Design and Analysis of the Gateway Relocation and Admission Control Algorithm in Mobile WiMAX Networks

Zong-Hua Liu, *Student Member*, *IEEE*, and Jyh-Cheng Chen, *Senior Member*, *IEEE*

Abstract—The WiMAX Forum has defined a two-tiered mobility management to minimize handover delay and packet loss. However, it leads to another problem: When to perform ASN GW relocation? The standards only define the ASN GW relocation procedures without specifying when the ASN GW relocation should be performed. It is left for vendors and operators to develop their own proprietary solutions. In this paper, we propose an algorithm, which incorporates traditional Admission Control (AC) and Wiener Process (WP)-based prediction algorithms to determine when to carry out ASN GW relocation. We further develop an analytical model to analyze the proposed algorithm. Simulations are also conducted to evaluate the performance of the proposed algorithm. The results show that the proposed algorithm can improve the performance significantly in terms of blocking probability, dropping probability, average serving rate, and average signaling overhead.

Index Terms—Mobility management, resource management, admission control, WiMAX networks, statistics and stochastic process, and wireless networks.

1 Introduction

The IEEE 802.16-series standards [1], [2] are expected to provide broadband wireless access for a variety of multimedia services. Like other IEEE 802-series standards, IEEE 802.16 working group standardizes physical (PHY) layer and Medium Access Control (MAC) layer only. To build a complete system, higher layers are still necessary. One of the major objectives of WiMAX Forum [3], thus, is to develop and standardize the *WiMAX Forum Network Architecture* [4], [5], [6], [7], which is evolving into Internet Protocol (IP)-based wireless network. The architecture is depicted in Fig. 1. In Fig. 1, the *Access Service Network (ASN)* provides wireless radio access for WiMAX subscribers. It consists of one ASN Gateway (ASN GW) and many base stations (BSs). Each ASN is connected to *Connectivity Service Network (CSN)*, which provides IP connectivity services. To support IP mobility, Mobile IP (MIP)[1] is adopted. The Home Agent (HA) of a Mobile Station (MS) is located in the CSN of the MS's Home Network Service Provider (H-NSP). ASN GW supports the Foreign Agent (FA) functionality.

The WiMAX Forum has defined a two-tiered mobility management: *ASN Anchored Mobility* and *CSN Anchored Mobility*:

- ASN Anchored Mobility refers to the procedures associated with the MS's movement between BSs, which may belong to the same or different ASN GWs. In ASN Anchored Mobility, the context of the designated MS is transferred from the previous BS to the new BS. Without performing CSN Anchored Mobility, ASN Anchored Mobility can minimize handover delay and packet loss. For example, an MS may perform *intra-ASN handover* (e.g., changing from *Flow (1)* to *Flow (2)* in Fig. 1) while still attaching to the same ASN GW. In addition, an MS may perform *inter-ASN handover* (e.g., changing from *Flow (2)* to *Flow (3)* in Fig. 1) where the ASN GW A is the traffic anchor point and responsible for ASN-CSN tunneling. That is, traffic is still sent to ASN GW A, which then further tunnels traffic to ASN GW B. In Flow (1) and Flow (2), the MS is called *Serving MS* of ASN GW A. In Flow (3), the MS is called *Anchored MS* of ASN GW A and *handover MS* of ASN GW B. In such case, the ASN GW A and ASN GW B are called *anchored ASN GW* and *Serving ASN GW*, respectively.
- CSN Anchored Mobility refers to the process of changing the traffic anchor point and is independent of the MS's link layer handover [4]. It is also called *ASN GW relocation*. For example, if CSN Anchored Mobility is not performed, when the MS roams from ASN GW B to ASN GW C in Fig. 1, ASN GW A will tunnel traffic to ASN GW C. The MS is still served by two ASN GWs (ASN GW A and ASN GW C). As aforementioned discussion, the MS is called Anchored MS of ASN GW A. Later on, the ASN GW A may request the MS to carry out CSN Anchored Mobility, i.e., ASN GW relocation. This may happen due to the heavy load of the ASN GW A [8], to

[1] Without loss of generality, we only discuss MIP in this paper. Although Client MIP (CMIP) and Proxy MIP (PMIP) are discussed in the WiMAX standards, they are variants of MIP. The technical problems and solutions discussed in this paper apply to both CMIP and PMIP as well.

- *Z.-H. Liu is with the Department of Computer Science, National Tsing Hua University, Hsinchu 30010, Taiwan. E-mail: horselui@gmail.com.*
- *J.-C. Chen is with the Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan. E-mail: jcc@cs.nctu.edu.tw.*

Manuscript received 21 Apr. 2010; revised 11 Sept. 2010; accepted 25 Oct. 2010; published online 20 Dec. 2010.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2010-04-0185.
Digital Object Identifier no. 10.1109/TMC.2010.265.

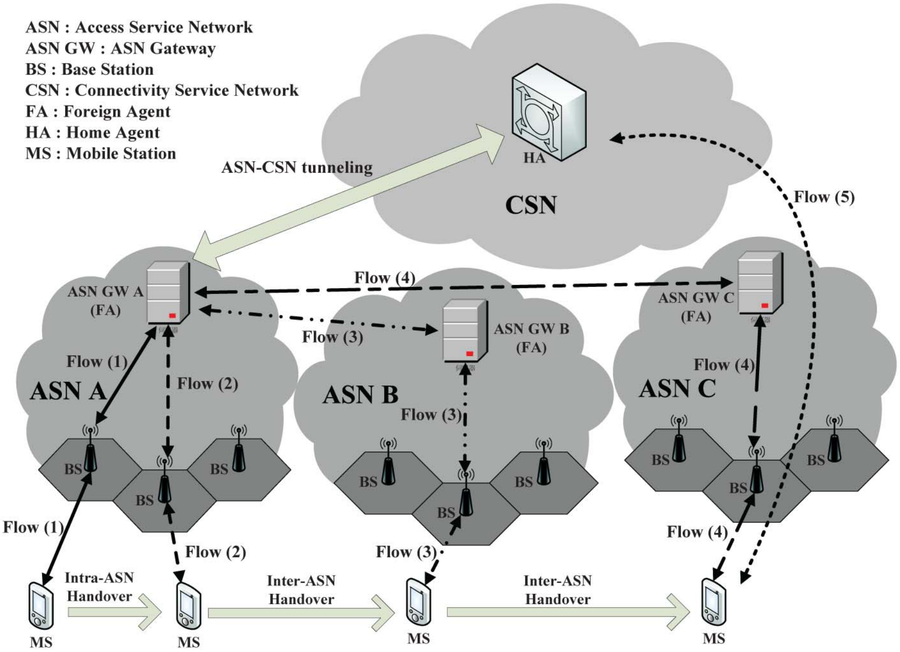

Fig. 1. ASN Anchored Mobility and CSN Anchored Mobility in WiMAX networks.

reduce end-to-end latency, or for resource optimization purposes [4], [5]. After performing ASN GW relocation, the traffic anchor point is changed to ASN GW C. The MS then is not served by ASN GW A. This is shown in Fig. 1 after changing from *Flow (4)* to *Flow (5)*.

Although the two-tiered mobility management defined in WiMAX potentially can minimize handover delay and packet loss, it leads to another problem: When to perform ASN GW relocation? The WiMAX standards, however, only define the procedures for ASN Anchored Mobility and CSN Anchored Mobility. The standards do not address when the Anchored MSs should perform ASN GW relocation to relocate the traffic anchor point from the anchored ASN GW to the serving ASN GW. The problem is left for vendors and operators to develop their own proprietary solutions. Besides, the problem is closely related to Admission Control (AC), which is widely used in wireless networks to ensure service quality and reduce network congestion by limiting the number of MSs served in the network. However, traditional AC algorithms [9], [10], [11], [12], [13], [14], [15], [16], [17] cannot be used directly when the two-tiered mobility management is deployed in WiMAX. As aforementioned discussion, some MSs may be served by two ASN GWs. The resources are required in both ASN GWs. Therefore, those MSs will be counted twice in two ASN GWs by the AC algorithm. If there are many Anchored MSs, new incoming users will likely be rejected due to the lack of resources. If the ASN GW relocation can be performed before the system becomes overloaded, the system may be able to accommodate more MSs. Therefore, a well-designed AC algorithm should cooperate with the ASN GW relocation algorithm closely.

In this paper, we propose *Gateway Relocation AC (GRAC)*, which combines ASN GW relocation and AC algorithm to maximize system capacity. In GRAC, the AC algorithm cooperates with the ASN GW relocation. When a new MS arrives and there is no resource for the newly arrived MS, the proposed GRAC will request an Anchored MS to perform ASN GW relocation if there are Anchored MSs in the system. Moreover, because handover MSs are sensitive to call dropping and handover latency, we also propose a prediction algorithm based on Wiener Process [18] to request Anchored MSs to perform ASN GW relocation early. Thus, handover MSs are not dropped when the system load is full. In addition, handover MSs do not need to wait for the completion of ASN GW relocation so handover latency can be reduced. Furthermore, we develop an analytical model to investigate the performance of the proposed GRAC. The model analyzes the performance bounds of the system. Extensive simulations are conducted to validate the analysis. The results show that the proposed GRAC can effectively reduce the blocking probability of new MSs and the dropping probability of handover MSs. The average signaling overhead is also reduced. The average serving rate is increased.

The contributions of this paper include: 1) The proposed GRAC provides a systematic way to solve the problem effectively. 2) The proposed GRAC is fully compatible with the WiMAX standards, and can be used with other AC algorithms. 3) We derive the performance bounds mathematically and show that the performance of the proposed GRAC approaches the lower bound.

The rest of this paper is organized as follows: Section 2 reviews the related work. The proposed GRAC is presented in Section 3. In Section 4, we propose an analytical model to evaluate the performance bounds of the proposed GRAC. The numerical results are discussed in Section 5. Section 6 concludes this paper.

2 Related Work

Many issues in mobile WiMAX have been studied [6], [7], [19], [20], [21], [22], [23]. In [6], [7], [19], [20], the authors provide an overview of the WiMAX technology and WiMAX network architecture. In [19], the authors discuss the mobility management in WiMAX networks and several optimization procedures for ASN Anchored Mobility management. For example, the data path is extended from the old ASN GW to a new ASN GW to reduce the impact of the delay caused by IP-layer handover. In [21], the authors propose a fast intra-network and cross-layer handover protocol to support fast and efficient handover in WiMAX. In [22], a seamless IP mobility scheme is proposed and evaluated in the flat architecture of WiMAX networks. In [23], the authors propose an analytical model to study the cost of Anchor Paging Controller (APC) reassignment in ASN GW for location update. The APC relocation problem has a great impact on signaling overhead for location tracking.

Moreover, the two-tiered mobility management defined in WiMAX is similar to that in Hierarchical MIP (HMIP) [24]. In HMIP, the multiple levels of FA hierarchy can reduce handover latency and localize the MIP signaling traffic. In addition, Xie et al. [25] propose a distributed dynamical regional location management to determine the size of a regional network based on the MS's traffic load and mobility patterns [25]. In [26], the authors design a dynamical HMIP scheme for MIP networks. Each MS dynamically determines the hierarchy of FAs according to the call-to-mobility ratio. The MIP registration update is only performed when a threshold is reached. Therefore, the signaling overhead incurred by MIP can be reduced significantly. The similar idea of chain-like architecture is also applied to the location update in cellular networks [27]. In WiMAX, however, an MS is served at most by two ASN GWs (FAs) simultaneously due to the specific two-tiered mobility management procedures.

In the literature, some gateway relocation approaches or load control techniques for cellular networks or IP-based mobile networks have been proposed [28], [29], [30], [31], [32]. In [31], the Serving Radio Network Controller (SRNC) relocation is discussed for the Universal Mobile Telecommunications System (UMTS). The SRNC in UMTS networks is similar to the ASN GW in WiMAX. They all control and manage the connections in the radio access network. When an MS no longer connects to the BS under the RNC, which is serving the MS currently, SRNC relocation is immediately initiated by the new SRNC. Hence, the SRNC can decide when to perform SRNC relocation. However, in WiMAX, if the MS connects to the BS which is under another ASN, the MS only performs ASN Anchored Mobility. This is because that if both ASN Anchored Mobility and CSN Anchored Mobility are performed simultaneously, the handover delay will become too long.

In MIP, load balancing and load control mechanisms have been proposed [28], [29], [30], [32]. The idea is that according to different criteria, MSs are equally served by HAs or Mobility Anchor Points (MAPs). However, if the approaches discussed in [28], [29], [30], [32] are used in WiMAX, the loads of the anchored and serving ASN GWs are all affected. The MSs may also need to perform both ASN Anchored Mobility and CSN Anchored Mobility during an inter-ASN handover. The long handover latency and high packet loss will degrade the service quality. On the other hand, in WiMAX, when performing ASN GW relocation, the load of the anchored ASN GW is reduced but the load of the serving ASN GW is not affected. Although the aforementioned techniques can reduce the load of the old serving ASN GW, the load of the new serving ASN GW is increased. Therefore, only the *Anchored MS* needs to perform ASN GW relocation to reduce the load of the *Anchored ASN GW*. The load of the *Serving ASN GW* is irrelevant.

Admission Control (AC) is one of the resource management techniques to limit maximum amount of traffic in the network to guarantee service quality for subscribers. In wireless and mobile networks, the AC algorithms are much more complicated due to the movement of MSs. An MS served in current network may move to another network. The connection of the MS may be dropped if the required resources in the target network cannot be supported. It is generally agreed that keeping an ongoing connection unbroken is more important than admitting a new MS. Therefore, a handover MS is given higher priority to access the network resources. For this purpose, the overall resources are partitioned and some resources are preserved for the handover MSs only. This is called *priority-based AC*. Various priority-based AC algorithms have been proposed [9], [10], [11], [12], [13], [14], [15], [16], [17]. Here, we discuss two commonly used priority-based AC algorithms: *cutoff priority algorithm* [10], [11], [12] and *new call bounding algorithm* [9].

Fig. 2 illustrates the resource allocation in the cutoff priority algorithm and new call bounding algorithm. In the cutoff priority algorithm, both new MS and handover MS can be admitted if the total number of new MSs and handover MSs in the network is equal to or less than a predefined threshold, T_{cp}, which is less than the total capacity C. Once the number of new MSs and handover MSs in the network reaches T_{cp}, new MSs are blocked. Only handover MSs are admitted. Once the total number of MSs exceeds C, handover MSs are dropped. In the new call bounding algorithm, there is a limit, T_{ncb}, for the number of new MSs admitted into the network, which is also less than the total capacity C. The handover MSs use the resources in $C - T_{ncb}$ first. If the number of new MSs is less than T_{ncb}, handover MSs can use more resources than $C - T_{ncb}$. However, the number of new MSs is always less than T_{ncb} or the remainder resources the handover MSs have not used. This is shown as $X < \min(T_{ncb}, C - Y)$ in Fig. 2b. To show the difference between the two algorithms, we assume C equals 50, and both T_{cp} and T_{ncb} are 30. We also assume in both algorithms, there are now 20 new MSs and 10 handover MSs. In the cutoff priority algorithm, a newly arrived MS will be blocked and a handover MS will be admitted. In the new call bounding algorithm, however, both a new MS and a handover MS will be admitted.

There are still many other AC algorithms. The ideas are similar although they may have different names. Nevertheless, they cannot be applied to WiMAX networks directly. As aforementioned discussion, due to the specific mobility management techniques in WiMAX, an MS may be served by two ASN GWs simultaneously. Hence, the required resources of an Anchored MS are reserved in both ASN GWs. Besides, the Anchored MS will be counted twice in two ASN GWs in the AC algorithm. Thus, when many MSs are served by two ASN GWs in the system, a newly arrived MS or handover MS may be easily blocked or dropped by the AC algorithm. Without considering ASN GW relocation in the AC algorithm, the network performance will be degraded significantly.

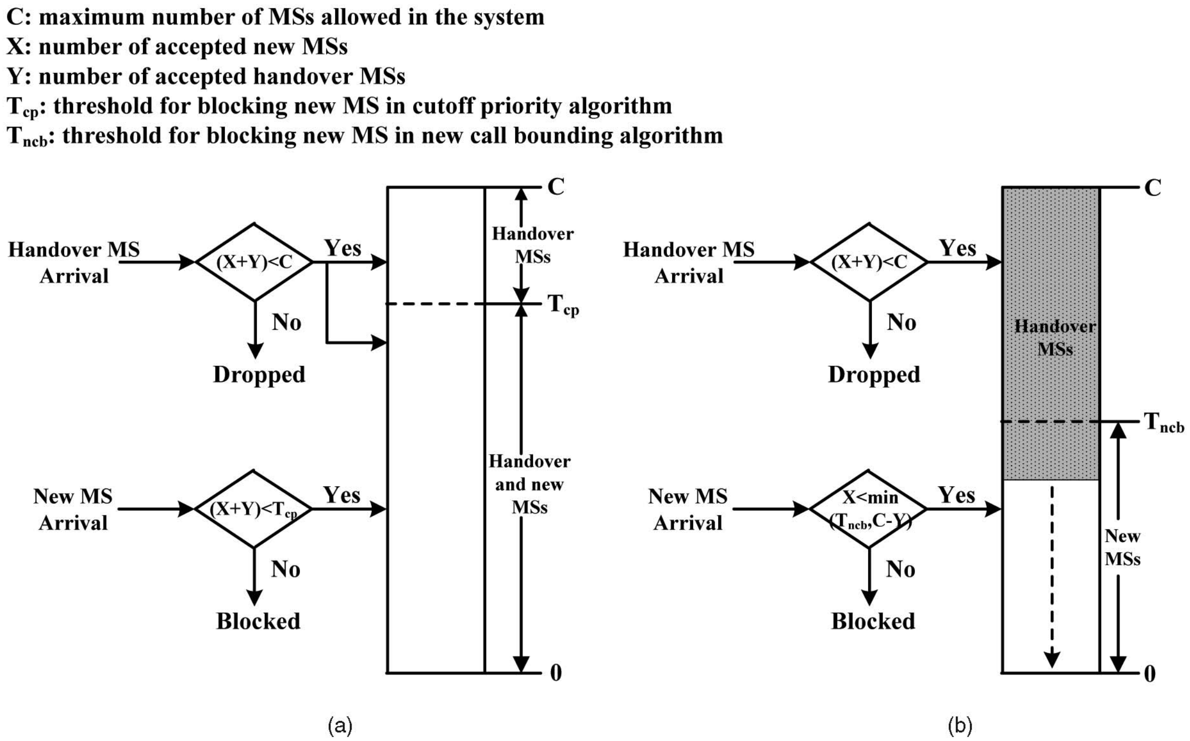
C: maximum number of MSs allowed in the system
X: number of accepted new MSs
Y: number of accepted handover MSs
T_cp: threshold for blocking new MS in cutoff priority algorithm
T_ncb: threshold for blocking new MS in new call bounding algorithm
Handover MS Arrival
(X+Y)<C
Yes
No
Dropped
New MS Arrival
(X+Y)<T_cp
Yes
No
Blocked
C
Handover MSs
T_cp
Handover and new MSs
0
(a)
Handover MS Arrival
(X+Y)<C
Yes
No
Dropped
New MS Arrival
X<min (T_ncb,C-Y)
Yes
No
Blocked
C
Handover MSs
T_ncb
New MSs
0
(b)

Fig. 2. Resource allocation. (a) Cutoff priority algorithm. (b) New call bounding algorithm.

3 Proposed Gateway Relocation Admission Control (GRAC)

The ASN GW relocation may be initiated at different times with different reasons. For example, as aforementioned discussion, an MS may perform ASN Anchored Mobility without performing CSN Anchored Mobility to reduce handover latency. After the handover is completed (i.e., the handover delay has been reduced), the MS may perform ASN GW relocation immediately so the number of Anchored MSs can be kept small. However, it may not be a good strategy always to relocate an Anchored MS so quick. For example, an MS may move fast and keep changing its Serving ASN GW. In this example, it might be better to keep the Anchored ASN GW unchanged. In some other examples, if the system load is light, there is no emergent need to perform ASN GW relocation. However, when more and more MSs are served by two ASN GWs, the system load will become heavy. New users may be blocked. Handover users may be dropped as well. The network performance may be reduced significantly. Therefore, performing ASN GW relocation is essential.

In WiMAX standards [4], [5], it is specified that ASN GW can decide when to perform ASN GW relocation. In this paper, we consider that the system load is heavy so Anchored MSs are forced to perform ASN GW relocation. The proposed GRAC determines when to request Anchored MSs to perform ASN GW relocation and how many Anchored MSs should be relocated. After ASN GW relocation, resources are released and system performance is improved.

Because WiMAX is based on all-IP network architecture, a variety of services, including voice and data services, can be deployed. Unlike voice traffic, data traffic tends to be bursty. Therefore, it is hard to estimate the resource required in an ASN GW to fulfill the requirements of the MSs the ASN GW is currently serving. If the resource in one ASN GW is overprovisioned, the ASN GW may become a performance bottleneck. Another approach is that the number of BSs controlled by each ASN GW can be scaled down to prevent the resource overprovision. However, because the number of BSs controlled by each ASN GW is reduced, this will cause many inter-ASN handovers. As a result, this approach will incur high cost. In [22], the authors discuss the flat mobile WiMAX network architecture. The paper shows that the resource management problem in the ASN GW has a great impact on the performance of WiMAX network architecture. Besides, in WiMAX, the AC algorithm can be deployed in each ASN GW to limit the maximum number of MSs to ensure network service quality. Our goal is to design a *stand-alone* algorithm such that each ASN GW can determine when to request Anchored MSs to perform ASN GW relocation. The proposed algorithm does not need to exchange information between neighboring ASN GWs. It also does not require centralized coordination and any assistance from extra servers. In addition, the proposed algorithm does not need to predict the movement of the mobile stations. It combines AC algorithm with a prediction technique to determine when is necessary to perform ASN GW relocation. Thus, it is called *Gateway Relocation AC (GRAC)*. The corresponding parameters used in this section are listed in Table 1.

The proposed GRAC consists of two components. The first one is AC algorithm, which is discussed in Section 3.1. The prediction algorithm based on Wiener Process (WP) is then presented in Section 3.2.

TABLE 1
List of Parameters

C	Maximum number of MSs in one ASN GW
T_{ncb}	Threshold for blocking a new MS
T_{wnr}	Threshold for carrying out WP-based prediction
$W(t)$	Number of MSs in one ASN GW at time t
$N_S(t)$	Number of serving MSs in one ASN GW at time t
$N_A(t)$	Number of anchored MSs in one ASN GW at time t
$N_H(t)$	Number of handover MSs in one ASN GW at time t
α	Standard normal random variable
Δt	Prediction time interval
τ	Sampling time interval
k	Number of latest samples
λ_n	Arrival rate of new MSs
λ_h	Arrival rate of handover MSs
$1/\mu_c$	Average connection holding time for new MSs
$1/\mu_n$	Average network residence time for new MSs and handover MSs
p^u_{nb}	Blocking probability of new MSs in upper-bound analysis
p^l_{nb}	Blocking probability of new MSs in lower-bound analysis
p^u_{hd}	Dropping probability of handover MSs in upper-bound analysis
p^l_{hd}	Dropping probability of handover MSs in lower-bound analysis
Θ_u	Average serving rate in upper-bound analysis
Θ_l	Average serving rate in lower-bound analysis
Λ_u	Average signaling overhead in upper-bound analysis
Λ_l	Average signaling overhead in lower-bound analysis

3.1 New Call Bounding AC with ASN GW Relocation

As discussed in Section 2, the ideas of AC algorithms are similar, although they have different names. Basically, the overall resources are partitioned and some resource are preserved for the handover MSs only. The proposed GRAC can work with any AC algorithm. In this section, we simply pick up the new call bounding algorithm. For simplicity, here we assume that the resource assigned to each MS in one ASN GW is equal. The main point is not on a specific AC algorithm. The focus is on how to modify an AC algorithm for the two-tier mobility management in WiMAX.

The proposed GRAC with the new call bounding algorithm is presented in Algorithm 1. In Algorithm 1, we limit the number of Serving MSs and Anchored MSs in one ASN GW. As shown in Fig. 2, C is the maximum number of MSs in the network and T_{ncb} is the limit for the number of new MSs, which have been admitted into the network. Let $W(t)$ denote the total number of MSs in the ASN GW at time t. $W(t)$ consists of $N_S(t)$, $N_A(t)$, and $N_H(t)$, which represent the number of Serving MSs, the number of Anchored MSs, and the number of handover MSs, respectively, at time t. As aforementioned discussion, a new MS admitted into the ASN GW is regarded as a Serving MS. After the MS performs inter-ASN handover to a neighboring ASN, the MS becomes an Anchored MS of the ASN GW. Thus, $N_A(t)$ is increased by 1 but $N_S(t)$ is decreased by 1.

Algorithm 1. New call bounding AC with ASN GW relocation

```
Require: A new or handover MS is requesting to connect with the ASN GW at time $t$.
if a new MS arrives then
  if $N_S(t) + N_A(t) < \min(T_{ncb}, C - N_H(t))$ then
    $N_S(t) \longleftarrow N_S(t) + 1$ /* The new MS is accepted. */
  else if $N_S(t) + N_A(t) = \min(T_{ncb}, C - N_H(t))$ then
    if $N_A(t) > 0$ then
      $N_A(t) \longleftarrow N_A(t) - 1$ /* Requesting one of the
      Anchored MSs to perform ASN GW relocation. */
      $N_S(t) \longleftarrow N_S(t) + 1$ /* The new MS is accepted. */
    else
      The new MS is blocked.
    end if
  end if
else if a handover MS arrives then
  if $W(t) < C$ then
    $N_H(t) \longleftarrow N_H(t) + 1$ /* The handover MS is
    accepted. */
  else
    The handover MS is dropped.
  end if
end if
```

To adapt the new call bounding algorithm into WiMAX networks, the algorithm is modified as:

If $N_S(t) + N_A(t) < \min(T'_{ncb}, C' - N_H(t))$ *and a new MS arrives, the new MS is accepted.*

where $T'_{ncb} \leq T_{ncb}$, $C' \leq C$. How to choose the value of T'_{ncb} and C' will be discussed later. However,

When $N_S(t) + N_A(t) = \min(T'_{ncb}, C' - N_H(t))$ *and* $N_A(t) > 0$, *one anchored MS is requested to perform ASN GW relocation.*

Because one Anchored MS is relocated, the new MS can be accepted. Otherwise, the new MS is blocked. Furthermore, if a handover MS arrives at time t, it is always accepted unless $W(t) = C'$.

As aforementioned discussion, in this paper, we consider that the system load is heavy. Therefore, Anchored MSs are forced to perform ASN GW relocation to accommodate new coming users. Based on this principle, we can set T'_{ncb} as T_{ncb} and C' as C. Thus, an Anchored MS is requested to perform ASN GW relocation only when no more resource is available for a new coming MS. The proposed GRAC does not limit the selection of other parameters for other conditions.

3.2 WP-Based Prediction Algorithm

In the above algorithm, we can set C' as C because a new coming MS can be queued until the resource is available after ASN GW relocation is completed. However, this approach cannot be applied to handover MSs because handover MSs are sensitive to handover latency. The acceptable handover delay is much less than the queuing delay of a new MS. Assuming that a handover MS arrives and C is reached. If the handover MS needs to wait for the ASN GW relocation of one Anchored MS, the handover latency will be too high. Actually, if ASN GW relocation is performed just when a handover MS arrives, it is equivalent to performing both ASN Anchored Mobility and CSN Anchored Mobility. The handover latency cannot be reduced. On the other hand, one may perform ASN GW relocation much earlier than C is reached. However, this may force many Anchored MSs to perform ASN GW relocation, which may not be preferable as already discussed earlier. Thus, for handover MSs, it is critical to perform ASN GW relocation at an appropriate time. Therefore, we propose a prediction algorithm based on *Wiener Process* (WP) which provides a systematic way to determine when to request Anchored MSs to perform ASN GW relocation. In addition, the algorithm can also estimate how many Anchored MSs should be relocated. As we will see later, the proposed algorithm is simple and accurate. The algorithm is described in Algorithm 2.

Algorithm 2. WP-based prediction algorithm

```
Require: At each $\tau$ time interval.
if the number of samples is equal to $k$ then
    the oldest sample is discarded.
end if
$W(t)$ is recorded.
if $W(t) \geq T_{wnr}$ then
    /* Stage 1: Generating the expected drift rate and the standard deviation rate */
    $\hat{\mu}$ is computed by using $k$ samples, $W(t - i\tau)$, and (3).
    $\hat{\delta}$ is computed by using $k$ samples, $W(t - i\tau)$, $\hat{\mu}$, and (4).
    /* Stage 2: Estimating the number of MSs */
    Computing $\Delta W$ by using $\hat{\mu}$, $\hat{\delta}$, and (2).
    Computing $\hat{W}(t + \Delta t)$ by using $\Delta W$, $W(t)$, and (1).
    /* Stage 3: Determining when and how many to perform ASN GW relocation */
    if $\hat{W}(t + \Delta t) > C$ then
        $n \longleftarrow \lceil \hat{W}(t + \Delta t) - C + 1 \rceil$
        $n \longleftarrow \min(n, N_A(t))$
        Requesting $n$ Anchored MSs to perform ASN GW relocation.
    end if
end if
```

Wiener Process has been proven effective in modeling stochastic processes where the values of the random variables are affected by a large number of independent or weakly dependent factors, each with a relatively small impact [18]. The $W(t)$ we want to model is impacted by a large number of factors. These factors are either independent or weakly dependent of each other. For example, $W(t)$ is impacted by the arrival rate of new MSs, arrival rate of handover MSs, average connection holding time, average network residence time, and so on. Based on the definitions and properties of Wiener Process, $W(t)$ is continuous and ΔW $(\Delta W = W(t) - W(t - \Delta t))$ follows normal distribution. However, according to the Central Limit Theorem (CLT) [33], we can use a normal distribution to approximate ΔW although $W(t)$ we want to model is integer-valued. The reason is that the variation of ΔW is large and the probability distribution of ΔW approaches to normal distribution. This suggests that Wiener Process can be used to estimate future value of $W(t)$ based on current and past samples of $W(t)$. By using Wiener Process, $W(t)$ can be modeled as

$$\Delta W = W(t) - W(t - \Delta t) = \alpha\sqrt{\Delta t}, \tag{1}$$

where α is a standard normal random variable and Δt is the prediction time interval. ΔW is the variation from $(t - \Delta t)$ to t. Thus, the quantity of ΔW can be computed. ΔW can be further modeled as a normal random variable for any given Δt. The variation of Wiener Process in the following equation allows the mean and standard deviation of ΔW to be changed over time:

$$\Delta W = \mu\Delta t + \alpha\delta\sqrt{\Delta t}, \tag{2}$$

where μ and δ are constants. As shown in (2), ΔW becomes a normal distributed random variable with mean $\mu\Delta t$ and standard deviation $\delta\sqrt{\Delta t}$. It suggests that the mean and the standard deviation of ΔW, for any given Δt, can be calculated directly from μ and δ. Therefore, μ and δ are referred to as the *expected drift rate* and the *standard deviation rate* of ΔW, respectively. For any given time interval τ, the μ and δ can be estimated based on the mean and variance of the sample values in previous k time intervals. The samples in the time interval $[t - i\tau - \tau, t - i\tau]$ are $W(t - i\tau) - W(t - i\tau - \tau)$, for $i = 0, \ldots, k-1$. Hence, μ can be estimated by $\hat{\mu}$

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=0}^{k-1}(W(t-i\tau) - W(t-i\tau-\tau))}{k\tau} \\ &= \frac{W(t) - W(t-k\tau)}{k\tau}.\end{aligned} \tag{3}$$

Also, $\hat{\delta}$, the estimation of δ, is given by

$$\hat{\delta} = \sqrt{\frac{\sum_{i=0}^{k-1}(W(t-i\tau) - W(t-i\tau-\tau) - \hat{\mu}\tau)^2}{k\tau}}. \tag{4}$$

By using (2)-(4), we can estimate ΔW closely when $k \geq 25$ [34]. Therefore, Wiener Process can estimate $\hat{W}(t + \Delta t)$ in the near future by using current $W(t)$ and ΔW.

Algorithm 2 shows the prediction based on Wiener Process. It is executed at each time interval τ. At the sampling time t, the ASN GW records the sample, $W(t)$. After that, only the latest k samples are recorded in the ASN GW. Also, we define a threshold, T_{wnr}, to determine whether the WP-based prediction should be executed or not. If $W(t) \geq T_{wnr}$, our algorithm uses the k samples and (1)-(4) to estimate $\hat{W}(t + \Delta t)$ to determine whether to perform ASN GW relocation or not. We also use the difference between $\hat{W}(t + \Delta t)$ and C to determine how many Anchored MSs need to perform ASN GW relocation.

The proposed WP-based prediction algorithm directly estimates the average number of MSs, $\hat{W}(t + \Delta t)$, in the near future. According to (2), ΔW is determined by μ, δ, and Δt. In addition, by (3) and (4), μ and δ are affected by $W(t - i\tau)$, $\forall i = 0, \ldots, k-1$. Thus, ΔW is sensitive to the variation of the samples of $W(t)$ and Δt. Therefore, when ΔW is increasing, the algorithm can easily predict that $\hat{W}(t + \Delta t)$ is going to be higher than C because $\hat{W}(t + \Delta t) = W(t) + \Delta W$. Thus, it can request Anchored MSs to relocate their anchor points earlier. Also, by using the difference between $\hat{W}(t + \Delta t)$ and C, the algorithm can easily determine how many Anchored MSs should be relocated. The calculations in Wiener Process are simple. The number of sampling data to be recorded is minimal. It is easy to implement. The proposed WP-based prediction algorithm provides a systematic way to perform ASN GW relocation when system load is heavy.

3.3 Discussion

Comparing with traditional AC algorithms, the proposed GRAC decreases the blocking probability of new MSs in WiMAX networks. To see the reasons behind it, we use *Flow (3)* in Fig. 1 as an example. When deploying traditional AC algorithms without considering ASN GW relocation, a new incoming MS is blocked when $N_S(t) + N_A(t) = \min(T_{ncb}, C - N_H(t))$ in ASN GW A. However, there may be some Anchored MSs served in ASN GW A, that is, $N_A(t) > 0$. In contrast, in the proposed GRAC, ASN GW A will request one of the Anchored MSs to perform ASN GW relocation to relocate the traffic anchor point from ASN GW A to ASN GW B. Therefore, $N_A(t)$ in ASN GW A is decreased. Thus, $N_S(t) + N_A(t) < \min(T_{ncb}, C - N_H(t))$. Therefore, a new incoming MS can be accepted by ASN GW A. Besides, the $W(t)$ in ASN GW B is not increased.

Moreover, in order to reduce the dropping probability of handover MSs, the WP-based prediction algorithm prevents the handover MS from being dropped. When $\hat{W}(t + \Delta t)$ is expected to be larger than C, some Anchored MSs are requested to perform ASN GW relocation. In addition, the number of Anchored MSs requested to perform ASN GW relocation is estimated as $\hat{W}(t + \Delta t) - C$. On the other hand, if $W(t)$ is approaching C but the variation of the samples of $W(t)$ is smooth, the WP-based prediction will not request Anchored MSs to perform ASN GW relocation. This is because the system is not expected to be overloaded although $W(t)$ is approaching C. Thus, even if the system load remains high, the Anchored MSs are not necessary to relocate their traffic anchor points. Thus, signaling overhead can be reduced even though the system load is heavy. In short, the proposed WP-based algorithm can trigger ASN GW relocation at an appropriate time. It can also estimate how many Anchored MSs should be relocated. Thus, the dropping probability of handover MSs can be reduced significantly.

In addition to the WP-based prediction algorithm, in our earlier paper [8], we also discussed some other *predictive* and *nonpredictive* ASN GW relocation algorithms. Because the WP-based prediction algorithm performs best, we adopt it in the proposed GRAC.

4 Performance Analysis

In this section, we propose an analytical model to investigate the performance of the proposed algorithm. In the analysis, the *connection holding time* is defined as the time from an MS connects to the network until it is disconnected. The *network residence time* is the time an MS is served by an ASN GW. We assume each ASN GW has two arrival processes which are Poisson distributed with rate λ_n and λ_h for new MSs and handover MSs, respectively. If a new MS is admitted into the network, we assume the connection holding time and network residence time follow exponential distribution with mean $1/\mu_c$ and $1/\mu_n$, respectively. For a handover MS, only network residence time is required. It is also assumed to be exponentially distributed with mean $1/\mu_n$. The corresponding parameters are also listed in Table 1.

To analyze the proposed GRAC, there are three major factors to be considered—the number of Serving MSs, the number of handover MSs, and the number of Anchored MSs. Intuitively, a 3-D Markov chain may be used to investigate the performance. Unfortunately, the computational complexity of a 3-D Markov chain will be increased dramatically when the number of MSs in the system becomes large. It is computationally infeasible to calculate the solution. Therefore, instead of solving the 3-D Markov chain, we derive the upper bound and lower bound by considering the following two extreme cases:

1. **Upper bound:** If we assume each MS *never* performs ASN GW relocation, it will always be served by two ASN GWs. For each ASN GW, the average service time of new MSs is $1/\mu_c$. That is, the MSs will stay in the ASN GW for the duration of whole connection holding time. It will result in the highest blocking probability for new MSs and dropping probability for handover MSs.

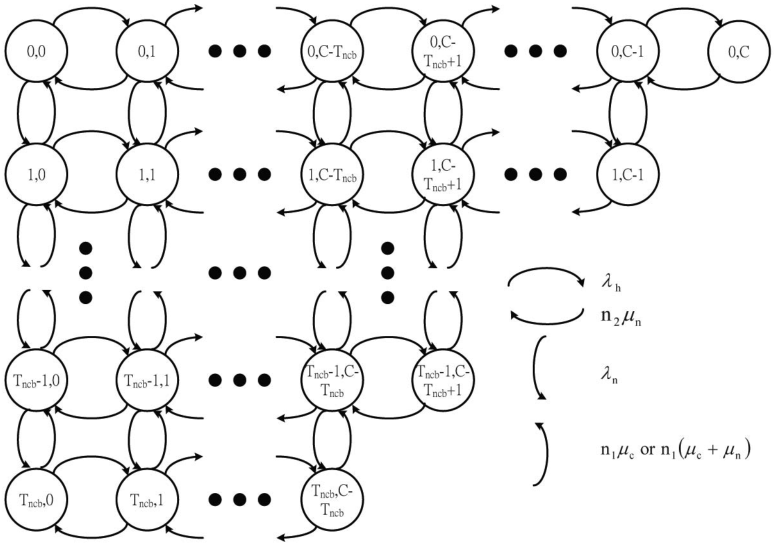

Fig. 3. State transition diagram for the new call bounding algorithm.

2. **Lower bound:** If each MS always performs ASN GW relocation *immediately* after each inter-ASN handover, the average service time of new MSs becomes $1/(\mu_c + \mu_n)$ for each ASN GW. Thus, we will have the lowest blocking probability for new MSs and dropping probability for handover MSs.

In both of the above cases, the average service time of handover MSs is $1/\mu_n$ for each ASN GW. As we will see later, the upper and lower bounds can be derived by a 2D Markov chain.

Next, we calculate the upper bound and lower bound of the proposed GRAC. Because we aim at getting the upper and lower bounds, the WP-based prediction algorithm is irrelevant to the mathematical analysis. The simulation results of the WP-based prediction algorithm will be discussed in Section 5. The following analysis is based on that in [9].

We denote ρ^u_n as λ_n/μ_c for the upper-bound analysis. In the lower-bound analysis, similarly, ρ^l_n equals $\lambda_n/(\mu_c + \mu_n)$. In addition, $\rho_h = \lambda_h/\mu_n$. Let $p^u_{n_1,n_2}$ denote the steady state probability that there are n_1 Serving and Anchored MSs, and n_2 handover MSs in the ASN GW for the upper-bound case. According to the new call bounding AC described above, the state transition diagram can be drawn as a 2D Markov chain, as shown in Fig. 3. Hence, for the upper-bound case, we can obtain the global balance equations as follows:

$$\begin{aligned}(\lambda_n + \lambda_h)p^u_{n_1,n_2} &= (n_2+1)\mu_n p^u_{n_1,(n_2+1)}\\ &\quad + (n_1+1)\mu_c p^u_{(n_1+1),n_2}, \quad n_1 = 0, n_2 = 0.\end{aligned} \tag{5a}$$

$$\begin{aligned}(\lambda_n+\lambda_h+n_2\mu_n)p^u_{n_1,n_2} &= \lambda_h p^u_{n_1,(n_2-1)} + (n_2+1)\mu_n p^u_{n_1,(n_2+1)}\\ &\quad + (n_1+1)\mu_c p^u_{(n_1+1),n_2}, \qquad n_1 = 0, 1 \le n_2 < C.\end{aligned} \tag{5b}$$

$$(n_2\mu_n)p^u_{n_1,n_2} = \lambda_h p^u_{n_1,(n_2-1)}, \quad n_1 = 0, n_2 = C. \tag{5c}$$

$$\begin{aligned}(\lambda_n+\lambda_h+n_1\mu_c)p^u_{n_1,n_2} &= \lambda_n p^u_{(n_1-1),n_2} + (n_2+1)\mu_n p^u_{n_1,(n_2+1)}\\ &\quad + (n_1+1)\mu_c p^u_{(n_1+1),n_2}, \qquad 1 \le n_1 < T_{ncb}, n_2 = 0.\end{aligned} \tag{5d}$$

$$\begin{aligned}(\lambda_n+\lambda_h+n_1\mu_c + n_2\mu_n)p^u_{n_1,n_2} &= \lambda_n p^u_{(n_1-1),n_2} + \lambda_h p^u_{n_1,(n_2-1)}\\ &\quad + (n_2+1)\mu_n p^u_{n_1,(n_2+1)} + (n_1+1)\mu_c p^u_{(n_1+1),n_2},\\ &\qquad 1 \le n_1 < T_{ncb},\\ &\qquad 1 \le n_2 < C - n_1.\end{aligned} \tag{5e}$$

$$\begin{aligned}(n_1\mu_c + n_2\mu_n)p^u_{n_1,n_2} &= \lambda_n p^u_{(n_1-1),n_2} + \lambda_h p^u_{n_1,(n_2-1)},\\ &\qquad 1 \le n_1 \le T_{ncb}, \qquad n_2 = C - n_1.\end{aligned} \tag{5f}$$

$$\begin{aligned}(\lambda_h + n_1\mu_c)p^u_{n_1,n_2} &= \lambda_n p^u_{(n_1-1),n_2} + (n_2+1)\mu_n p^u_{n_1,(n_2+1)},\\ &\qquad n_1 = T_{ncb}, n_2 = 0.\end{aligned} \tag{5g}$$

$$\begin{aligned}(\lambda_h + n_1\mu_c + n_2\mu_n)p^u_{n_1,n_2} &= \lambda_n p^u_{(n_1-1),n_2} + \lambda_h p^u_{n_1,(n_2-1)}\\ &\quad + (n_2+1)\mu_n p^u_{n_1,(n_2+1)}, \qquad n_1 = T_{ncb},\\ &\qquad 1 \le n_2 < C - n_1.\end{aligned} \tag{5h}$$

$$\sum_{n_1=0}^{T_{ncb}} \sum_{n_2=0}^{C-n_1} p^u_{n_1,n_2} = 1 \tag{5i}$$

By solving (5a)-(5i), the steady-state probability distribution can be obtained as follows:

$$p^u_{n_1,n_2} = \frac{(\rho^u_n)^{n_1}}{n_1!} \cdot \frac{(\rho_h)^{n_2}}{n_2!} \cdot p^u_{0,0}, \qquad 0 \le n_1 \le T_{ncb}, \quad 0 \le n_1 + n_2 \le C, \tag{6}$$

where

$$\begin{aligned}p^u_{0,0} &= \left[\sum_{0\le n_1 \le T_{ncb}, 0 \le n_1+n_2 \le C} \frac{(\rho^u_n)^{n_1}}{n_1!} \cdot \frac{(\rho_h)^{n_2}}{n_2!}\right]^{-1}\\ &= \left[\sum_{n_1=0}^{T_{ncb}} \frac{(\rho^u_n)^{n_1}}{n_1!} \cdot \sum_{n_2=0}^{C-n_1} \frac{(\rho_h)^{n_2}}{n_2!}\right]^{-1}.\end{aligned} \tag{7}$$

Thus, we can obtain the blocking probability, p^u_{nb}, of new MSs and the dropping probability, p^u_{hd}, of handover MSs as follows:

$$p^u_{nb} = \frac{\sum_{n_2=0}^{C-T_{ncb}} \frac{(\rho^u_n)^{T_{ncb}}}{T_{ncb}!} \cdot \frac{(\rho_h)^{n_2}}{n_2!} + \sum_{n_1=0}^{T_{ncb}-1} \frac{(\rho^u_n)^{n_1}}{n_1!} \cdot \frac{(\rho_h)^{C-n_1}}{(C-n_1)!}}{\sum_{n_1=0}^{T_{ncb}} \frac{(\rho^u_n)^{n_1}}{n_1!} \sum_{n_2=0}^{C-n_1} \frac{(\rho_h)^{n_2}}{n_2!}}, \qquad (8)$$

$$p^u_{hd} = \frac{\sum_{n_1=0}^{T_{ncb}} \frac{(\rho^u_n)^{n_1}}{n_1!} \cdot \frac{(\rho_h)^{C-n_1}}{(C-n_1)!}}{\sum_{n_1=0}^{T_{ncb}} \frac{(\rho^u_n)^{n_1}}{n_1!} \sum_{n_2=0}^{C-n_1} \frac{(\rho_h)^{n_2}}{n_2!}}. \qquad (9)$$

With the steady-state probabilities, the average serving rate, Θ_u, is given by

$$\Theta_u = \sum_{n_1=0}^{T_{ncb}} \sum_{n_2=0}^{C-n_1} \left(\frac{\lambda_n}{\rho^u_n} \cdot n_1 + \mu_n \cdot n_2 \right) \cdot p^u_{n_1,n_2}. \qquad (10)$$

For the lower-bound case, we can use the similar approach to obtain $p^l_{n_1,n_2}$, p^l_{nb}, p^l_{hd}, and Θ_l by replacing μ_c with $(\mu_c + \mu_n)$ in (5a)-(5h) and ρ^u_n with ρ^l_n in (6)-(10).

We also consider the signaling overhead generated by executing ASN GW relocation. The average signaling overhead is defined as the average number of ASN GW relocation performed in the system. It is denoted as Λ_u and Λ_l for the upper-bound case and lower-bound case, respectively. As discussed above, the ASN GW relocation is never performed in the upper-bound case. Thus, the average signaling overhead is given by

$$\Lambda_u = 0. \qquad (11)$$

On the other hand, in the lower-bound case, the ASN GW relocation is immediately executed after an MS performs inter-ASN handover. The average signaling overhead is given by

$$\Lambda_l = \sum_{n_1=0}^{T_{ncb}} \sum_{n_2=0}^{C-n_1} \mu_n \cdot n_1 \cdot p^l_{n_1,n_2}. \qquad (12)$$

The computational complexity in the proposed analytical models is low. In our analytical model, μ_n and μ_c are two important factors for the performance results. As mentioned above, the average service times of a new MS for upper-bound case and lower-bound case are $1/\mu_c$ and $1/(\mu_c + \mu_n)$, respectively. If $\mu_c \gg \mu_n$, μ_c becomes the dominating factor and the upper and lower bounds are almost equal. Besides, if the difference between μ_n and μ_c in a new MS is large, the average service times of upper and lower bounds are relatively different. Thus, the difference in the performance results between upper bound and lower bound is large. Moreover, when μ_n is increased but μ_c keeps unchanged, the blocking and dropping probabilities of the lower-bound case also become relatively lower than that of the upper-bound case according to (8) and (9). Therefore, our analytical model suggests that when to perform ASN GW relocation has a great impact on the performance of WiMAX networks.

5 Numerical Results

This section provides the numerical results for the analysis presented in Section 4. The analysis is validated by extensive simulations by using *Network Simulator-version 2 (ns-2)* [35]. The analytical results of both upper-bound and lower-bound cases are close to the simulation results. In addition to the upper-bound analysis and lower-bound analysis, we also provide simulation results for the proposed GRAC with WP-based prediction. The parameters and values used in simulations are listed in Table 2. The following sections present the results with various performance metrics. The results are based on exponential distribution for connection holding time and network residence time. We have also conducted simulations by using *gamma distribution* to model connection holding time and network residence time with mean $1/\mu_c$ and $1/\mu_n$. The results are similar to those shown in Figs. 4, 5, 6, 7, 8, 9, 10, 11. Due to space limitation, we only present the results by using exponential distribution.

TABLE 2
Parameters for Simulation

Parameter	Value
C	50
T_{ncb}	25
$1/\mu_c$	1000 (s)
T_{wnr}	45
τ	5 (s)
k	25
α	$N(0,1)$

5.1 Blocking Probability of New MSs

Fig. 4 depicts the blocking probability of new MSs when λ_n is varied from 0.01 (1/s) to 0.1 (1/s). We set $\lambda_h = 0.04$ (1/s) and $1/\mu_n = 400$ (s). As expected, for both upper-bound and lower-bound cases, the blocking probability increases significantly when λ_n increases. Nevertheless, Fig. 4 shows that the blocking probability of the proposed GRAC is close to that of the lower-bound case regardless of the value of Δt. This is because our algorithm can appropriately request Anchored MSs to perform ASN GW relocation when a new MS arrives.

We also investigate the blocking probability with different mean network residence time, $1/\mu_n$, as shown in Fig. 5. In this case, we choose $\lambda_n = 0.04$ (1/s) and $\lambda_h = 0.04$ (1/s). When $1/\mu_n$ increases, the MSs will be served by the ASN GW longer. Thus, they perform inter-ASN handover less. Therefore, the blocking probability in the lower-bound case and the proposed GRAC is increased even if λ_n and λ_h are fixed. On the other hand, because the new MSs never perform ASN GW relocation, the blocking probability of the upper-bound case is irrelevant to $1/\mu_n$. Therefore, it remains constant. Comparing the upper-bound case with the lower-bound case, when $1/\mu_n$ is much lower than $1/\mu_c$, many new MSs become Anchored MSs. The incoming new MSs can be accepted easily by requesting the Anchored MSs to perform ASN GW relocation in the lower-bound case. Thus, the difference of the blocking probability between upper-bound case and lower-bound case is relatively large.

5.2 Dropping Probability of Handover MSs

Fig. 6 illustrates the dropping probability of handover MSs when λ_n is varied from 0.01 (1/s) to 0.1 (1/s). As that in Section 5.1, we set $\lambda_h = 0.04$ (1/s) and $1/\mu_n = 400$ (s). When

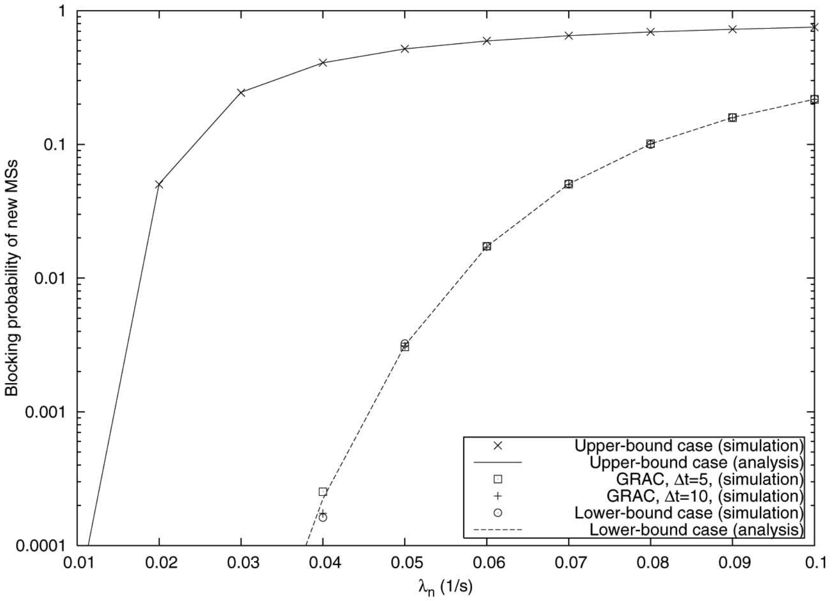

Fig. 4. Blocking probability versus arrival rate (λ_n) for new incoming MSs.

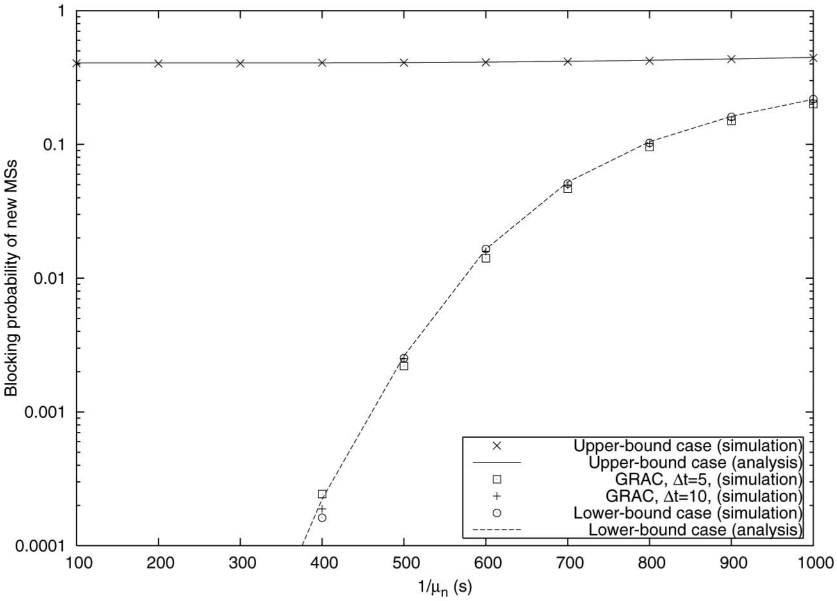

Fig. 5. Blocking probability versus mean network residence time ($1/\mu_n$) for new incoming MSs.

λ_n increases, i.e., there are more MSs in the system, the dropping probability increases too. The handover MS is dropped when C in the AC algorithm is reached. In the proposed GRAC, however, the WP-based prediction is sensitive to the variation of the samples. The Anchored MSs are requested to perform ASN GW relocation when the system is expected to be overloaded. Thus, the dropping probability of handover MSs is reduced significantly. In Fig. 6, we show the simulation results of the proposed GRAC with $\Delta t = 5$ (s) and $\Delta t = 10$ (s). We observe that when $\Delta t = 10$ (s), the dropping probability of the proposed GRAC is very close to that of the lower-bound case. This is because the WP-based prediction is sensitive to the variation of the samples of $W(t)$. According to (1)-(4), when $\Delta W > 0$ and Δt increases, it is easier for the WP-based prediction algorithm to determine whether the system will be overloaded or not in the near future. As a result, the Anchored MSs are requested to perform ASN GW relocation earlier. However, if we set the value of Δt to be a relatively large value, e.g., $\Delta t = 1{,}000$ (s), Δt becomes the only dominating factor and the impact of the variation of the samples is small. Thus, the WP-based prediction algorithm becomes useless. As we can see, when Δt is equal to twice of the sampling interval, τ, the result is very close to that of the lower-bound case. This is sufficient for the WP-based prediction algorithm.

We also investigate the dropping probability with different mean network residence time, $1/\mu_n$, as shown in Fig. 7. In this case, we choose $\lambda_n = 0.04$ (1/s) and $\lambda_h = 0.04$ (1/s). When $1/\mu_n$ increases, the new MSs and handover MSs are served by one ASN GW longer. Thus, they perform inter-ASN handover less. Therefore, the dropping probability is increased even if λ_n and λ_h are fixed. However, unlike the blocking probability shown in Fig. 5,

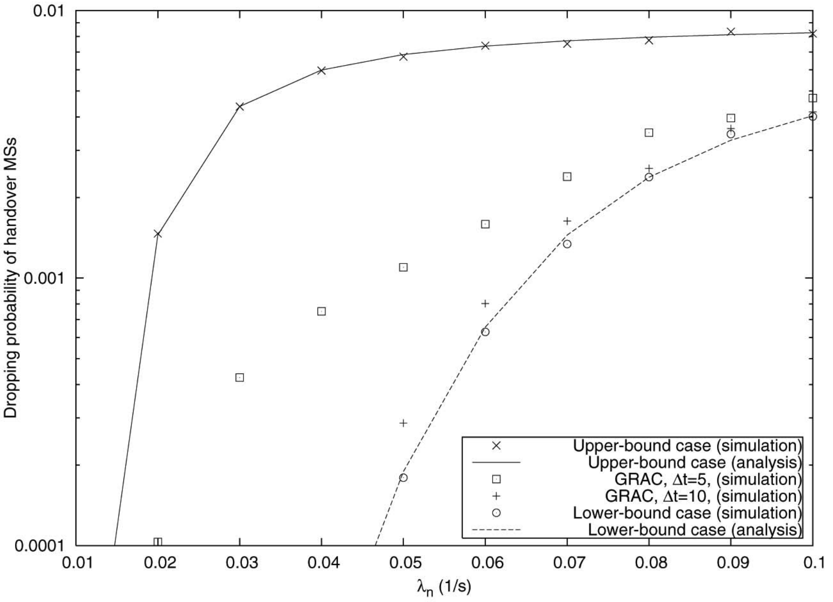

Fig. 6. Dropping probability versus new MS arrival rate (λ_n) for handover MSs.

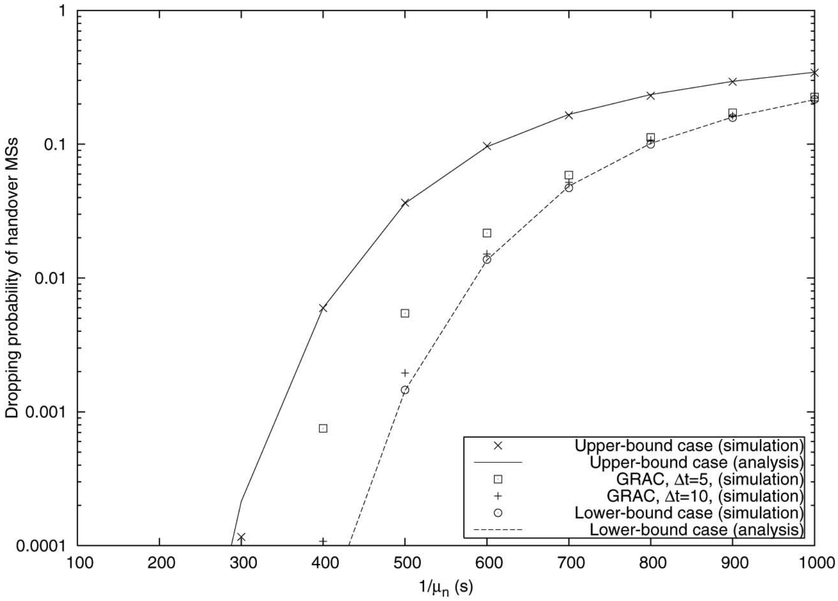

Fig. 7. Dropping probability versus mean network residence time ($1/\mu_n$) for handover MSs.

the dropping probability of the upper-bound case is also increased. This is because the handover MSs are also served by one ASN GW longer. In addition, in the proposed GRAC, the dropping probability of $\Delta t = 10$ (s) is lower than that of $\Delta t = 5$ (s).

5.3 Average Serving Rate

The average serving rate is defined as *the average number of MSs served by an ASN GW per minute*. It includes both new MSs and handover MSs. Fig. 8 presents the average serving rate versus λ_n, where λ_n is varied from 0.01 (1/s) to 0.1 (1/s). We choose $\lambda_h = 0.04$ (1/s) and $1/\mu_n = 400$ (s). As shown in the figure, the upper-bound case and lower-bound case are almost equal when $\lambda_n \leq 0.02$ (1/s). This is because the blocking and dropping probabilities are small in both cases. However, when λ_n increases, the average serving rate of lower-bound case increases faster than that of upper-bound case. This is because the blocking and dropping probabilities in the upper-bound case are higher than those in the lower-bound case. Thus, less MSs are served in the upper-bound case. Please also note that the average serving rate of the proposed GRAC is very close to that of the lower-bound case.

In Fig. 9, we investigate the average serving rate with different mean network residence time, $1/\mu_n$. We also set $\lambda_n = 0.04$ (1/s) and $\lambda_h = 0.04$ (1/s). Fig. 9 shows that the average serving rate decreases when $1/\mu_n$ increases. This is because both new and handover MSs perform inter-ASN handover less. Besides, the average serving rate of the proposed GRAC is close to that of the lower-bound case.

5.4 Average Signaling Overhead

Fig. 10 illustrates the average signaling overhead per minute versus λ_n, where λ_n is varied from 0.01 (1/s) to

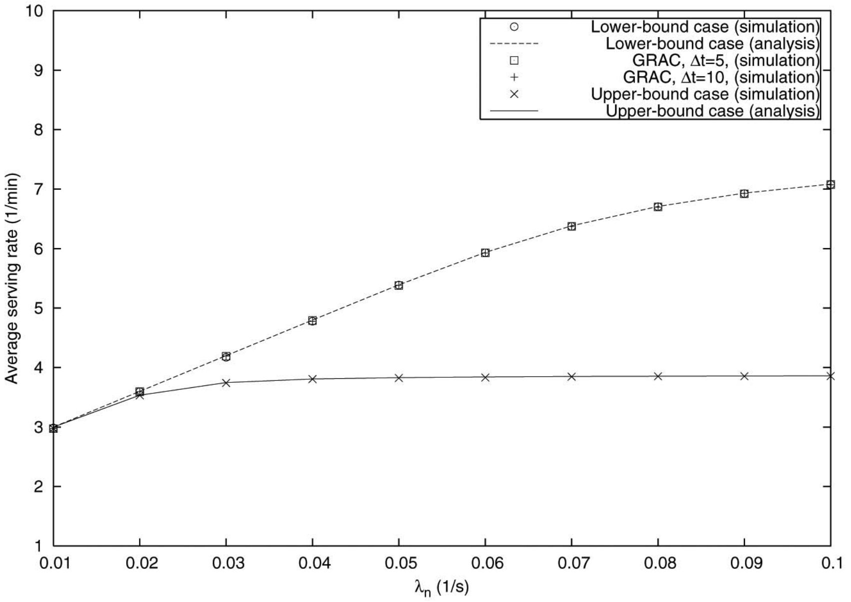

Fig. 8. Average serving rate versus new MS arrival rate (λ_n).

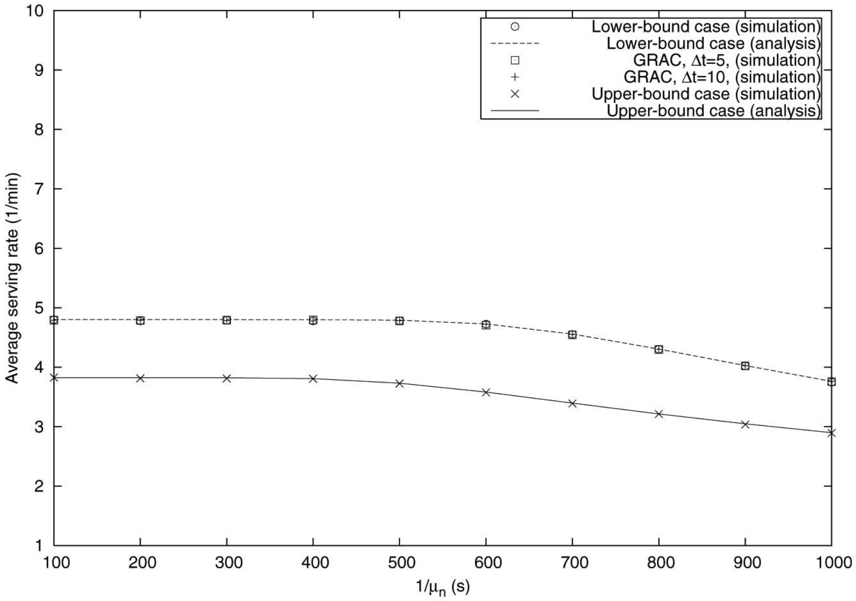

Fig. 9. Average serving rate versus mean network residence time ($1/\mu_n$).

0.1 (1/s). We set $\lambda_h = 0.04$ (1/s) and $1/\mu_n = 400$ (s). The amount of signaling traffic generated by executing CSN Anchored Mobility can be measured by the number of ASN GW relocation performed in the system. As shown in the figure, the signaling overhead of the upper-bound case is 0, because new MSs never perform ASN GW relocation in the upper-bound case. In the lower-bound case, the signaling overhead is increased when λ_n increases. However, the signaling overhead of the proposed GRAC is always lower than that of the lower-bound case. This is because with WP-based prediction, the proposed GRAC can request ASN GW relocation only when the system is expected to be overloaded as that discussed in Section 3.3.

Furthermore, we also investigate the average signaling overhead with different mean network residence time, $1/\mu_n$, as shown in Fig. 11. We still set $\lambda_n = 0.04$ (1/s) and $\lambda_h = 0.04$ (1/s). Again, the signaling overhead of the upper-bound case is 0. For the lower-bound case, when $1/\mu_n$ is small, the signaling overhead is relatively high because the MSs are more likely to perform inter-ASN handover. However, regardless of the variation of $1/\mu_n$, the average signaling overhead of the proposed GRAC almost remains constant.

6 SUMMARY

In WiMAX standards, an ASN GW can decide when to perform ASN GW relocation. In this paper, we consider that the system load is heavy, so Anchored MSs are forced to perform ASN GW relocation. We propose GRAC which considers admission control and ASN GW relocation jointly to improve the performance of WiMAX networks. The traditional AC algorithms cannot be used directly when the

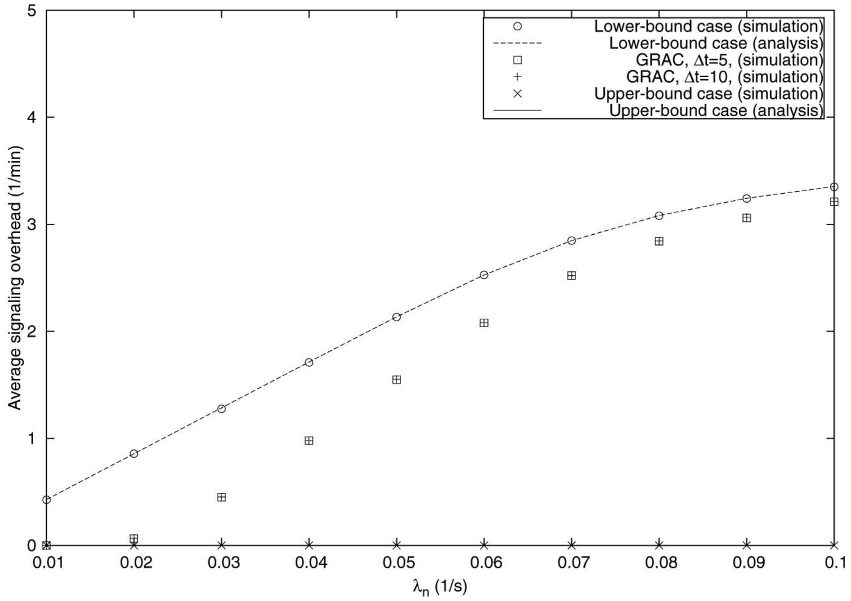

Fig. 10. Average signaling overhead versus new MS arrival rate (λ_n).

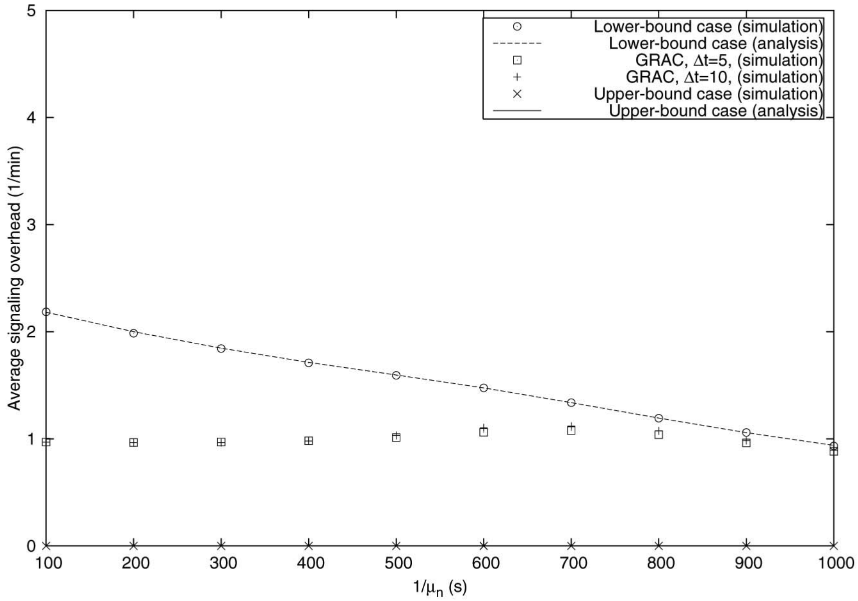

Fig. 11. Average signaling overhead versus mean network residence time ($1/\mu_n$).

two-tiered mobility management is deployed in WiMAX because some MSs may be served by two ASN GWs. If there are many Anchored MSs, new incoming users will likely be rejected due to the lack of resources. In the proposed GRAC, the AC algorithm cooperates with the ASN GW relocation. When a new MS arrives and there is no resource for the newly arrived MS, the proposed GRAC will request an Anchored MS to perform ASN GW relocation. In addition, for handover MSs, the WP-based prediction algorithm can trigger the ASN GW relocation at an appropriate time. It can also estimate how many Anchored MSs should be relocated. We develop an analytical model to investigate the performance of the proposed GRAC. The model analyzes the performance bounds of the system. Extensive simulations are also conducted to validate the analysis and evaluate the performance of the proposed GRAC. The numerical results show that the proposed algorithm can effectively reduce the blocking probability, dropping probability, and average signaling overhead. It also increases the average serving rate.

Acknowledgments

This work was supported in part by the National Science Council (NSC) under Grant 98-2221-E-007-065-MY3.

References

[1] *IEEE 802.16-2004 Std., Air Interface for Fixed Broadband Wireless Access Systems,* IEEE, Oct. 2004.

[2] *IEEE 802.16e-2005 Std., 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, Feb. 2006.

[3] WiMAX Forum, http://www.wimaxforum.org, 2011.

[4] *WiMAX Forum Std. 1.0, Rev. 4, WiMAX Forum Network Architecture (Stage 2: Architecture Tenets, Reference Model and Reference Points),* WiMAX, Feb. 2009.

[5] *WiMAX forum Std. 1.0, Rev. 4, WiMAX Forum Network Architecture (Stage 3: Detailed Protocols and Procedures),* WiMAX, Feb. 2009.

[6] L. Nuaymi, *WiMAX: Technology for Broadband Wireless Access.* John Wiley, 2007.

[7] K. Etemad, "Overview of Mobile WiMAX Technology and Evolution," *IEEE Comm. Mag.,* vol. 46, no. 10, pp. 31-40, Oct. 2008.

[8] Z.-H. Liu, S.-Y. Pan, and J.-C. Chen, "Access Service Network (ASN) Gateway Relocation Algorithms in WiMAX Networks," *Proc. IEEE Int'l Conf. Comm. (ICC '08),* pp. 2674-2679, May 2008.

[9] Y. Fang and Y. Zhang, "Call Admission Control Schemes and Performance Analysis in Wireless Mobile Networks," *IEEE Trans. Vehicular Technology,* vol. 51, no. 2, pp. 371-382, Mar. 2002.

[10] D. Hong and S.S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures," *IEEE Trans. Vehicular Technology,* vol. 35, no. 3, pp. 77-92, Aug. 1986.

[11] Y.-B. Lin, S. Mohan, and A. Noerpel, "Queueing Priority Channel Assignment Strategies for PCS Hand-Off and Initial Access," *IEEE Trans. Vehicular Technology,* vol. 43, no. 3, pp. 704-712, Mar. 1994.

[12] B. Li, S.T. Chanson, and C. Lin, "Analysis of a Hybrid Cutoff Priority Scheme for Multiple Classes of Traffic in Multimedia Wireless Networks," *Wireless Networks,* vol. 4, no. 4, pp. 279-290, 1998.

[13] R. Ramjee, D. Towsley, and R. Nagarajan, "On Optimal Call Admission Control in Cellular Networks," *Wireless Networks,* vol. 3, no. 1, pp. 29-41, 1997.

[14] M.D. Kulavaratharasah and A.H. Aghvami, "Teletraffic Performance Evaluation of Microcellular Personal Communication Networks (PCN's) with Prioritized Handoff Procedures," *IEEE Trans. Vehicular Technology,* vol. 48, no. 1, pp. 137-152, Jan. 1999.

[15] R. Garg and H. Saran, "Fair Bandwidth Sharing Among Virtual Networks: A Capacity Resizing Approach," *Proc. IEEE INFOCOM,* pp. 255-264, Mar. 2000.

[16] J. Yao, J.W. Mark, T.C. Wong, Y.H. Chew, K.M. Lye, and K.-C. Chua, "Virtual Partitioning Resource Allocation for Multiclass Traffic in Cellular Systems with QoS Constraints," *IEEE Trans. Vehicular Technology,* vol. 53, no. 3, pp. 847-864, May 2004.

[17] S.C. Borst and D. Mitra, "Virtual Partitioning for Robust Resource Sharing: Computational Techniques for Heterogeneous Traffic," *IEEE J. Selected Areas in Comm.,* vol. 16, no. 5, pp. 668-678, June 1998.

[18] J.C. Hull, *Options, Futures, and Other Derivatives.* Prentice-Hall, 2000.

[19] P. Iyer, N. Natarajan, M. Venkatachalam, A. Bedekar, E. Gonen, K. Etemad, and P. Taaghol, "All-IP Network Architecture for Mobile WiMAX," *Proc. IEEE Mobile WiMAX Symp.,* pp. 54-59, 2007.

[20] F. Wang, A. Ghosh, C. Sankaran, P. Fleming, F. Hsieh, and S. Benes, "Mobile WiMAX Systems: Performance and Evolution," *IEEE Comm. Mag.,* vol. 46, no. 10, pp. 41-49, Oct. 2008.

[21] J.-H. Yeh, J.-C. Chen, and P. Agrawal, "Fast Intra-Network and Cross-Layer Handover (FINCH) for WiMAX and Mobile Internet," *IEEE Trans. Mobile Computing,* vol. 8, no. 4, pp. 558-574, Apr. 2009.

[22] S. Sim, S.-j. Han, J.-s. Park, and S.-c. Lee, "Seamless IP Mobility Support for Flat Architecture Mobile WiMAX Networks," *IEEE Comm. Mag.,* vol. 47, no. 6, pp. 142-148, June 2009.

[23] Y.-B. Lin and Y.-C. Lin, "WiMAX Location Update for Vehicle Applications," *Mobile Networks and Applications,* vol. 15, no. 1, pp. 148-159, 2010.

[24] E. Fogelstroem, A. Jonsson, and C. Perkins, *Mobile IPv4 Regional Registration,* IETF RFC 4857, June 2007.

[25] J. Xie and I.E. Akyildiz, "A Distributed Dynamic Regional Location Management Scheme for Mobile IP," *Proc. IEEE INFOCOM,* vol. 2, pp. 1069-1078, 2002.

[26] W. Ma and Y. Fang, "Dynamic Hierarchical Mobility Management Strategy for Mobile IP Networks," *IEEE J. Selected Areas in Comm.,* vol. 22, no. 4, pp. 664-676, May 2004.

[27] W. Ma and Y. Fang, "A Pointer Forwarding Based Local Anchoring (POFLA) Scheme for Wireless Networks," *IEEE Trans. Vehicular Technologies,* vol. 54, no. 3, pp. 1135-1146, May 2005.

[28] J.P. Jue and D. Ghosal, "Design and Analysis of a Replicated Server Architecture for Supporting IP-Host Mobility," *Cluster Computing,* vol. 1, no. 2, pp. 249-260, 1998.

[29] A. Vasilache, J. Li, and H. Kameda, "Threshold-Based Load Balancing for Multiple Home Agents in Mobile IP Networks," *Telecomm. Systems,* vol. 22, nos. 1-4, pp. 11-31, Apr. 2003.

[30] H. Deng, X. Huang, K. Zhang, Z. Niu, and M. Ojima, "A Hybrid Load Balance Mechanism for Distributed Home Agents in Mobile IPv6," *Proc. IEEE Int'l Symp. Personal, Indoor, and Mobile Radio Comm. (PIMRC '03),* pp. 2842-2846, Sept. 2003.

[31] A.-C. Pang, Y.-B. Lin, H.-M. Tsai, and P. Agrawal, "Serving Radio Network Controller Relocation for UMTS All-IP Network," *IEEE J. Selected Areas in Comm.,* vol. 22, no. 4, pp. 617-629, May 2004.

[32] S. Pack, T. Kwon, and Y. Choi, "A Mobility-Based Load Control Scheme at Mobility Anchor Point in Hierarchical Mobile IPv6 Networks," *Proc. IEEE GlobeCom,* pp. 3431-3435, Nov./Dec. 2004.

[33] R.V. Hogg and E.A. Tanis, *Probability and Statistical Inference,* seventh ed. Prentice Hall, 2006.

[34] P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods.* Springer Verlag, 1991.

[35] "The Network Simulator - ns-2," http://www.isi.edu/nsnam/ns, 2011.

Zong-Hua Liu received the PhD degree from the Department of Computer Science, National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2010. He was a summer intern at Telcordia Technologies, Piscataway, New Jersey, in 2009. His research interests include mobility management, admission control, resource management, and performance evaluation of wireless networks. Dr. Liu is a student member of the IEEE.

Jyh-Cheng Chen received the PhD degree from the State University of New York at Buffalo in 1998. He is the director of the Institute of Network Engineering and a professor in the Department of Computer Science, National Chiao Tung University (NCTU), Hsinchu, Taiwan. He was with Bellcore/Telcordia Technologies, Morristown, New Jersey, during 1998-2001, and with Telcordia Technologies, Piscataway, New Jersey, during 2008-2010. He has also been with the Department of Computer Science, National Tsing Hua University (NTHU), Hsinchu, Taiwan, since 2001, as an assistant/associate/full/adjunct professor. He coauthored the book *IP-Based Next-Generation Wireless Networks* (Wiley, 2004). He has published more than 80 papers and is the holder of 19 US patents, six ROC patents, and four PRC patents. He received the 2000 Telcordia CEO Award, the 2001 SAIC ESTC (Executive Science and Technology Council) Publication Award, the 2004 NTHU New Faculty Research Award, the 2006 NTHU Outstanding Teaching Award, and the 2007 Best Paper Award for Young Scholars, IEEE Communications Society Taipei and Tainan Chapters & IEEE Information Theory Society Taipei Chapter. He is a technical editor of the *IEEE Wireless Communications*. He was a guest editor of the *IEEE Journal on Selected Areas in Communications* special issue on "All-IP Wireless Networks," May 2004. He was the technical program cochair of the Ninth IFIP/IEEE International Conference on Mobile and Wireless Communications Networks (MWCN 200707) held in Ireland, September 2007. He was the technical program cochair of the Third IEEE International Conference on Information Technology: Research and Education (ITRE 2005). He has been on the technical program committee of numerous international conferences, including IEEE INFOCOM 2005-2006, IEEE GlobeCom 2005-2009, and IEEE ICC 2007-2010. He was a tutorial speaker at IEEE GlobeCom 2002, 2003, and 2006 on the subject of next-generation wireless networks. He leads the development of WIRE1x, which is one of the most important implementations of the IEEE 802.1x supplicant. He is a senior member of the IEEE and a senior member of the ACM.