ELSEVIER

Computer Networks 38 (2002) 631–643

COMPUTER NETWORKS

www.elsevier.com/locate/comnet

Optimization for adaptive bandwidth reservation in wireless multimedia networks

Ki-Dong Lee[a,*], Sehun Kim[b,1]

[a] *Radio and Broadcasting Laboratory, Electronics and Telecommunications Research Institute (ETRI), 161 Kajong-dong, Yusong-gu, Daejon, 305-350 South Korea*

[b] *Department of Industrial Engineering, Korea Advanced Institute of Science and Technology (KAIST), 373-1 Kusong-dong, Yusong-gu, Daejon, 305-701 South Korea*

Received 1 September 2001; accepted 2 October 2001

Responsible Editor: I.F. Akyildiz

Abstract

In future wireless multimedia networks, user mobility management for seamless connection regarding realtime multimedia applications is one of the most important problems. In this paper we propose an opportunity-cost concept-based approach for adaptive bandwidth reservation with admission control for handover calls utilizing network traffic information. Excessive reservation guarantees low blocking probability of handover calls at the cost of high blocking probability of new calls. According to our survey, however, it may degrade bandwidth utilization while no prioritization for handover admissions degrades quality of service (QoS) for ongoing calls. We consider both QoS assurance and bandwidth utilization in order to optimize the amount of bandwidth to reserve for handover admissions. We believe that our scheme could be utilized as a guideline for cost-effective radio resource allocation in mobile multimedia networks.

Keywords: Resource management; Admission control; Optimization; Wireless multimedia

1. Introduction

In wireless communication networks, radio resource management is one of the most interesting issues. Especially, resource management for user mobility has been extensively studied [1–16]. Bandwidth reservation for prioritized handover management is one of hot issues [1,2,5,9,12,14,16] together with admission control [10,11,13]. The evolution toward wireless multimedia networks makes it more important to support a target quality of service (QoS) for each class while maintaining high bandwidth utilization.

There are extensive studies on this issue. Prioritized channel assignment schemes for accommodating handover attempts and optimization studies for assigning channels to given priority classes have been presented in Refs. [1,2]. Utilizing inevitable overlapping area of radio coverage, *directed retry* which remarkably reduces call blocking

* Corresponding author. Tel.: +82-42-860-5225; fax: +82-42-860-5454.

E-mail addresses: kdlee@etri.re.kr (K.-D. Lee), shkim@kaist.ac.kr (S. Kim).

[1] Tel.: +82-42-869-2914; fax: +82-42-869-3110.

1389-1286/01/$ - see front matter
PII: S1389-1286(01)00276-6

probability has been proposed [3]. Performance analysis has been provided using a one-dimensional Poisson approximation of multiple traffic arrivals from neighboring cells and a simplified mobility model [4]. In Ref. [5], Lagrange and Jabbari have proposed an improved scheme to give fairness to mobile users in non-overlapping area and those in overlapping area in a microcellular network based on the proposed scheme in Ref. [3]. Since utilizing the overlapping area is beneficial for traffic performance, it has been applied to *reuse partitioning* in order to improve *reuse efficiency* and blocking probability [6], and to *base station selection* in order to accommodate temporarily uneven traffic and reduce blocking probability of handover requests [7,8]. The authors in Ref. [9] have presented a non-preemptive priority queuing method for handover requests based on a mobile subscriber's power measurements in order to improve QoS while maintaining high spectrum utilization.

In addition, admission policies in multiple-class communication networks have been proposed [10,11]. Sampath and Holtzman [10] have considered probabilistic admission policies of data traffic with dynamically changing the permission probability. The authors in Ref. [12] have proposed an algorithm for bandwidth utilization using the prediction of users' future locations. Using the concept of shadow cluster produced by user movement [12], has shown how base stations determine the probabilities that a user will be active in other cells at future times, and predict resource demands based on the concept. The authors in Ref. [13] have proposed a distributed and adaptive admission control scheme to limit the handover dropping probability utilizing the traffic information of adjacent cells. In a single class service with prioritized handover procedure [21], has newly formulated three problems and proposed efficient solution algorithms with an invented probabilistic admission policy. In the paper, so-called *discounted return formulation* is employed where benefit (cost) in near future is more significant than that in far future because mainly of uncertainty [22]. This is quite different from our algorithm of this paper that does not consider uncertain events in near and far future, but uses estimation reports for very near future traffic condition only and event-sensitive admission control. Oliveria et al. [14] have proposed an adaptive and distributed bandwidth reservation scheme using both local information and remote information in multimedia wireless networks. In Ref. [15], an adaptive call admission control (CAC) algorithm is studied in consideration of interference level. Ramanathan et al. in Ref. [16] have presented strategies for accommodating continuous service to mobile users through estimating resource requirements of potential handover connections based on an invented measure of handover request probability.

Classical models for bandwidth reservation and admission control using Poisson process models provide sufficient accuracy in determining the amount of bandwidth to reserve using a simple blocking probability formula in 2G cellular networks [1–3]. However, the assumptions become somewhat invalid because microcell evolution and various traffic types in 3G multimedia networks make traffic generation and stochastic process of handovers be much more different from Poisson processes.

Recently, several algorithms for adaptive bandwidth reservation have been proposed to guarantee a reasonable QoS to handover calls [15,16]. In some papers, adaptive features of bandwidth reservation are achieved by using estimated probability distributions on the number of future traffic arrivals (i.e., using current network traffic condition and estimated future condition). According to our survey, however, there are no papers on this issue which utilize the full information of (estimated) probability distributions on the number of future traffic arrivals in optimizing the amount of bandwidth to reserve for handovers. Unlike in previous work, we suggest an improved scheme where bandwidth reservation amount is optimized using full information on near future traffic arrivals and an event-sensitive rescheduling method. In optimization for bandwidth reservation, we invent an objective function (called *penalty*) based on inventory control (IC) under uncertain demands. We assume that there is a regression analyzer (an estimation module) within each base station that reports estimators on probability distribution functions (PDFs) on the number of near future traffic arrivals to the bandwidth allocation scheduler.

To optimize the amount of bandwidth reservation using this concept, we formulate an optimization model and suggest an efficient solution algorithm. Performance analysis shows that our approach based on optimization using full traffic information and event-sensitive rescheduling performs very well. The suggested optimization approach for satisfying QoS requirements of multiple classes and achieving high bandwidth utilization under uncertain traffic demand could be utilized as a guideline for cost effective radio resource allocation in wireless multimedia networks.

The rest of this paper is organized as follows. Section 2 presents our model for bandwidth allocation. We discuss traffic estimation in Section 3. Our mathematical formulation for optimizing bandwidth reservation, the property analysis and optimality conditions for the formulation, and the algorithm are presented in Section 4. Performance analysis and discussions can be found in Section 5, and Section 6 provides our conclusions.

2. Model description

2.1. System model

Let us consider a microcell-based network providing the subscribers with multimedia services including delay tolerant and/or intolerant ones. There is a total of s classes of services where we have s_1 classes of realtime services and s_2 classes of non-realtime services ($s = s_1 + s_2$). Without loss of generality, we may assume that class k has the kth highest priority. Our major focus is to minimize the opportunity loss regarding the realtime handover requests of s_1 classes by adaptive admission control sensitive to highly variable network traffic condition in microcellular environments. In our system model, we assume that each base station has a *regression analyzer*: a sub-module that monitors network traffic condition and processes regression analysis about the PDFs of the number of future traffic arrivals for the multiple classes. When a base station requires a set of data, a regression analyzer will report a set of data (the PDFs and some other data regarding network traffic condition) to the corresponding base station.

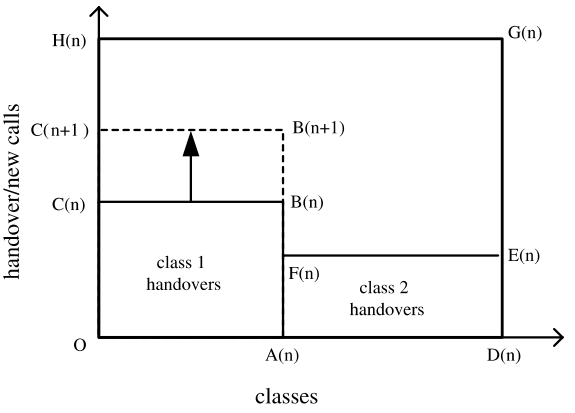

Fig. 1. Bandwidth allocation model: a two-class example.

Fig. 1 shows an example of our bandwidth allocation model for a reference base (station). In the example, the available bandwidth at time step n (exclusive of the bandwidth in use) is denoted by area $[O, D(n), G(n), H(n)]$ and two classes are shown. The figure shows that the amount of bandwidth denoted by area $[O, A(n), B(n), C(n)]$ is reserved for class 1 handover admission at step n and the amount of area $[A(n), D(n), E(n), F(n)]$ is for class 2 handover admission. Suppose that it is estimated that the number of handover requests of class 1 at step $n+1$ is greater than that at step n. Then a bandwidth allocation (BwAlloc) scheduler will reserve more amount of bandwidth at step $n+1$ for class 1 handover admission, i.e., area $[O, A(n), B(n+1), C(n+1)]$ where the partial amount $[B(n), B(n+1), C(n+1), C(n)]$ is additionally reserved comparing to the reserved amount at step n. This is trivial. However, what is focused on in this paper is how much amount of bandwidth reservation should be newly determined for handover admission at every subsequent step for each class according to the changed new traffic condition. Our major focus is to optimize the amount of bandwidth for handover admission under an economically reasonable criterion based on realtime network traffic condition and estimation.

2.2. Traffic model

We consider general traffic arrival and departure processes. Under the assumption in previous

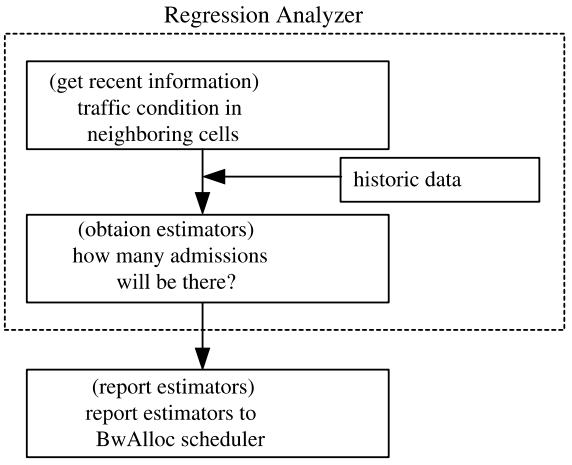

Fig. 2. Role of a regression analyzer: estimation and report procedures of $f_k(x_k|\mathbf{m}_k)$ to a BwAlloc scheduler within a base.

section, PDFs of the number of future traffic arrivals for multiple classes are provided at every step by a regression analyzer (see Fig. 2). All users are able to make calls for a voice-, a data-, a video communication and the like. Each user can make more than one call at the same time. Each class k ($k = 1, \ldots, s_1$) is represented by its effective bandwidth ϕ_k. Realtime traffics have higher priority than non-realtime ones.

3. Traffic estimation

Let us consider a continuous time space where multiple users generate multiple types of calls and complete them. Also, let us consider a series of contiguous time intervals of a constant length (t_w: update period). This generates a new quantized time space. The approach like this is widely used in modeling and analysis. For example, a geometric lifetime model is one of the most popular quantization approaches. In addition to this, we can easily find *discrete-time Markov chain* models which are frequently used in the area of system modeling and analysis. In this paper, we use the concept of the quantization approach for time space.

3.1. Traffic information on handover arrivals

Consider a cell i in a general cellular system. Let N_i be the set of neighboring cells of cell i. Base (station) i is provided with PDFs $f_k(x_k|\mathbf{m}_k(n))$ for $k = 1, \ldots, s_1$ where x_k is the number of arrivals from a regression analyzer at every time step ($n = 1, 2, \ldots$). $f_k(x_k|\mathbf{m}_k(n))$ is the conditional probability distribution for the number of handover requests in cell i for class k given that there are $m_k^j(n)$ ongoing calls for class k in cells $j \in N_i$ where $\mathbf{m}_k(n)$ is a vector which consists of $m_k^j(n)$ for $j \in N_i$. The regression approach to obtain the PDFs for each cell using historic data is desirable because the PDF of channel holding time in each cell might be quite different from the PDFs of other cells depending on many site-environmental factors and the resulting propagation characteristics of a given cell: the layout of roads, the average velocity of mobile users, the average fraction of mobile users in each cell, and the like. But the regression analysis to obtain the PDFs is not a focus of this paper. As mentioned above, we assume that the PDFs are provided for each base by a *regression analyzer* (an estimation module) as shown in Fig. 2. Such an assumption where traffic parameters are given is generally used in the literature [2,4,13]. The differentiability of the distribution functions with the continuous variable is not a necessary condition to apply our scheme. However, we assume they are *twice differentiable* on interval $(0, \infty)$ for the sake of tractability in deriving optimality conditions of our optimization problem. Our approach could be generalized after releasing such assumptions.

3.2. Benefits of utilizing traffic information

Recently, several CAC algorithms have been proposed which utilize network traffic information [12,14,16]. This is because network traffic information is very useful to make an optimal decision for call admission.

If a regression analyzer is based on unbiased and minimum-variance estimation, we have some benefits in using the conditional distributions in such traffic estimation. Most important benefit would be exactness in estimation. Since the conditional distributions are based on *best unbiased* (meaning *minimum variance* and *unbiasedness* in regression theory [20]) estimation and they utilize the current traffic condition, they can provide

adaptive and best information on future traffic condition with sufficient accuracy. With these two concepts, it is expected to make a safe (minimum error) and adaptive decision for bandwidth reservation and admission control.

4. Optimal bandwidth reservation with admission control

4.1. Description of the proposed bandwidth reservation

Let us consider a cell i (we leave out cell index if not confusing). Let BW be the total amount of bandwidth allocated to the base. Then $x_k\phi_k$ is the amount of reserved bandwidth for admitting class k handover requests where x_k is the number of planned admissions. In this bandwidth allocation policy, therefore, each handover request of class k is accepted only if $x_k \geqslant 1$, and a new call request is accepted if there is available bandwidth (excluding the reserved bandwidth). New call requests, regardless of the class, are accepted if there is available bandwidth excluding the reserved bandwidth for handover admission. If great values for $x_1, \ldots, x_{s_1}$ are scheduled, high blocking probabilities of new calls are expected. In contrast, if small values are scheduled, low blocking probabilities of new calls are expected. These facts mean that controlling handover admission with $x_1, \ldots, x_{s_1}$ affects new CAC simultaneously.

Definition 1 (*Lower-priority*). For a class k, each class j (for $j = k+1, \ldots, s_1$) is the lower-priority class of class k. Also, the call admission of class j is called lower-priority admission of class k.

Our main objective is to minimize a penalty function in an economic aspect (presented later in this section) by controlling variables x_ks. In optimizing the values x_ks, we employ *opportunity cost* concept well known in economics. Suppose that x_k is scheduled to be small. Then it is obvious that the blocking probability of class k handovers is high because little opportunity is given to class k handovers. However, it is also obvious that the amount of opportunity for the handovers of its lower-priority classes and the new calls will be great (as a result, lower blocking probabilities are expected). This is because the rest of opportunity not to be given to class k handovers will be contributed to the handovers of the other classes of lower priorities and to the new calls. Conversely, suppose that x_k is scheduled to be large. Then it is obvious that the blocking probability of class k handovers is low because more opportunity is given to class k handovers. However, it is also obvious that the amount of opportunity for the handovers of its lower-priority classes and the new calls will be small (as a result, higher blocking probabilities are expected).

At this point we need to think about two kinds of opportunity loss. If great amount of x_k is scheduled, there will be more opportunity loss of its lower-priority admission. And if little amount of x_k is scheduled, there will be more opportunity loss of class k handover admission. If there is a *stable* vector point $(x_1, \ldots, x_{s_1})$, it is reasonable that the point be used as an optimal solution for bandwidth reservation. We focus on applying this concept to bandwidth reservation problems in multiple-class systems.

The penalty (in the sense of opportunity cost) due to making a reserved bandwidth unit idle (when there are fewer arrivals than reserved) is denoted by a_k (positive real), and the penalty due to making a handover request of class k blocked (the loss of one handover call when there are more arrivals than reserved) is denoted by b_k (positive real) for class k. The values of a_ks and b_ks could be determined based on an economic theory. Since the opportunity cost concept is applied, the values of a_ks would be determined by the values of a_js and b_j for $j = 1, \ldots, k-1$. On the other hand, b_ks are related to the service provider's policy and decision. There could be various versions of valuing b_ks and many different schemes of valuing a_ks. In this work, both of them are assumed to be given. In Fig. 3, the absolute values for the slopes of the two lines are the corresponding penalty factors. In the figure, the penalty for class k when there are x_k arrivals, denoted by $p_k(x_k)$, is given by

$$p_k(x_k) = \frac{a_k + b_k}{2}|x_k - x_k^*| + \frac{b_k - a_k}{2}(x_k - x_k^*) \qquad (1)$$

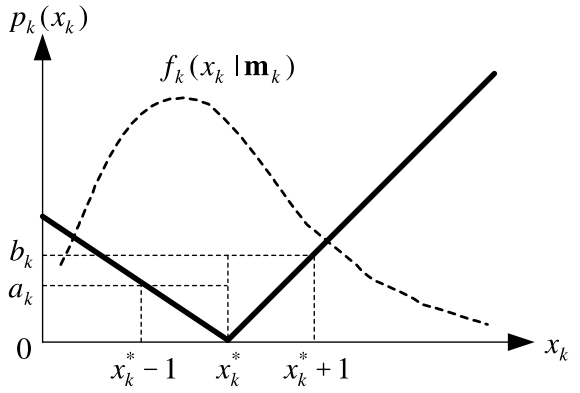

Fig. 3. Penalty function for a given probability density function.

given that x_k^* is the number of reserved bandwidth units. Thus, for example, we can easily find that $x_k^* = \arg\min_{x_k}\{p_k(x_k)\}$, $p_k(x_k^* - n) = na_k$, and $p_k(x_k^* + n) = nb_k$. Similarly, in determining the optimal values of the control variables for multiple classes, we use the penalty vector and the probability distributions. From Eq. (1), the *expected penalty* for class k ($k = 1, \ldots, s_1$) is expressed with a function $H_k : \mathscr{R}^+ \cup \{0\} \to \mathscr{R}$ as

$$H_k(x_k) = a_k \int_0^{x_k} (x_k - y) f_k(y|\mathbf{m}_k)\,\mathrm{d}y + b_k \int_{x_k}^{\infty} (y - x_k) f_k(y|\mathbf{m}_k)\,\mathrm{d}y \tag{2}$$

and the total *expected penalty* is given by a function $H : [\mathscr{R}^+ \cup \{0\}]^{s_1} \to \mathscr{R}$ as

$$H(x_1, \ldots, x_{s_1}) = \sum_{k=1}^{s_1} H_k(x_k). \tag{3}$$

Eq. (2) is a function of penalty under uncertain traffic demand of class k where a concept of IC under uncertain customer demand [19] is applied.

4.2. Formulation of bandwidth reservation problem

Based on the total penalty function defined in Eq. (3), our bandwidth reservation problem is reduced to a penalty minimization problem (PMP_n). We formulate the PMP_n as a *nonlinear convex programming* [18,19] for each cell. Let $\mathbf{x} = (x_1, \ldots, x_{s_1})^{\mathrm{t}}$, $\phi = (\phi_1, \ldots, \phi_{s_1})^{\mathrm{t}}$, and $\mathbf{X} = \{\mathbf{x} | \mathbf{x} \geqslant \mathbf{0}\}$.

(PMP_n)
minimize $\quad H(\mathbf{x})$,
subject to $\quad h(\mathbf{x}) \leqslant 0, \ \mathbf{x} \in \mathbf{X}$

where $h(\mathbf{x}) = (\mathbf{x} + \mathbf{m})^{\mathrm{t}}\phi + E(B(t)|\mathbf{m}) - E_n - \mathrm{BW}$ and vector $\mathbf{m}$ of size s_1 denotes the numbers of ongoing calls in N_i for class $1, \ldots, s_1$ for a reference cell i, i.e., $(m_1, \ldots, m_{s_1})^{\mathrm{t}}$. There could be released bandwidth due to handover departure and call completion during an update period (the length of a time step). If there is released bandwidth, it should be rescheduled to be utilized for potential traffic arrivals. The above formulation (PMP_n) considers such a reallocation problem by introducing a parameter E_n which denotes the amount of increase in bandwidth due purely to bandwidth release. Note that $\sum_{n=1}^{n^*} E_n \leqslant \sum_{k=1}^{s_1} m_k\phi_k$ where equality holds if and only if $n^* = \sum_{k=1}^{s_1} m_k$. For every update period, the initial value of E_0 is zero. If there is an event of bandwidth release, E_1 will have positive value, and so on. For example, if there have been five events of bandwidth release after scheduling for (PMP_0), problem (PMP_5) is to be solved with updated parameter values for bandwidth reallocation.

4.3. Property analysis of (PMP_n)

Some properties of (PMP_n) are analyzed to develop a solution algorithm in this section.

Theorem 1 (*Convexity of* H_k). *$H_k(x) : \mathscr{R}^+ \cup \{0\} \to \mathscr{R}^+$ is a strictly convex function.*

Proof. The first derivative of $H_k(x_k)$ is written as

$$\frac{\mathrm{d}H_k(x_k)}{\mathrm{d}x_k} = (a_k + b_k) \int_0^{x_k} f_k(y|\mathbf{m}_k)\,\mathrm{d}y - b_k \tag{4}$$

and the second derivative is written as

$$\frac{\mathrm{d}^2 H_k(x_k)}{\mathrm{d}x_k^2} = (a_k + b_k) f_k(x_k|\mathbf{m}_k). \tag{5}$$

Since $a_k > 0$ and $b_k > 0$, the second derivative is nonnegative (positive if $f_k(x_k) > 0$) for all $x_k \in (0, \infty)$. This proves the theorem. □

Corollary 1. *$H(\mathbf{x})$ is strictly convex over $\mathscr{R}^{s_1} \geqslant \mathbf{0}$.*

Proof. Any positive sum of strictly convex functions is also strictly convex [18]. □

Note that if there exists an optimum of (PMP_n), then it is the unique optimum. This is because the objective function is strictly convex by Corollary 1. With the convexity of the objective function, the problem is solved easily.

Definition 2. A constraint $h(\mathbf{x}) \leqslant 0$ is said to be '*binding*' at $\mathbf{x}^*$ if $h(\mathbf{x}^*) = 0$ for an optimal vector $\mathbf{x}^*$.

Definition 3. A square matrix Q is said to be 'positive definite' if $\mathbf{y}^{\mathrm{t}} Q \mathbf{y} > 0$ for all $\mathbf{y} \neq \mathbf{0}$.

To solve (PMP_n) we apply *Lagrange multiplier theory* [18]. We first define a *Lagrangian function* $L : \mathscr{R}^{s_1+1} \to \mathscr{R}$ for (PMP_n) as

$$L(\mathbf{x}, \lambda) = H(\mathbf{x}) + \lambda h(\mathbf{x}), \tag{6}$$

where λ is a nonnegative scalar called *a Lagrange multiplier*.

Theorem 2 (*Sufficient conditions for optimality*)**.** *Let $\mathbf{x}^*$ and λ^* satisfy*

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*) = 0, \quad \nabla_{\lambda} L(\mathbf{x}^*, \lambda^*) = \mathbf{0}, \tag{7}$$

$$\mathbf{y}^{\mathrm{t}} \nabla_{\mathbf{xx}} L(\mathbf{x}^*, \lambda^*) \mathbf{y} > 0, \tag{8}$$

for all $\mathbf{y} \neq \mathbf{0}$ with $\nabla h(\mathbf{x}^*)^{\mathrm{t}} \mathbf{y} = 0$. Then $\mathbf{x}^*$ is a strict local minimum.

Proof (*abbreviated*)**.** Our Lagrangian function defined in Eq. (6) is convex. In this special case the proof is very simple. By Eq. (5), $\nabla_{\mathbf{xx}} L(\mathbf{x}^*, \lambda^*)$ is *positive definite*. Thus $(\mathbf{x}^*, \lambda^*)$ satisfying Eq. (7) is the strict local minimum of $L(\mathbf{x}, \lambda)$, which means that $\mathbf{x}^*$ is also the strict local minimum of $H(\mathbf{x})$ subject to $h(\mathbf{x}) \leqslant 0$. (A complete proof can be found in Ref. [18].) □

4.4. Solution procedure for (PMP_n)

We first consider a property of the problem and show a solution algorithm.

Step 1: (Solve $\arg\min_{\mathbf{x} \in \mathbf{X}} H(\mathbf{x})$)

Assume that there is no constraint, then solve (PMP_n) as follows: Since $H(\mathbf{x})$ is strictly convex, $\mathbf{x}^*$ is the optimum such that

$$\nabla_{\mathbf{x}} H(\mathbf{x}^*) = \mathbf{0}. \tag{9}$$

In this case, for each k we can simply find

$$x_k^* = F_k^{-1}\left(\frac{b_k}{a_k + b_k}\right), \tag{10}$$

where F_k is the distribution function [17] for f_k and $F_k^{-1}(x) = \inf\{y | F_k(y) = x\}$. Note that $F_k^{-1}(x)$ is simply the inverse function of $F_k(x)$ if it is *invertible*. If $h(\mathbf{x}^*) \leqslant 0$, then current $\mathbf{x}^*$ is feasible in original problem, thus, it is optimal (Stop). Else, it is infeasible in original problem, and it is clear that $h(\mathbf{x}^*) = 0$, i.e., the constraint is *binding* at the optimum $\mathbf{x}^*$ (go to Step 2).

Step 2: (Solve $\arg\min_{\mathbf{x} \in \{\mathbf{x} | h(\mathbf{x}) = 0, \mathbf{x} \in \mathbf{X}\}} H(\mathbf{x})$)

If the constraint is binding, there exists a nonnegative scalar λ^* such that

$$h(\mathbf{x}) = 0, \tag{11}$$

that is,

$$\mathbf{x}^{\mathrm{t}} \phi = \mathrm{BW} - \mathbf{m}^{\mathrm{t}} \phi - E(B(t) | \mathbf{m}) + E_n, \tag{12}$$

where x_k is represented with λ as

$$x_k = F_k^{-1}\left(\frac{b_k - \lambda \phi_k}{a_k + b_k}\right). \tag{13}$$

This is due to *Lagrange multiplier theorem* or basically to *Farkas' Lemma* [18]. Furthermore, since Eq. (12) is a linear equation of $\mathbf{x}$ and there exists a one-to-one mapping between x_k and λ, there always exists a unique λ^* satisfying the equation. With this, we can find the optimal solution $\mathbf{x}^*$ where

$$x_k^* = F_k^{-1}\left(\frac{b_k - \lambda^* \phi_k}{a_k + b_k}\right). \tag{14}$$

Step 3: (obtain optimal integer solutions)

(3.1) $e_k := x_k^* - [x_k^*]$, for $k = 1, \ldots, s_1$, where $[x] = \max_n \{n \leqslant x, n \in \mathbf{Z}\}$.

(3.2) If $\sum_{k=1}^{s_1} e_k > 0$, allocate currently available bandwidth, i.e., $\sum_{k=1}^{s_1} e_k \phi_k$ again. Solve a binary integer programming (BIP) problem of $\mathbf{x}$ with the objective $H(\mathbf{x})$ and a constraint $\sum_{k=1}^{s_1} (x_k - e_k)\phi_k \leqslant 0$. Let the optimal integer solution for class k be r_k^*.

(3.3) $x_k^* := [x_k^*] + r_k^*$, for $k = 1, \ldots, s_1$.

In Step (3.2), we mention BIP. Efficient solution algorithms for such a BIP are easily found in the literature, for example, Section 13.7 in Ref. [19] could be referred to. For a mobile multimedia network with less than 10 classes of realtime services, *branch-and-bound technique* and *dynamic programming approach* could solve this BIP within a very short time less than a few milliseconds.

Example 1. To explain how our algorithm runs we have a simple example. Let $\mathbf{m} = (3, 5)$, $\phi = (2, 1)$, BW $= 15$, $B(t) = 0$,

$$f_1(x|\mathbf{m}_1) = \tfrac{1}{2}\mathrm{e}^{-x/2}, \tag{15}$$

$$f_2(x|\mathbf{m}_2) = \tfrac{1}{3}\mathrm{e}^{-x/3}. \tag{16}$$

We can easily find that this example is under a congested case. Then Eq. (12) becomes

$$-2\phi_1 \ln\frac{a_1 + \lambda\phi_1}{a_1 + b_1} - 2\phi_2 \ln\frac{a_2 + \lambda\phi_2}{a_2 + b_2} = 4, \tag{17}$$

we obtain

$$\left(\frac{a_1 + b_1}{a_1 + \lambda\phi_1}\right)^2 \left(\frac{a_2 + b_2}{a_2 + \lambda\phi_2}\right) = \mathrm{e}^2. \tag{18}$$

Given that $(a_1, b_1, a_2, b_2) = (1, 4, 0.5, 2)$, we obtain

$$\lambda^* = \frac{5\mathrm{e}^{-2/3} - 1}{2}. \tag{19}$$

Then we obtain the optimal solution for (PMP_0) is

$$x_1^* = \tfrac{4}{3}, \quad x_2^* = \tfrac{4}{3}.$$

Thus $8/3$ units and $4/3$ units are reserved to admit handover requests of classes 1 and 2, respectively, at update period start. If an integer solution vector is required, the optimal solution will be $x_1^* = 1$ and $x_2^* = 2$ because the corresponding BIP has the optimal solution $(e_1^*, e_2^*) = (0, 1)$.

4.5. Bandwidth reservation algorithm with admission control

In our algorithm, the amount of bandwidth reservation for each class is determined in Step 1, and the amount of additional reservation due to each event of bandwidth (channel) release is determined in Step 2. During a given update period, Step 3 runs for admission control following the thresholds determined in Steps 1 and 2.

Step 1 (Solve (PMP_0)): Obtain an optimal solution $\mathbf{x}^*$.

Step 2 (Solve (PMP_n): event-sensitive rescheduling)

(2.1) Obtain an optimal solution $x_{n,1}, \ldots, x_{n,s_1}$ for each case of n.

(2.2) If $x_{i,k} > x_{i+1,k}$, resolve (PMP_{i+1}) with fixing $x_k^* = x_{i,k}$ and obtain a new optimal solution. Repeat (2.2) until $x_{i,k} \leqslant x_{i+1,k}$, $\forall k = 1, \ldots, s_1$.

(2.3) If $x_{i,k} \leqslant x_{i+1,k}$, $\forall k = 1, \ldots, s_1$, build a matrix

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,s_1} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,s_1} \end{bmatrix}.$$

Step 3 (Admission control):
If (class k arrival)
  If $x_k \geqslant 1$
    Accept connection;
    $x_k := x_k - 1$;
  Else Reject connection;
Else If (departure)
  $x_k := x_k + \Delta_{i+1,k}$ for $\forall k\ \Delta_{i+1,k} > 0$;
where $\Delta_{i+1,k} = (x_{i+1,k} - x_{i,k})$.

5. Performance analysis

5.1. Simulation model

We consider four performance measures: the blocking probability of new calls, the blocking (dropping) probability of handover calls, the penalty newly defined in this paper, and the amount of idle bandwidth. To compare the performance, we consider two previous schemes: scheme-1 [14], and scheme-2 [16]. A 19-cell system model is considered during rush hour and during non-rush hour. In this simulation, the traffic arrival events are generated from truncated (by $[0, z]$) gamma-2 distributions, e.g.,

$$f(x) = \frac{f_{\mathrm{gamma}}(x)}{\int_0^z f_{\mathrm{gamma}}(y)\,\mathrm{d}y}$$

with different values of z. We consider two cases of penalty factors: (Case 1) $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 2.0$; (Case 2) $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 1.8$. In this paper, the penalty factors for three classes considered are just examples. In Case 1, we

consider one of the simplest cases that the average loss of one excessive arrival is twice penalized comparing with the loss of opportunity cost due to excessive reservation. In Case 2, we consider a case that each opportunity cost is greater than that used in Case 1. In our simulation, all of the distributions in 19 cells follow truncated gamma-2 distributions (the parameter(s) could be different for each cell).

5.2. Simulation results and discussion

As shown in our simulation results, it is observed that our scheme is dominant over the other two schemes. This is because, unlike our scheme, schemes-1 and -2 do not utilize full information included in the probability distributions in their decision making. The bandwidth utilization of our scheme and that of scheme-2 dominate that of scheme-1 because of sensible and selective bandwidth reservation. There are relatively too much bandwidth reserved in neighboring cells in scheme-1.

We think that the following example gives easy understanding about the performance of our scheme from simulations.

Example 2. Let us consider a case for class k where the expected number of handover arrivals is 11, the expected number of departures (handover departure and call completion) is 4, and the probability distribution of pure increase of ongoing calls follows an exponential distribution with mean 7 $(= 11 - 4)$. Under the condition that there are appropriate sufficiency of available bandwidth, scheme-2 [16] determines that required bandwidth for handover admission is $6.955\phi_k$, i.e.,

$$\overline{Y}_k = \sum_{s\in S(4,11)} Y_k(s)\frac{1}{|S(4,11)|} = \sum_{s\in S(4,11)} Y_k(s)\frac{1}{1365} = 6.955.$$

However, the decision based on our scheme is

$$x_k^* = 7\ln\left(\frac{a_k}{a_k+b_k}\right),$$

and if $(a_k, b_k) = (1, 2)$ is known, then $x_k^* = 7.690$ (if $(a_k, b_k) = (1.2, 2)$, then $x_k^* = 6.866$).

The difference between the results of scheme-2 and our scheme results from the following fact: scheme-2 uses the expected values (partial information of a PDF) for arrivals and departures as a consideration bound, thus there is no further concern regarding excessive/fewer arrivals and excessive/fewer departures with respect to the PDF, respectively. If the variance of the distribution considered is very small or close to zero, the bound is sufficient for an optimal decision. In general cases, however, it is more beneficial in decision making to use full information of a PDF than to use its partial information. In this case,

$$H_k(x) = a_k x - \tfrac{1}{7}a_k + (a_k + b_k)\tfrac{1}{7}e^{-x/7},$$

and then we obtain $H_k(x_k^*) = 7.69$ and $H(\overline{Y}_k) = 7.73$.

Aggregated blocking probabilities on our three realtime classes are shown in Figs. 4 and 5. In the figures, it is observed that scheme-2 and ours show dominant performance over scheme-1, i.e., it shows that all of the performance measures considered of scheme-1 are not better than those of scheme-2 and our scheme. This is because scheme-1 reserves too much bandwidth of neighboring cells which is possibly unnecessary. Also, since we consider non-uniform traffic condition, the amount of reserved bandwidth in such cells with low traffic load does not contribute to reducing blocking probability and increasing bandwidth utilization, rather it contributes to degrading those two measures in cells with relatively heavy traffic load.

In Fig. 5(b), the blocking probability of handover calls in our scheme is slightly dominated by the other two schemes. However, smaller blocking probabilities of handover calls achieved with schemes-1 and -2 do not compensate for the opportunity loss of lower-priority admission at $((b_1 + b_2 + b_3)/(a_1 + a_2 + a_3)) = 1.8$. Our scheme dominates the other two schemes in the other three measures. Also, since our decision is based on the third measure (i.e., penalty), being ranked over the other two schemes in penalty (see Fig. 6) should be focused. In Fig. 7, it is observed that our algorithm shows highly adaptive feature to temporarily uneven traffic condition. This is owing to an event-sensitive rescheduling function of the proposed scheme.

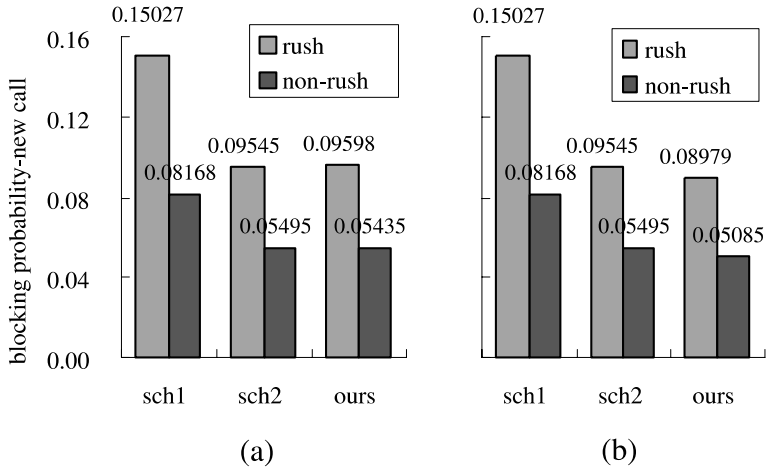

Fig. 4. Blocking probability at $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 2.0$: (a) new call blocking and (b) handover dropping.

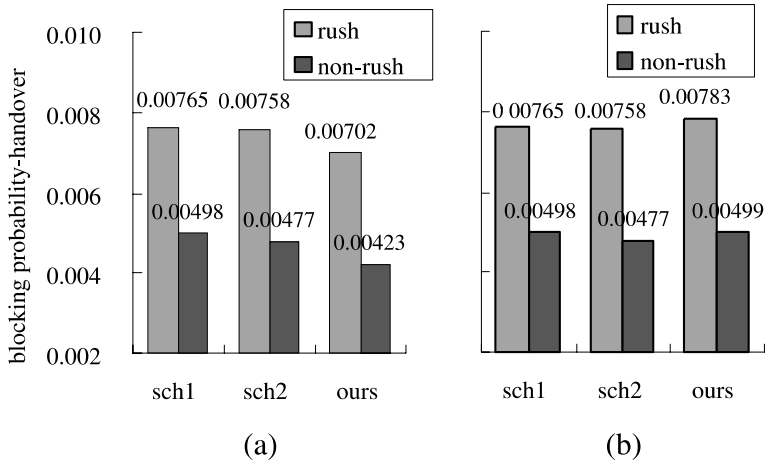

Fig. 5. Blocking probability at $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 1.8$: (a) new call blocking and (b) handover dropping.

6. Summary and conclusions

This paper has focused on optimizing the amount of bandwidth to reserve for handover admissions in wireless multimedia networks under uncertain traffic demand. The proposed scheme is quite different from the previous ones because both QoS guarantee and resource utilization are reflected in the objective function in order to be optimized using estimated full information of future traffic condition. In addition, two different penalty factor classes are considered in our objective: penalty due to making a reserved bandwidth unit idle (in case that there are too much bandwidth reserved), and penalty due to making a handover request blocked (in case that there are too little bandwidth reserved, i.e., lack of reserved bandwidth). Owing to utilizing sufficient information on estimated traffic parameters, more exact and reasonable solutions for bandwidth

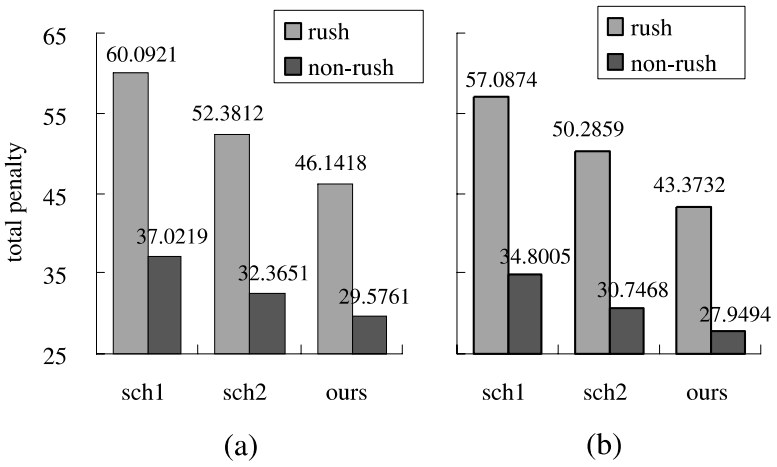

Fig. 6. Total penalty: (a) $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 2.0$ and (b) $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 1.8$.

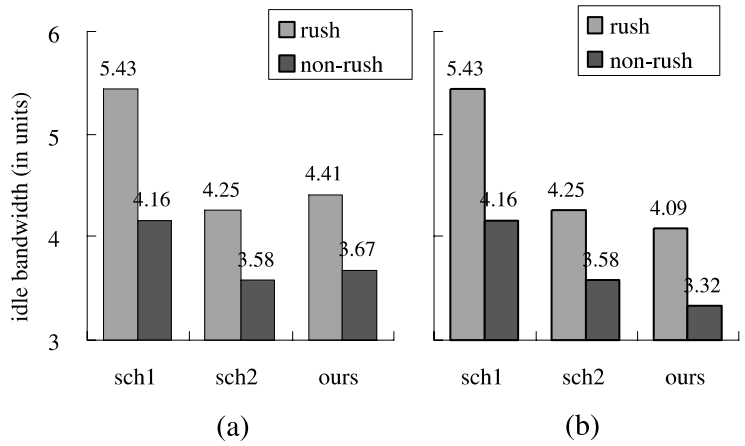

Fig. 7. Idle bandwidth: (a) $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 2.0$, (b) $(b_1 + b_2 + b_3)/(a_1 + a_2 + a_3) = 1.8$.

reservation under uncertain traffic condition are expected. In addition, our algorithm is highly adaptive to temporarily uneven network traffic condition because it has an event-sensitive re-scheduling function, which improves the utilization of valuable bandwidth resource.

Such a set of probable phenomena in a system with limited available resource is formulated as a mathematical programming, and an exact and efficient solution algorithm is studied. Performance analysis shows that our scheme performs very well. We believe that our scheme could be utilized as a decision guideline for optimal radio resource allocation in mobility management and target QoS achievement.

Acknowledgements

The authors are very grateful to the anonymous reviewers and (Dr.) H.Y. Kim for their valuable comments. This work was prepared during the

authors' work at KAIST and supported in part by the Brain Korea 21 (BK21) Project from Ministry of Education. It was completed at Radio & Broadcasting Laboratory, ETRI. It is a pleasure to acknowledge their kind hospitality.

Appendix A

A.1. Economic meaning of $H(\mathbf{x})$

In this section, we present an economic meaning of our penalty function that is used for the objective function of our optimization model for bandwidth reservation. At first, we introduce an inventory control problem under uncertain demand (ICPUD) [19,23]. Also, we make it sure that an ICPUD is applicable to bandwidth reservation model in order to optimize the amount of bandwidth to reserve.

Consider a service provider and the customer set. The demand of customers generally has uncertainty or randomness [23]. The service provider should optimize the amount of inventory, i.e., he should determine the economic quantity of order. If he orders too many items, there will be too much inventory and high inventory cost. In contrast, if he orders too few items, there will be loss of sale due to lack of inventory. Under uncertainty or randomness of customer demand, a PDF, if any, would be very helpful to the service provider.

In optimizing the amount of bandwidth to reserve, if we reserve too much bandwidth for a class handover, then there will be much loss of admission opportunity for lower priority calls and might be much idle bandwidth although reserved. In contrast, if we reserve too little bandwidth, then there will be much loss of admission opportunity of the handover although there will be many lower priority calls to be admitted. There are two different kinds of costs: inventory cost (in case of reserving bandwidth too much); opportunity loss cost (in case of reserving bandwidth too little). The inventory cost is a factor of penalty to the service provider, and the opportunity loss cost is a factor of penalty to the service provider and the customers (because the customer who arrives at the store to buy an item at the time of no inventory probably dislikes it and the service provider might lose the customer). With these concepts as in Refs. [19,23,24], we have formulated the penalty functions shown in Eqs. (2) and (3) in order to optimize the resource allocation under uncertainty of demand.

References

[1] D. Hong, S.S. Rappaport, Traffic model and performance analysis of cellular radio telephone systems with prioritized and nonprioritized hand-off procedures, IEEE Trans. Veh. Technol. VT-35 (1986) 77–92.
[2] S.-H. Oh, D.W. Tcha, Prioritized channel assignment in a cellular radio network, IEEE Trans. Commun. 40 (7) (1992) 1259–1269.
[3] B. Eklundh, Channel utilization and blocking probability in a cellular mobile telephone system with directed retry, IEEE Trans. Commun. COM-34 (4) (1986) 329–337.
[4] T.-S.P. Yum, K.L. Yeung, Blocking and handoff performance analysis of directed retry in cellular mobile systems, IEEE Trans. Veh. Technol. 44 (3) (1995) 645–650.
[5] X. Lagrange, B. Jabbari, Fairness in wireless microcellular networks, IEEE Trans. Veh. Technol. 47 (2) (1998) 472–479.
[6] T.-P. Chu, S.S. Rappaport, Overlapping coverage with reuse partitioning in cellular communication systems, IEEE Trans. Veh. Technol. 46 (1) (1997) 41–54.
[7] H. Takanashi, S.S. Rappaport, Dynamic base station selection for personal communication systems with distributed control schemes, in: Proc. 47th VTC, vol. 3, 1997, pp. 1787–1791.
[8] M. Norfal, M. Au, R. Steele, Base station selection algorithms in microcellular mobile radio networks, in: Proceedings of the 15th National Radio Science Conference, Cairo, Egypt, 1998, pp. c11.1–c11.8.
[9] S. Tekinay, B. Jabbari, A measurement-based prioritization scheme for handover in mobile cellular networks, IEEE J. Select. Areas Commun. 10 (1992) 1343–1350.
[10] A. Sampath, J.M. Holtzman, Access control of data in integrated voice/data CDMA systems: benefits and tradeoffs, IEEE J. Select. Areas Commun. 15 (8) (1997) 1511–1526.
[11] C. Chao, W. Chen, Connection admission control for mobile multiple-class personal communications networks, IEEE J. Select. Areas Commun. 15 (8) (1997) 1618–1626.
[12] D.A. Levine, I.F. Akyildiz, M. Naghshineh, A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept, IEEE/ACM Trans. Networking 5 (1) (1997) 1–12.
[13] M. Naghshineh, M. Schwartz, Distributed call admission control in mobile/wireless networks, IEEE J. Select. Areas Commun. 14 (1996) 711–717.

[14] C. Oliveria, J.B. Kim, T. Suda, An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks, IEEE J. Select. Areas Commun. 16 (6) (1998) 858–874.
[15] M. Andersin, Z. Rosberg, J. Zander, Soft and safe admission control in cellular networks, IEEE/ACM Trans. Networking 5 (2) (1997) 255–265.
[16] P. Ramanathan, K.M. Sivalingam, P. Agrawal, S. Kishore, Dynamic resource allocation schemes during handoff for mobile multimedia wireless networks, IEEE J. Select. Areas Commun. 17 (7) (1999) 1270–1283.
[17] S. Ross, Stochastic Processes, second ed., Wiley, New York, 1996.
[18] D. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA, 1995.
[19] F.S. Hillier, G.J. Lieberman, Introduction to Operations Research, fifth ed., McGraw-Hill, New York, 1990.
[20] G.C. Chow, Econometrics, McGraw-Hill, Singapore, 1988.
[21] R. Ramjee, R. Nagarajan, D. Towsley, On optimal call admission control in cellular networks, in: Proc. INFOCOM'96, 1996, pp. 43–50.
[22] Denardo E.V. Denardo, Dynamic Programming: Models and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1982.
[23] J. Qui, R. Loulou, Multiproduct production/inventory control under random demands, IEEE Trans. Auto. Control 40 (2) (1995) 350–356.
[24] K.D. Lee, Stochastic optimal resource management for prioritized admission in broadband wireless communications, Ph.D. Thesis, Korea Advanced Institute of Science and Technology (KAIST), 2001.

Ki-Dong Lee received the BS and MS degrees in Operations Research (Management Science) and the Ph.D. degree in Industrial Engineering (with applications to wireless communications) in 1995, 1997, and 2001, respectively, all from KAIST. As a senior member of technical staff, he joined Radio & Broadcasting Laboratory, Electronics & Telecommunications Research Institute (ETRI) starting upon his graduation from KAIST. His research interests are in the area spanning queueing and optimization theories with applications to MAC and radio resource management for broadband wireless/cellular/satellite communication networks.

Sehun Kim received the BS degree in Physics from Seoul National University in 1972, and the MS and Ph.D. degrees in Operations Research from Stanford University in 1978 and 1981, respectively. He was a visiting Associate Professor in Arizona State University in 1986. He has been with Korea Advanced Institute of Science & Technology (KAIST) as a faculty member since 1982. He served for Korean Operations Research and Management Science Society as the Editor-in-Chief and published a number of papers in Mathematical Programming, Operations Research Letters, IEEE Transactions on Vehicular Technologies, and International Journal of Satellite Communications. His research interests are in the area of mathematical programming, telecommunication systems optimization, and system security.