B_{\mathrm{LMS},1}$ for LMS(n, 16) is exactly the same as the $B_{\mathrm{OMS},1}$ for OMS(16) for all values of n ($n \ge 16$). Therefore, we have

$$B_{\mathrm{LMS},i} \text{ of LMS}(n,k) = B_{\mathrm{OMS},i} \text{ of OMS}(k), \qquad n \ge k.$$

The LMS decouples the requirement for the wireless-link bandwidth from the size of an MBS zone to balance a tradeoff between the bandwidth usage and the handover delay, as compared with the OMS. If we enforce the constraint that keeps the average handover delay below a specified value, the LMS can provide a bandwidth-efficient solution. On the other hand, if we wish to ensure that the total bandwidth usage cannot exceed a specified value, it yields a less-disruptive solution.

B. Deciding N_Z and N_L

Recall that the problem of the MBS-zone planning for the LMS is to search for the values of N_Z and N_L, which satisfy $\Pr[Y > D_{\mathrm{th}}]_i \le \delta$, where D_{th} is the handover-delay threshold and δ is the tolerable disruption ratio of session i. The probability that receiving session i is disrupted can be written as

$$\Pr[Y > D_{\mathrm{th}}]_i = \begin{cases} 1, & 0 \le D_{\mathrm{th}} < D_{L3} \\ \frac{E[Z_{1,i}]+E[Z_{2,i}]+E[L_{1,i}]}{E[C]}, & D_{L3} \le D_{\mathrm{th}} < D_{L1} \\ \frac{E[Z_{1,i}]+E[Z_{2,i}]}{E[C]}, & D_{L1} \le D_{\mathrm{th}} < D_{Z2} \\ \frac{E[Z_{1,i}]}{E[C]}, & D_{Z2} \le D_{\mathrm{th}} < D_{Z1} \\ 0, & D_{Z1} \le D_{\mathrm{th}}. \end{cases}$$

The resulting disruption probabilities are summarized in Table III, where $D_{L3} \le D_{\mathrm{th}} < D_{L1}$, $D_{L1} \le D_{\mathrm{th}} < D_{Z2}$, and $D_{Z2} \le D_{\mathrm{th}} < D_{Z1}$.

Fig. 6(a) shows the values of N_Z and N_L, which are determined by the LMA-based MBS-zone-planning algorithm

[8] $T_{\mathrm{LMS},1} = 1263$ and $T_{\mathrm{LMS},100} = 1298$ for OMS(1). $T_{\mathrm{LMS},1} = 259$ and $T_{\mathrm{LMS},100} = 321$ for OMS(16).

TABLE III
DISRUPTION PROBABILITIES FOR LMS ($\alpha = 0.8$, $\rho^* = 1$ USER/CELL, AND $v = 60$ km/h)

N_L	D_{th}	$Pr[Y > D_{th}]_i$							
		$N_Z = 16$		$N_Z = 64$		$N_Z = 256$		$N_Z = 400$	
		$i = 1$	$i = 100$	$i = 1$	$i = 100$	$i = 1$	$i = 100$	$i = 1$	$i = 100$
1	$D_{L3} \le D_{th} < D_{L1}$	0.9132	0.9977	0.8988	0.9973	0.8916	0.9971	0.8901	0.9971
	$D_{L1} \le D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \le D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026	0.0000	0.0284	0.0000	0.0145
4	$D_{L3} \le D_{th} < D_{L1}$	0.4029	0.4969	0.3543	0.4954	0.3301	0.4946	0.3252	0.4945
	$D_{L1} \le D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \le D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026	0.0000	0.0284	0.0000	0.0145
9	$D_{L3} \le D_{th} < D_{L1}$	0.2776	0.3310	0.1939	0.3276	0.1521	0.3259	0.1437	0.3256
	$D_{L1} \le D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \le D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026	0.0000	0.0284	0.0000	0.0145
16	$D_{L3} \le D_{th} < D_{L1}$	0.2500	0.2500	0.1425	0.2440	0.0887	0.2410	0.0780	0.2404
	$D_{L1} \le D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \le D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026	0.0000	0.0284	0.0000	0.0145
25	$D_{L3} \le D_{th} < D_{L1}$	N/A	N/A	0.1285	0.1944	0.0689	0.1898	0.0569	0.1889
	$D_{L1} \le D_{th} < D_{Z2}$	N/A	N/A	0.1250	0.1250	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \le D_{th} < D_{Z1}$	N/A	N/A	0.0000	0.1026	0.0000	0.0284	0.0000	0.0145

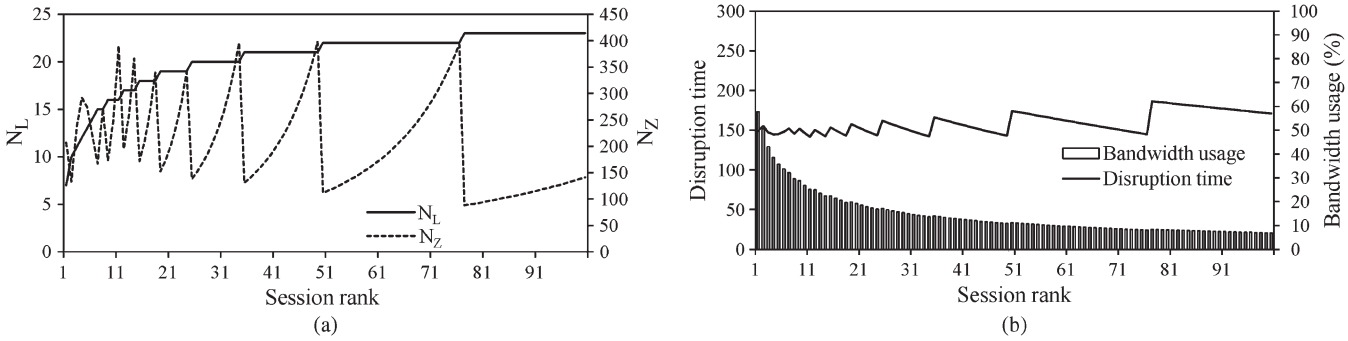

Fig. 6. LMA and MBS-zone-planning results ($N_{L,\min} = 1$, $N_{Z,\min} = 1$, $N_{Z,\max} = 400$, $\rho^* = 1$ user/cell, $v = 60$ km/h, and P-FAIR). (a) Deciding N_L and N_Z. (b) Disruption time and bandwidth usage.

TABLE IV
BLOCKING PROBABILITIES ($m = 20$ SESSIONS, $\alpha = 0.8$, AND $\rho^* = 1$ USER/CELL)

Best result	$i = 1$	$i = 4$	$i = 7$	$i = 10$	$i = 20$	$i = 30$	$i = 50$	$i = 100$
LMS	0.0%	0.0%	0.4%	0.7%	2.8%	4.2%	8.5%	9.6%
OMS	0.3%	4.5%	7.2%	8.6%	10.7%	11.6%	12.5%	13.3%

given in Fig. 4, where $\delta = 0.2$, and $D_{L3} \le D_{\text{th}} < D_{L1}$ for all MBS sessions. Then, the determination of N_L is affected by the popularity of the session. As shown in Fig. 6(a), a less-popular session has the higher values of N_L. Intuitively, it is expected that the less-popular session has fewer receivers; therefore, the bandwidth usage will be lower. In addition, an MS that is receiving an unpopular session has less chance to encounter LMAs with the session on air, which results in a larger disruption time. Therefore, to mitigate this effect, N_L should be increased as the session popularity decreases. The determination of N_Z, however, is hardly affected by the session popularity. Note that the N_Z curve shown in Fig. 6(a) represents the minimum values of N_Z. For the LMS, the bandwidth usage does not rely on the value of N_Z; therefore, a larger N_Z is always better. For example, a minimum value of N_Z between $i = 50$ and 76 increases as the session popularity decreases because N_L is fixed. At $i = 76$ in particular, the minimum value of N_Z approaches $N_{Z,\max}$. Therefore, when i reaches 77, a smaller N_Z becomes feasible because of the increase in N_L. Additionally, the algorithm in Fig. 4 usually selects a large N_Z (> 100) to avoid inter-MBS-zone handovers. Fig. 6(b) shows the disruption time and the bandwidth usage based on the planning results, and Table IV shows the blocking probabilities for $m = 20$.[9] Compared with OMS(25), which satisfies $\delta = 0.2$ that is minimizing its bandwidth usage,[10] the disruption time for the LMS is reduced by 36% on the average. In addition, the bandwidth usage is slightly reduced; thus, the blocking probabilities are also improved.

C. Effect of Different Delay Profiles

Fig. 7 shows how the disruption times for the OMS and the LMS are affected by the five different delay profiles. The results are normalized to the P-FAIR profile with $i = 1$ and $i = 100$. Except for P-REAL, P-ASN generally exhibits the highest values for both the OMS and the LMS, which implies that the disruption time is largely a result of the increased delay

[9]Regarding an MBS session of 384 Kb/s, a WiMAX BS can simultaneously transmit about 20 sessions with a modulation of 16 quadrature amplitude modulation (QAM) 1/2 and a downlink/uplink ratio of 2 using a 10-MHz channel [5].

[10]By putting $N_L = N_Z$ in Fig. 4, the OMS planning result can be similarly obtained.

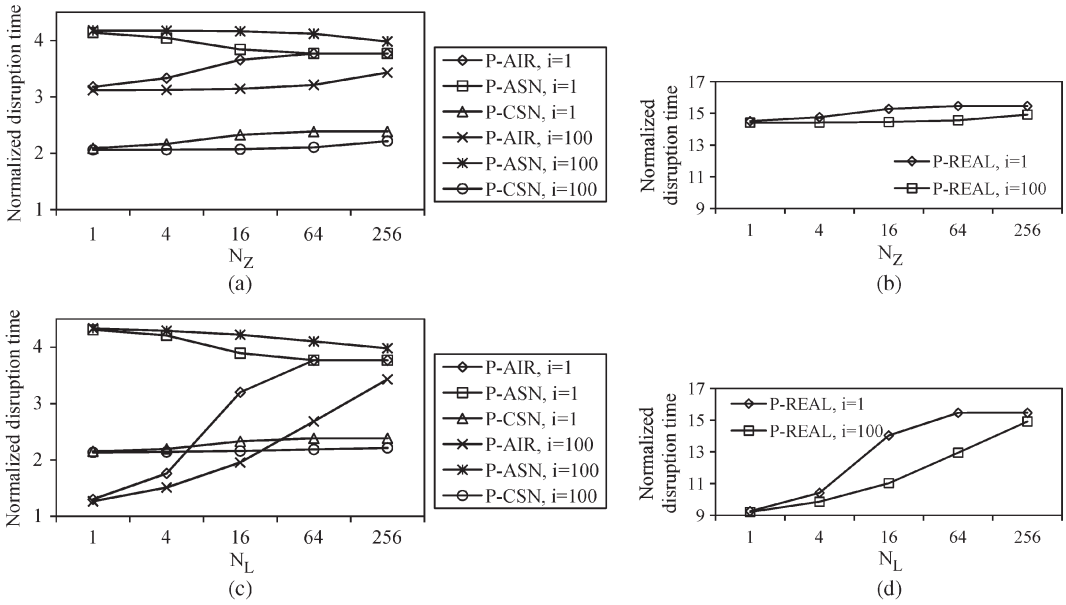

Fig. 7. Effect of different delay profiles on the disruption time ($\alpha = 0.8$, $\rho^* = 1$ user/cell, $v = 60$ km/h). These results are normalized to the values of P-FAIR. (a) P-AIR, P-ASN, and P-CSN in OMS(N_Z). (b) P-REAL in OMS(N_Z). (c) P-AIR, P-ASN, and P-CSN in LMS(256, N_L). (d) P-REAL in LMS(256, N_L).

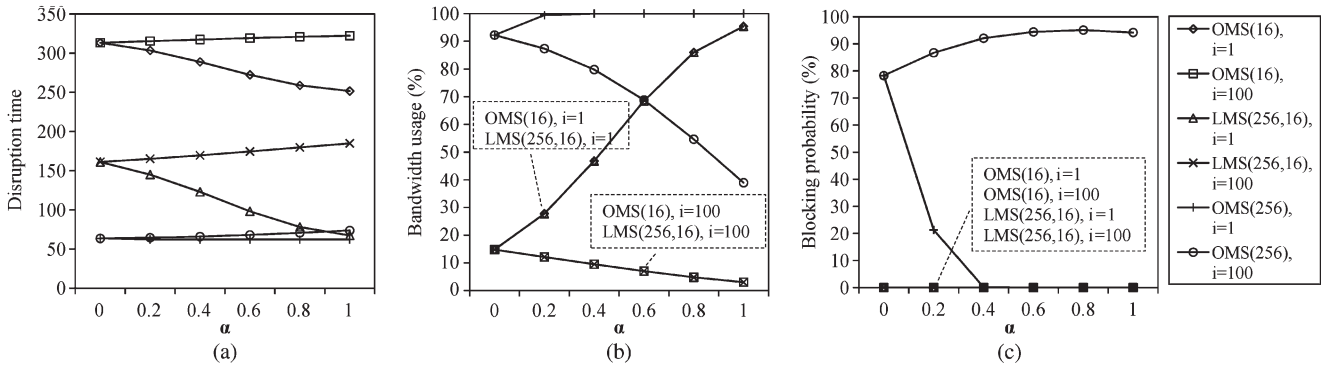

Fig. 8. Effect of α ($m = 20$, $\rho^* = 1$ user/cell, $v = 60$ km/h, and P-FAIR). (a) Disruption time. (b) Bandwidth usage. (c) Blocking probability.

between the MBSC (or the PC) and the BS. P-CSN has little effect on the disruption time with the lowest values in most cases. Clearly, the change of N_Z is less significant in Fig. 7(a) and (b). However, the curves for P-AIR in Fig. 7(c) show an interesting phenomenon, i.e., the disruption time for the LMS is insensitive to the increasing delay in the wireless links when N_L is small. When $N_L = 1$, the disruption time for P-AIR is about 1.3 times that for P-FAIR, although the transmission delay between a BS and an MS is increased by ten times for P-AIR. This phenomenon can be explained by the inter-LMA handover process, which requires no messages to be exchanged over the wireless link, except for the IEEE 802.16e MAC-layer handover process (step 1 in Fig. 3). As a result, P-REAL shows similar curves in Fig. 7(d) since its delay is 28 times longer in a wireless link than P-FAIR. Thus, we note that P-REAL and P-AIR bring out the advantages of the LMS over the OMS when N_L is small.

D. Effect of α

The average disruption time and bandwidth usage as functions of α are shown in Fig. 8. Values of α in the range of 0.64–0.98 have been reported [20]. When $\alpha = 0$, (1) can be simplified to $\beta_i = 1/S$ for all i, and the request rates for all MBS sessions become equivalent. Fig. 8(a) shows that $T_{\mathrm{OMS},1}$ for OMS(16) and $T_{\mathrm{LMS},1}$ for LMS(256, 16) substantially decrease as α increases. Meanwhile, $T_{\mathrm{OMS},100}$ for OMS(16) and $T_{\mathrm{LMS},100}$ of LMS(256, 16) slightly increase with α. In the case of large MBS zones [e.g., OMS(256)], the effect of α is insignificant. However, a change in α has a significant impact on the bandwidth usage, as shown in Fig. 8(b). As α increases, $U_{\mathrm{LMS},1}$ for LMS(256, 16) [which is the same as $U_{\mathrm{OMS},1}$ for OMS(16)] sharply increases, whereas $U_{\mathrm{OMS},100}$ for OMS(256) sharply decreases. Moreover, the more-popular session has a higher probability to be present among the already-transmitted sessions as α increases. Accordingly, the blocking probability of a popular session decreases. However, unpopular sessions are hardly accepted in a large MBS zone, as shown in Fig. 8(c).

VI. Simulation Results

Our simulations have two goals: One is to verify the previously discussed results in Section V-B, and the other is to show the impact of the variance of the MBS-zone and LMA residence

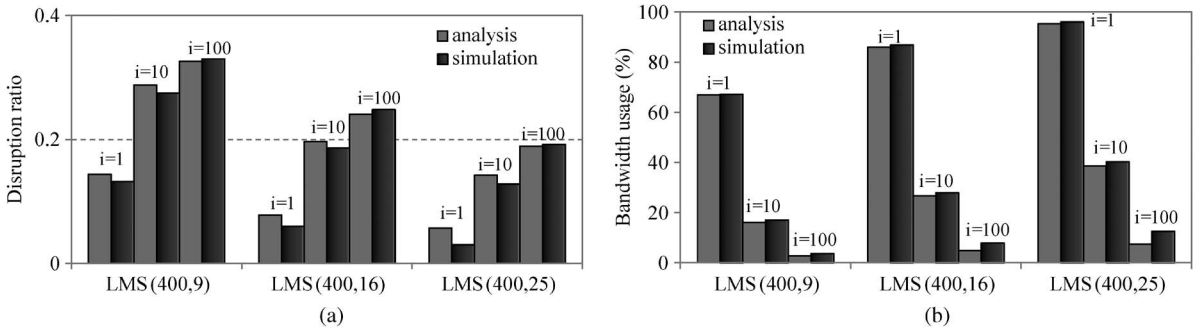

Fig. 9. Comparison of the simulation results with the analytical results. We conduct 100 simulation runs for each MBS zone and LMA configuration, and the average values are shown in the figures. (a) Disruption ratio. (b) Bandwidth usage.

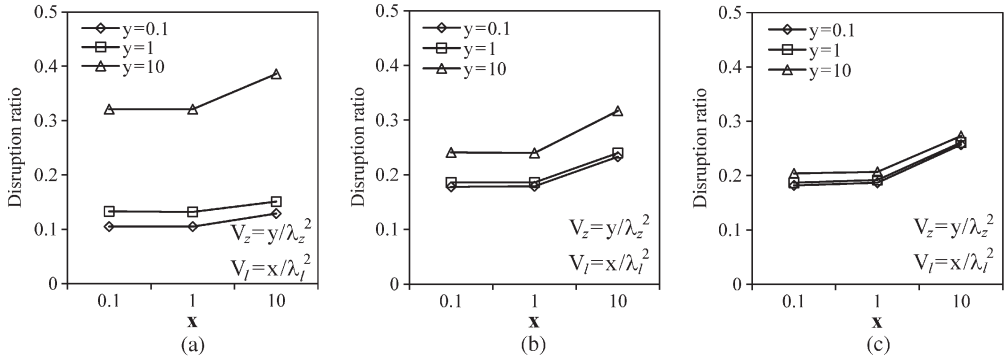

Fig. 10. Effects of V_l and V_z on disruption ratios. (a) $i = 1$ at LMS(400, 9). (b) $i = 10$ at LMS(400, 16). (c) $i = 100$ at LMS(400, 25).

times. First, we assumed an MBS service area (or a WiMAX network) that is comprised of 400 cells with 400 randomly located users. Each user moves according to a 2-D random-walk mobility model of which the speed range is uniformly distributed between 0 and 120 km/h and the service area is actually wrapping around to remove the boundary effect. To quantify handover delays, our simulation uses the MBS-handover message flows and delay values from our measurement experiments [25]. We further assume that all delay values are normally distributed as follows: A transmission delay between the MS and the BS is normally distributed with a mean of 28 ms and a variance of 4^2 (i.e., $N(28, 4^2)$), a transmission delay between the BS and the MBSC (or the PC) follows $N(21, 3^2)$, and all other transmission and processing delays follow $N(1, 0.1^2)$. The MAC-layer handover delay is fixed at 100 ms, and no message error is assumed. Then, the mean values of D_{Z1}, D_{Z2}, D_{Z3}, D_{L1}, D_{L2}, and D_{L3} are given by 345, 301, 100, 168, 100, and 100 ms, respectively. The handover threshold value is set to 130 ms. The simulation duration of each run is 120 min.

Fig. 9(a) shows the ratio for a noticeable session disruption that an MBS user experiences, whereas Fig. 9(b) shows the bandwidth usage. Recall that the MBS-zone and LMA planning results for $\delta = 0.2$ are shown in Fig. 6(a). In the simulations, three sizes of LMAs are examined; each LMA consists of 3×3, 4×4, or 5×5 cells. According to the analytical results in Table III and Fig. 6(a), LMS(400, 25) is feasible (satisfies $\delta = 0.2$) for all sessions, but LMS(400, 16) is not feasible for $i = 100$. In addition, LMS(400,9) is only feasible when $i = 1$. These match well with the simulation results shown in Fig. 9(a), although the disruption ratios are a little different between the simulations and the analytical results. This is because of the mobility of each MS, which may not guarantee that MBS users are uniformly distributed over the network. Since our algorithm finds the values of N_Z and N_L that minimize the use of bandwidth[11] while keeping a delay requirement, LMS(400,9) is the best for $i = 1$, whereas LMS(400, 16) is the best for $i = 10$. For $i = 100$, only LMS(400, 25) is feasible.

Second, we modified our simulations for each MS to have the residence times that follow a specific Gamma distribution, so that the impact of the variance of the residence times can be studied. Fig. 10 shows the average disruption ratios of MSs for Gamma residence-time distributions with different variance values as follows: the variance of the MBS-zone residence times $V_z = \{0.1/\lambda_z^2, 1/\lambda_z^2, 10/\lambda_z^2\}$ and the variance of the LMA residence times $V_l = \{0.1/\lambda_l^2, 1/\lambda_l^2, 10/\lambda_l^2\}$. Note that the exponential distribution is a special case of Gamma distributions with mean $1/\lambda$ and variance $1/\lambda^2$.

Overall, the figure indicates that the disruption ratios are substantially affected when the variance of the residence times of MBS zones and LMAs is high. For a high variance of the LMA residence times, our analysis may underestimate the ratio for a session disruption. In Fig. 10, most of the cases for $V_l = 10/\lambda_l^2$ do not satisfy $\delta = 0.2$ since the disruption ratios for $V_l = 10/\lambda_l^2$ are increased by 14%–37%, as compared with those for $V_l = 1/\lambda_l^2$. On the other hand, the variance of the MBS-zone residence times only affects the disruption ratios

[11]Fig. 9(b) shows that the bandwidth usage is overestimated in the simulations when $i = 100$. This is because the simulations use an integer (2 for $i = 100$) as the expected number of session users, rather than a real number (1.24 for $i = 100$, by analysis).

of the popular sessions.[12] As V_z is increased from $1/\lambda_z^2$ to $10/\lambda_z^2$, the disruption ratios are more than doubled in Fig. 10(a) $(i = 1)$, whereas they are hardly affected by V_z in Fig. 10(c) $(i = 100)$.

VII. Conclusion

Mobile WiMAX includes the MBS zone to reduce the MBS service disruption due to handovers, but this requires all the BSs to send the same packets in the MBS zone. This has motivated us to propose an MBS handover and zone-planning scheme based on LMAs to save the wireless-link bandwidth while keeping the service-disruption time at an acceptable level. We have presented a novel mathematical model of the service-disruption time, the bandwidth usage, and the blocking probability, which consider the user mobility, the user distribution, and the MBS session popularity. We have evaluated the performance of our scheme (LMS) and have compared it with the OMS with the results given here.

1) The inter-LMA handover delay is shorter than the inter-MBS-zone handover delay, whereas the intra-LMA and intra-MBS-zone handover delays are the same. As a result, the LMS outperforms the OMS in terms of the average disruption time when they use the same amount of bandwidth or in terms of bandwidth usage when their average disruption times are the same.
2) The service-disruption time is mostly dominated by the transmission delay between the MBSC (or the PC) and the BS. In particular, if the LMAs are small, the disruption time of the LMS is hardly affected at all by the wireless-transmission delay.
3) The MBS-user distribution and session popularity have significant effects on the bandwidth usage and the blocking probability, whereas the service-disruption time is mainly affected by the mobility (e.g., average user speed). Moreover, the disruption ratio can be significantly affected by the variance of the LMA and/or MBS-zone residence times.

We have demonstrated how to determine the MBS-zone and LMA sizes, which can make the best use of the bandwidth while maintaining the quality of the MBS services. Our results suggest that the LMA-based MBS-zone planning would deliver more-efficient multicast and broadcast services over Mobile WiMAX systems.

Appendix
Calculation of $B^m_{\mathrm{LMS},i}$

The LMS with S available sessions and m admitted sessions can be modeled as $M/M/m/m/S$ Engset systems. The sessions in the LMS represent the users in an Engset system. In a generalized Engset system, the users are not identical; their arrival rates λ_i and departure rates μ_i, as well as the requested resources (c_i), can be different. For an Engset system with capacity C, the user blocking probability of user i is $B_i^C = (\sum_{j=C-c_i+1}^{C} \pi_j^{(i)})/(\sum_{j=0}^{C} \pi_j^{(i)})$, where $\pi_j^{(i)}$ is the probability that j capacity units are occupied in an infinite system with user i removed [27]. Probability $\pi_j^{(i)}$ can be calculated from the probability-generating function, i.e., $P_i(z) = \sum_{j=0}^{\infty} \pi_j^{(i)} z^j = \prod_{k \in S-\{i\}} (q_k + p_k z^{c_k})$, where $q_k = e^{-\lambda_k/\mu_k} = 1 - p_k$.

In the LMS, each session has a different ρ_i that represents the average number of MSs that are staying in the system for session i, and every session requests the same amount of resources (i.e., $c_i = 1$ and $C = m$). Then, the blocking probability can be expressed as $B_i^m = \pi_m^{(i)} / \sum_{j=0}^{m} \pi_j^{(i)}$, where $\pi_j^{(i)}$ is the probability that the capacity is occupied by j sessions in an infinite system with session i removed. Moreover, the probability-generating function is $P_i(z) = \sum_{j=0}^{\infty} \pi_j^{(i)} z^j = 1/(q_i + p_i z) \prod_{k=1}^{S} (q_k + p_k z)$, where $q_k = e^{-\rho_k A_l}$ in the LMS. Since we have $(d^j/dz^j) \sum_{j=0}^{\infty} \pi_j^{(i)} z^j|_{z=0} = (j!)\pi_j^{(i)}$, probability $\pi_j^{(i)}$ can be expressed as $\pi_j^{(i)} = (1/j!)(d^j/dz^j) P_i(z)|_{z=0}$. Therefore, $B^m_{\mathrm{LMS},i}$ can be computed by

$$B^m_{\mathrm{LMS},i} = \frac{\left(\frac{1}{m!}\right) \frac{d^m}{dz^m} \left[\frac{\prod_{k=1}^{S} \left\{ e^{-\rho_k A_l} + z(1 - e^{-\rho_k A_l}) \right\}}{e^{-\rho_i A_l} + z(1 - e^{-\rho_i A_l})} \right]_{|z=0}}{\sum_{j=0}^{m} \left(\frac{1}{j!}\right) \frac{d^j}{dz^j} \left[\frac{\prod_{k=1}^{S} \left\{ e^{-\rho_k A_l} + z(1 - e^{-\rho_k A_l}) \right\}}{e^{-\rho_i A_l} + z(1 - e^{-\rho_i A_l})} \right]_{|z=0}}.$$

Acknowledgment

Information and Communication Technology, Seoul National University, Seoul, Korea, provided the research facilities for this study.

References

[1] Multimedia Broadcast/Multicast Service (MBMS); Architecture and Functional Description (Rel. 8), Third Generation Partnership Project TS 23.246, v8.1.0, Dec. 2007. [Online]. Available: http://www.3gpp.org/
[2] Broadcast/Multicast Services (BCMCS)—Stage 1, Third Generation Partnership Project2 S.R0030-A, v1.0, Jan. 2004. [Online]. Available: http://www.3gpp2.org/
[3] IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Fixed Broadband Wireless Access Systems, IEEE Std. 802.16-2004, 2004. [Online]. Available: http://www.ieee.org/
[4] IEEE Standard for Local and Metropolitan Area Networks—Part 16: Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1, IEEE Std. 802.16e-2005 and IEEE Std. 802.16-2004/Cor 1-2005, 2005. [Online]. Available: http://www.ieee.org/
[5] Mobile WiMAX—Part I: A Technical Overview and Performance Evaluation, Aug. 2006. [Online]. Available: http://www.wimaxforum.org/documents/
[6] A. Dutta, J. Chennikara, W. Chen, O. Altintas, and H. Schulzrinne, "Multicasting streaming media to mobile users," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 81–89, Oct. 2003.
[7] R. Koodli, IETF RFC 4068, Fast Handovers for Mobile IPv6, Jul. 2005.
[8] R. Ramjee, K. Varadhan, L. Salgarelli, S. Thuel, S. Y. Wang, and T. L. Porta, "HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 3, pp. 396–410, Jun. 2002.
[9] M. Hauge and O. Kure, "Multicast in 3G networks: Employment of existing IP multicast protocols in UMTS," in *Proc. 5th ACM Int. Workshop Wireless Mobile Multimedia*, 2002, pp. 96–103.
[10] R. Rummler, Y. W. Chung, and A. H. Aghvami, "Modeling and analysis of an efficient multicast mechanism for UMTS," *IEEE Trans. Veh. Technol.*, vol. 54, no. 1, pp. 350–365, Jan. 2005.
[11] A. Alexiou and C. Bouras, "Multicast in UMTS: Evaluation and recommendations," *Wireless Commun. Mobile Comput.*, vol. 8, no. 4, pp. 463–481, May 2008.

[12] It may depend on the value of D_{th}. In the simulations, we assumed $D_{L3} \leq D_{\mathrm{th}} = 130$ ms $< D_{L1}$.

[12] A. Alexiou, D. Antonellis, and C. Bouras, "An efficient mechanism for multicast data transmission in UMTS," *Wireless Pers. Commun.*, vol. 44, no. 4, pp. 455–471, Mar. 2008.
[13] S. Sengupta, M. Chatterjee, and S. Ganguly, "Improving quality of VoIP streams over WiMAX," *IEEE Trans. Comput.*, vol. 57, no. 2, pp. 145–156, Feb. 2008.
[14] J. She, F. Hou, P.-H. Ho, and L.-L. Xie, "IPTV over WiMAX: Key success factors, challenges, and solutions," *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 87–93, Aug. 2007.
[15] W. Jiao, P. Jiang, and Y. Ma, "Fast handover scheme for real-time applications in mobile WiMAX," in *Proc. IEEE Int. Conf. Commun.*, 2007, pp. 6038–6042.
[16] J. Wang, M. Venkatachalam, and Y. Fang, "System architecture and cross-layer optimization of video broadcast over WiMAX," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 712–721, May 2007.
[17] S. Parkvall, E. Englund, M. Lundevall, and J. Torsner, "Evolving 3G mobile systems: Broadband and broadcast services in WCDMA," *IEEE Commun. Mag.*, vol. 44, no. 2, pp. 30–36, Feb. 2006.
[18] D.-N. Yang and M.-S. Chen, "Efficient resource allocation for wireless multicast," *IEEE Trans. Mobile Comput.*, vol. 7, no. 4, pp. 387–400, Apr. 2008.
[19] Y.-B. Lin, "A multicast mechanism for mobile networks," *IEEE Commun. Lett.*, vol. 5, no. 11, pp. 450–452, Nov. 2001.
[20] L. Berslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, Mar. 1999, pp. 126–134.
[21] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaa, and L. E. Meester, *A Modern Introduction to Probability and Statistics. Understanding Why and How*. New York: Springer-Verlag, 2005.
[22] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling of PCS networks," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1062–1072, Jul. 1999.
[23] S.-R. Yang and Y.-B. Lin, "Performance evaluation of location management in UMTS," *IEEE Trans. Veh. Technol.*, vol. 52, no. 6, pp. 1603–1615, Nov. 2003.
[24] Y.-B. Lin, "Reducing location update cost in a PCS network," *IEEE/ACM Trans. Netw.*, vol. 5, no. 1, pp. 25–33, Feb. 1997.
[25] J.-H. Lee, MBS Handover Delay Analysis, May 2010. [Online]. Available: http://mmlab.snu.ac.kr/~jhlee/TR003.pdf
[26] Y. Lu, F. Kuipers, M. Janic, and P. V. Mieghem, "E2E blocking probability of IPTV and P2PTV," in *Proc. 7th Int. IFIP-TC6 Netw. Conf. AdHoc Sens. Netw., Wireless Netw., Next Gener. Internet*, vol. 4982, *Lecture Notes in Computer Science*, 2008, pp. 445–456.
[27] J. Karvo, J. Virtamo, S. Aalto, and O. Martikainen, "Blocking of dynamic multicast connections in a single link," in *Proc. Int. Broadband Commun.*, Apr. 1998, pp. 473–483.
[28] X. Zhang, J. G. Castellanos, and A. T. Campbell, "P-MIP: Paging extensions for mobile IP," *Mobile Netw. Appl.*, vol. 7, no. 2, pp. 127–141, Mar. 2002.

Ji Hoon Lee (S'10) received the B.S. (double major) degree (*magna cum laude*) in industrial engineering and computer science from Pohang University of Science and Technology, Gyungbuk, Korea, in 2000. He is currently working toward the Ph.D. degree in the School of Computer Science and Engineering, Seoul National University, Seoul, Korea.

From 2000 to 2003, he was in the telecommunication industry and was involved in developing high-speed switches and routers. From 2004 to 2005, he was a Research Engineer for developing the media-access-control layer of the IEEE 802.16-based Worldwide-Interoperability-for-Microwave-Access base station with Postdata, USA. From 2007 to 2009, he was the Secretary for the Internet Engineering Task Force Internet Protocol with the IEEE 802.16 Networks (16ng) working group. His research interests include multimedia multicasting and mobility management in mobile wireless networks, as well as autonomous network optimization for femtocellular networks.

Sangheon Pack (S'03–M'05) received the B.S. (*magna cum laude*) and Ph.D. degrees in computer engineering from Seoul National University, Seoul, Korea, in 2000 and 2005, respectively.

From 2005 to 2006, he was a Postdoctoral Fellow with the Broadband Communications Research Group, University of Waterloo, Waterloo, ON, Canada. In 2003, he was a Visiting Researcher with the Fraunhofer Institute for Open Communication Systems (FOKUS), Berlin, Germany. Since March 2007, he has been an Assistant Professor with the School of Electrical Engineering, Korea University, Seoul. His research interests include mobility management, multimedia transmission, and quality-of-service provision issues in next-generation wireless/mobile networks.

Dr. Pack was a recipient of the Korea Foundation for Advanced Studies Computer Science and Information Technology Scholarship from 2002 to 2005. He was also a recipient of the IEEE Communications Society Asia-Pacific Outstanding Research Award 2009.

Taekyoung Kwon (A'00) received the B.S., M.S., and Ph.D. degrees in computer engineering from Seoul National University, Seoul, Korea, in 1993, 1995, and 2000, respectively.

He was a Visiting Student with the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1998 and a Visiting Scholar with the University of North Texas, Denton, in 1999. He is currently an Associate Professor with the Multimedia and Mobile Communications Laboratory, School of Computer Science and Engineering, Seoul National University. His recent research interests include radio resource management, wireless technology convergence, mobility management, and wireless sensor networks.

Yanghee Choi (SM'99) received the B.S. degree in electronics engineering from Seoul National University, Seoul, Korea, in 1975, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1977, and the Ph.D. degree of engineering in computer science from the Ecole Nationale Superieure des Telecommunications, Paris, France, in 1984.

From 1977 to 1991, he was with the Electronics and Telecommunications Research Institute, where he served as the Director of the Data Communication Section, and the Protocol Engineering Center. From 1981 to 1984, he was a Research Student with the Center National d'Etude des Telecommunications, Issy-les-Moulineaux. From 1988 to 1989, he was a Visiting Scientist with the IBM T. J. Watson Research Center, Yorktown Heights, NY. Since 1991, he has been with the School of Computer Engineering, Seoul National University, where he is currently leading the Multimedia and Mobile Communications Laboratory. His research interest includes future Internet.

Dr. Choi is the President of the Korean Institute of Information Scientists and Engineers. He is also the Chair of the Future Internet Forum. He is a regular member of the National Academy of Engineering of Korea and the Korean Academy of Science and Technology, Seongnam.