Optimal Resource Allocation and Adaptive Call Admission Control for Voice/Data Integrated Cellular Networks

Chi Wa Leong, Weihua Zhuang, *Senior Member, IEEE*, Yu Cheng, *Member, IEEE*, and Lei Wang, *Student Member, IEEE*

***Abstract*—Resource allocation and call admission control (CAC) are key management functions in future cellular networks, in order to provide multimedia applications to mobiles users with quality of service (QoS) guarantees and efficient resource utilization. In this paper, we propose and analyze a priority based resource sharing scheme for voice/data integrated cellular networks. The unique features of the proposed scheme are that 1) the maximum resource utilization can be achieved, since all the leftover capacity after serving the high priority voice traffic can be utilized by the data traffic; 2) a Markovian model for the proposed scheme is established, which takes account of the complex interaction of voice and data traffic sharing the total resources; 3) optimal CAC parameters for both voice and data calls are determined, from the perspective of minimizing resource requirement and maximizing new call admission rate, respectively; 4) load adaption and bandwidth allocation adjustment policies are proposed for adaptive CAC to cope with traffic load variations in a wireless mobile environment. Numerical results demonstrate that the proposed CAC scheme is able to simultaneously provide satisfactory QoS to both voice and data users and maintain a relatively high resource utilization in a dynamic traffic load environment. The recent measurement-based modeling shows that the Internet data file size follows a lognormal distribution, instead of the exponential distribution used in our analysis. We use computer simulations to demonstrate that the impact of the lognormal distribution can be compensated for by conservatively applying the Markovian analysis results.**

***Index Terms*—Call admission control (CAC), quality-of-service (QoS), resource allocation, voice and data services.**

I. Introduction

MOBILE CELLULAR networks are evolving into versatile Internet Protocol (IP)-based wireless networks that can provide multimedia services to mobile users. Successful support of multimedia applications requires quality-of-service (QoS) guarantees over wireless links [1], [2]. Due to limited radio resources, efficient resource allocation is extremely important for wireless networks in order to meet the rapidly increasing demands of mobile subscribers as well as to guarantee QoS [3]–[6]. One of the challenges in a multiservice system is that the radio resources should be properly distributed among multiple traffic classes so that the QoS requirements of each traffic class can be satisfied while the resources are utilized as efficiently as possible. In cellular systems, handoff calls resulting from user mobility make resource sharing even more complex. In this paper, we propose a call-level optimal resource allocation scheme and adaptive call admission control (CAC) policies for packet-switched voice/data integrated cellular networks, which can maximize resource utilization under QoS constraints of new call blocking probability (NBP) and handoff call dropping probability (HDP). Using the effective bandwidth technique [7], [8], the resources required for a packet switched network to service each call with a physical layer and link layer QoS guarantee can be determined. The total resources allocated to a service class can be equivalently represented in terms of the number of traffic flows that can be accepted. As a result, the call-level CAC concept originating from circuit-switched networks can be applied to packet-switched networks.

The existing resource sharing schemes can be broadly classified into three categories: complete sharing (CS) [9]–[12], complete partitioning (CP) [9], [11], [13], and virtual partitioning (VP) [14]–[16]. In complete sharing, a new user is always offered access to the network provided that there is sufficient bandwidth at the time of request, and all traffic classes share the total resources indiscriminately. Complete sharing achieves the highest resource utilization among the three categories, but the individual QoS requirement of a certain traffic class cannot be guaranteed in this scheme. On the other hand, complete partitioning can guarantee the resource commitment, and therefore the QoS, to each traffic class, but may underutilize the resources. If some underloaded traffic classes are not utilizing their allocations, the free capacity will be wasted as it can not be used by other heavily loaded traffic classes. Resource utilization of the complete partitioning schemes can be improved by a movable boundary allocation scheme [13], in which the resource allocation to a traffic class can be dynamically adjusted according to the traffic load variations. Virtual partitioning is a more flexible scheme than the movable boundary scheme to achieve a good tradeoff between QoS guarantee and resource utilization. While most of the free capacity from underloaded traffic classes can be utilized by the

Manuscript received November 3, 2004; revised June 17, 2005. This work was supported by the Ontario Government and the Natural Sciences and Engineering Research Council (NSERC) of Canada. The review of this paper was coordinated by Prof. T. Hou.

C. W. Leong is with with the Information Technology Department, Banco Comercial de Macau, Macau.

W. Zhuang and L. Wang are with the Centre for Wireless Communications (CWC), Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: wzhuang@bbcr.uwaterloo.ca; l29wang@bbcr.uwaterloo.ca).

Y. Cheng is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: y.cheng@utoronto.ca).

Digital Object Identifier 10.1109/TVT.2005.858186

overloaded traffic classes, trunk reservation [17], [18] is implemented to protect underloaded traffic classes from resource starvation. In the VP scheme, the QoS of the underloaded classes can not be guaranteed either, although the QoS violation is considerably alleviated compared to the CS scheme [19].

In this paper, we propose and analyze a priority based resource sharing scheme for integrated voice and data, different from the previously mentioned resource sharing schemes where resource preemption is not allowed. Assume a cell has total capacity C. In the proposed scheme, voice users have a preemptive priority in occupying the cell capacity up to $\Gamma(<C)$; all the residual capacity is then picked up by the available-rate (best-effort) data services for maximum resource utilization. Currently, most of the Internet data applications use the transmission control protocol (TCP), which has the characteristic of elastic rate adjustment according to available capacity. Therefore, the proposed resource sharing model is more proper for voice/data integration than the VP scheme, which is better suited to fixed rate resource allocation. In the priority-based system, the service time statistics of data users are related to both voice and data traffic loads in the system. One contribution of this paper is that a mathematical model is developed to characterize the interaction of voice and data calls, where the service time of a data call is represented as a function of the arrival and service time statistics of voice calls and the number of other admitted data calls. Although the similar approach to model the data call service time has been used in [20], we independently proposed such modeling in [21] and go considerably further in this paper to study the resource sharing system from the capacity planning perspective. Based on the mathematical model, we solve the optimal CAC parameters for both voice and data calls according to their QoS requirements. Voice CAC can satisfy the NBP and HDP requirements of voice users with minimal resource allocation, and therefore maximize the leftover capacity for data traffic; data CAC can maximize the new call admission rate while satisfying both the HDP requirement and the minimum throughput of the accepted data calls.

In cellular networks, handoff calls are normally given higher priority to access resources than the new calls in CAC, based on the fact that maintaining an ongoing call is more important than admitting a new arrival. The fractional guard channel policy [22], [23] is proven to be an optimal scheme to differentiate the resource allocation for handoff and new calls, which can maximize the resource utilization while guaranteeing the NBP and HDP requirements. The guard channel concept has also been extended to a multiclass environment [9], [24]. However, in all the previously mentioned schemes, the guard channels are static in that they do not adapt to changes in traffic load and mobility parameters, which may lead to inefficient resource utilization in practice. Most recently, the dynamic guard channel reservation has been extensively investigated in various adaptive resource allocation schemes [25]–[28] to improve resource utilization. In this paper, the guard channel approach is also exploited in the proposed priority resource sharing scheme to protect handoff calls. Particularly, we focus on the limited fractional guard channel policy (LFGCP) proposed in [22]. The policy is the optimal control policy to minimize NBP with a hard constraint on the HDP or to satisfy specified NBP and HDP requirements with minimal resource allocation. It is applied to both voice and data traffic. Designing optimal policy parameters for data traffic is challenging, as data serving capacity statistically depends on the voice service. Another contribution of this paper is that we propose an algorithm to solve the optimal CAC parameters for data LFGCP. Furthermore, the CAC parameters are dynamically adjusted with the traffic load variation.

Although resource allocation and CAC are investigated in this paper for a two-class voice/data integrated system, the proposed techniques can be readily extended to a multiclass multimedia system by grouping the multiple classes to a real-time (including voice)/non-realtime (including data) integrated system [16], as discussed in Section III. In wireless networks, rate adaption [29]–[31] is usually used to improve the sustain probability (defined as the probability that an accepted new call can successfully complete the communication) of real-time multimedia services. In the case of resource scarcity, the bandwidth allocated to a call can be decreased to some degree to increase the call admission rate as long as the QoS degradation is tolerable. The rate adaption is supported by the adaptive coding technique. A good example is the layered video coding format [32], [33], where a base layer contains the most important features of the video and some enhancement layers contain data refining the reconstructed video quality. The rate adaption can be achieved by discarding one or more enhancement layers when necessary. In this paper, the rate adaption of voice calls is also included in the proposed resource sharing scheme, which enhances the call level performance of both voice and data services as demonstrated by analysis and simulation results.

The remainder of this paper is organized as follows. Section II describes the system model under consideration. Section III proposes the LFGCP CAC policy with priority and analyzes its performance by Markov modeling. The procedure to determine optimal CAC parameters and the approaches for adaptive CAC are then presented in Section IV. Numerical results are given in Section V to illustrate the performance of the proposed resource allocation and CAC techniques. Section VI presents computer simulation results, where the performance of the proposed CAC scheme for the lognormally distributed data call length is compared with that for the exponentially distributed length. In Section VII, we provide concluding remarks.

II. System Model

A wireless cellular system supporting both voice and data services is considered. The system capacity (or bandwidth) allocation for both voice and data is normalized with respect to some basic unit. Each such basic unit is referred to as one channel, and nonintegral channel allocation is possible. For simplicity, we study a homogeneous system in statistical equilibrium, where any cell is statistically the same as any other cell, and the mean handoff arrival rate to a cell is equal to the mean handoff departure rate from the cell. Hence, we can decouple a cell from the rest of the system and evaluate the system performance by analyzing the performance of the cell. Such single cell analysis has been extensively used in modern CAC analysis [16], [20], [34], and [35]. In the following, the focus will be placed on CAC for a single test cell. The term “call” at air interface means not

TABLE I
SUMMARY OF THE NOTATIONS

C	the total capacity of a cell
Γ	the maximum capacity that can be used by voice calls
γ_{v}	the fixed bandwidth of a voice call
γ_{d}	the bandwidth used by a data call
c_{d}	the data call bandwidth threshold for minimum service quality
λ_{v}	the new call arrival rate of voice traffic
λ_{d}	the new call arrival rate of data traffic
X_{v}	the cell residence time of a voice call
Y_{v}	the lifetime of a voice call
$(\mu_{\mathrm{v}}^{X})^{-1}$	the mean of X_{v}
$(\mu_{\mathrm{v}}^{Y})^{-1}$	the mean of Y_{v}
X_{d}	the cell residence time of a data call
$(\mu_{\mathrm{d}}^{X})^{-1}$	the mean of X_{d}
L_{d}	the total length of a data call in packets
$(\mu_{\mathrm{d}}^{L})^{-1}$ or l_{d}	the mean of L_{d}
M_{v}	the total call-level channel capacity for voice call
T_{v}	the threshold to determine guard channel for voice call
β_{v}	the probability of accepting a new voice call when voice call channel occupancy is T_{v}
M_{d}	the total call-level channel capacity for data call
T_{d}	the threshold to determine guard channel for data call
β_{d}	the probability of accepting a new data call when data call channel occupancy is T_{d}
$g_{T_{\mathrm{v}},M_{\mathrm{v}}}^{\beta_{\mathrm{v}}}$	the limited fractional guard channel policy for voice calls
$g_{T_{\mathrm{d}},M_{\mathrm{d}}}^{\beta_{\mathrm{d}}}$	the limited fractional guard channel policy for data calls
B_{nv}	the new call blocking probability for voice calls
D_{hv}	the handoff call dropping probability for voice calls
B_{nd}	the new call blocking probability for data calls
D_{hd}	the handoff call dropping probability for data calls
Π_{od}	the overload probability for data calls
Q_{nv}	the QoS requirement on B_{nv}
Q_{hv}	the QoS requirement on D_{hv}
Q_{nd}	the QoS requirement on B_{nd}
Q_{hd}	the QoS requirement on D_{hd}
Q_{od}	the QoS requirement on Π_{od}
Θ_{v}	the total throughput at which all the admitted voice calls are being served
Θ_{d}	the total throughput at which all the admitted data calls are being served
Φ	the overall bandwidth utilization efficiency
Λ_{v}	the total voice call (including new call and handoff call) arrival rate to a cell
Λ_{d}	the total data call (including new call and handoff call) arrival rate to a cell
U_{v}	the voice call service rate (due to call completion or handoff)
U_{d}	the data call service rate (due to call completion or handoff)
h_{v}	handoff rate for voice calls
h_{d}	handoff rate for data calls
$p_{\mathrm{vd}}(i,j)$	the steady state probability that there are i voice calls and j data calls
$p_{\mathrm{v}}(i)$	the steady state probability of the 1-D model for voice
ρ_{v}	the total voice traffic load (in Erlang) in a cell
α_{v}	the fraction of the voice traffic load composed of handoff traffic
$\lambda_{\mathrm{v}i}$	the initial value of λ_{v}
$l_{\mathrm{d}i}$	the initial value of l_{d}
$\gamma_{\mathrm{v}i}$	the standard transmission rate assigned to a voice call
η_{v}	the ratio of the rate adjustment of a voice call to the increment of λ_{v}
η_{d}	the ratio of the rate adjustment of a voice call to the increment of l_{d}
$u(\cdot)$	the unit step function
α	the traffic compensation factor for lognormal data traffic model

only a voice call but also a connection for data service. For convenience, mathematical symbols used in this paper are summarized in Table I.

In the cell, each voice call consists of a constant-rate packet stream. The information carried in a voice call is real-time and can be characterized by low tolerance in transmission delay and medium tolerance in packet loss. When there is a small reduction in the resources available to a voice call, it can tolerate a small degree of packet dropping due to reduced bandwidth allocation. On the other hand, the information carried by a data call is nonreal-time and can be characterized by medium tolerance in transmission delay but low tolerance in packet loss. When there is a reduction in available resources to a data call, the transmission rate can be reduced without dropping information packets, resulting in a longer service (connection) time. Let C denote the capacity (i.e., total available radio bandwidth) of the cell. Due to their real-time nature, voice calls are given preemptive priority over data calls in obtaining resources up to a certain amount, denoted by $\Gamma(<C)$. Each admitted voice user is allocated with the required fixed amount of bandwidth, denoted by γ_{v}. Admitted data calls, at any time, equally share only the leftover bandwidth by voice calls. This method of allocating resources to different types of traffic was first proposed in [10]. The scheme guarantees that a certain amount of bandwidth (i.e., $C-\Gamma$) is always available to data calls by not allowing the high priority (voice) traffic to occupy the entire available bandwidth. The value of Γ should be large enough to satisfy the QoS requirements for voice calls but small enough to give data calls as many leftover

resources as possible. When there are i voice and j data users in the cell, the total leftover capacity after servicing the voice users is $C - i\gamma_{\mathrm{v}}$; the share of this amount allocated to each data user is therefore

$$\gamma_{\mathrm{d}} = \frac{C - i\gamma_{\mathrm{v}}}{j}. \tag{1}$$

The value of γ_{d} changes from time to time, depending on the instantaneous numbers of voice and data users in the cell. Consequently, there are chances that γ_{d} may drop below a critical threshold, denoted by c_{d}. The threshold is the minimum or bottleneck bandwidth required to maintain each data link efficiently with the minimum service quality. If $\gamma_{\mathrm{d}} < c_{\mathrm{d}}$, the service quality of the admitted data users is not satisfactory. This phenomenon is called overload, and the probability of its occurrence should be kept low by restricting the number of data users admitted to the cell.

There are four types of call arrivals in a cell: new voice and data calls originating within the cell, and handoff voice and data calls coming from adjacent cells. The voice and data call arrival processes are assumed to be independent of each other. After the admission process, these arrivals are either blocked (for new calls), dropped (for handoff calls), or admitted. There is no waiting room in the cell and all blocked and dropped calls are cleared.

For tractability in mathematical analysis, we use Markovian processes to model the voice and data call behaviors. It is widely agreed that the Poisson arrival process and exponentially distributed service time can be used to describe the voice calls [36], [37], and the Poisson model can also be used to describe the user-generated data call (session) arrival process [38]. Therefore, new call arrivals are assumed to be Poisson for both voice and data traffic. Let $\lambda_{\mathrm{v}}(\lambda_{\mathrm{d}})$ be the average new call arrival rate of voice (data) traffic. For voice calls, the cell residence time before handoff, denoted by X_{v}, and the total length of a call in time, denoted by Y_{v}, are assumed to be exponentially distributed with means $(\mu_{\mathrm{v}}^{X})^{-1}$ and $(\mu_{\mathrm{v}}^{Y})^{-1}$, respectively. Because the bandwidth allocated to a voice call is constant, there is a one-to-one relationship between the length of a voice call in packets and that in time. Consequently, given γ_{v}, it is necessary to specify only one of the length distributions. The channel holding time of a voice call in the cell is $\min(X_{\mathrm{v}}, Y_{\mathrm{v}})$, which also has an exponential distribution with mean $(\mu_{\mathrm{v}}^{X} + \mu_{\mathrm{v}}^{Y})^{-1}$. The exponential channel holding time and Poisson new call arrival process lead to a Poisson handoff arrival process of voice calls, where the handoff rate is denoted as h_{v}.

For data calls, the cell residence time before handoff, denoted by X_{d}, can also be assumed to be exponentially distributed with mean $(\mu_{\mathrm{d}}^{X})^{-1}$. However, the total length of a data call in packets (the data file size), denoted by L_{d}, is found to follow a lognormal distribution, according to recent measurement-based modeling [39]. For tractability, we assume that L_{d} is exponentially distributed with mean $(\mu_{\mathrm{d}}^{L})^{-1}$ in the mathematical analysis and then use computer simulations to examine the deviation of the true values with the lognormal distribution from the corresponding Markovian mathematical results with the exponential distribution.[1] The relationship between the length of a data call in packets and that in time depends on the numbers of both voice and data calls in the cell. Given that there are i voice and j data users in the system, the bandwidth allocated to each data user is given in (1). From the memoryless property of the exponential distribution, no matter how much data has been transferred, the leftover data packets for this call is still exponentially distributed with mean $(\mu_{\mathrm{d}}^{L})^{-1}$. Therefore, at the state (i, j), the data call has a state-dependent exponential service rate $((C - i\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^{L})/j$ and, hence, the state-dependent exponential channel holding time with mean $[\mu_{\mathrm{d}}^{X} + (C - i\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^{L}/j]^{-1}$. The state-dependent exponential property does not mean the total channel holding time averaged over all the states is also exponential. To facilitate further analysis, we assume that the data call channel holding time is exponentially distributed, which then leads to a Poisson handoff arrival process of data calls, with the handoff rate denoted as h_{d}. Note that the exponential distribution assumption has been widely used in literature [9]–[14], [16], [20], [24]–[26], [29], [34]to provide approximate solutions for cellular systems.

With the system model, the objective of this research is to develop an adaptive CAC policy and find the optimal CAC parameters, so that the QoS requirements of mobile users can be satisfied with high resource utilization in a dynamic traffic load environment.

III. LFGCP With Priority

The proposed CAC policy for the system consists of LFGCPs: one for handling the admission of voice calls and the other for data calls. The LFGCP was originally proposed in [22] for a test cell in a system supporting only constant-rate voice services. The LFGCP can be denoted by $g_{T,M}^{\beta}$, where M is total call-level channel capacity, $T(<M)$ is the channel occupancy threshold over which no new calls are accepted, and β is a constant denoting the probability of accepting a new call when the channel occupancy is T calls. By introducing β, the actual threshold for the new calls (corresponding to T) is equivalent to a continuous variable, leading to an optimized policy. The LFGCP discriminates against new calls since they are not accepted after the channel occupancy reaches $T + 1$. This is to ensure that the remaining channels, called guard channels, are reserved for potential handoff users who have a higher priority in obtaining resources. It has been demonstrated that the LFGCP has the capability of guaranteeing service quality to voice users in the cellular environment [22]. Here, the concept of LFGCP is extended to the integrated voide/data system, and the interaction between the high-priority voice and the available-rate data is taken into account by the mathematical model for accurate performance analysis.[2] Let $g_{T_{\mathrm{v}},M_{\mathrm{v}}}^{\beta_{\mathrm{v}}}$ and $g_{T_{\mathrm{d}},M_{\mathrm{d}}}^{\beta_{\mathrm{d}}}$ denote the LFGCPs for voice

[1] The simulation results presented in Section VI demonstrate that all the QoS metrics obtained with a lognormally distributed data file size degrade slightly from those obtained by theoretical analysis with an exponentially distributed data file size.

[2] Compared with our previous work [40], the CAC for the integrated voice/data services presented in this paper is based on the 2-D Markovian model for both performance analysis and CAC parameter determination, which captures the complex correlation between the two services in a more realistic way.

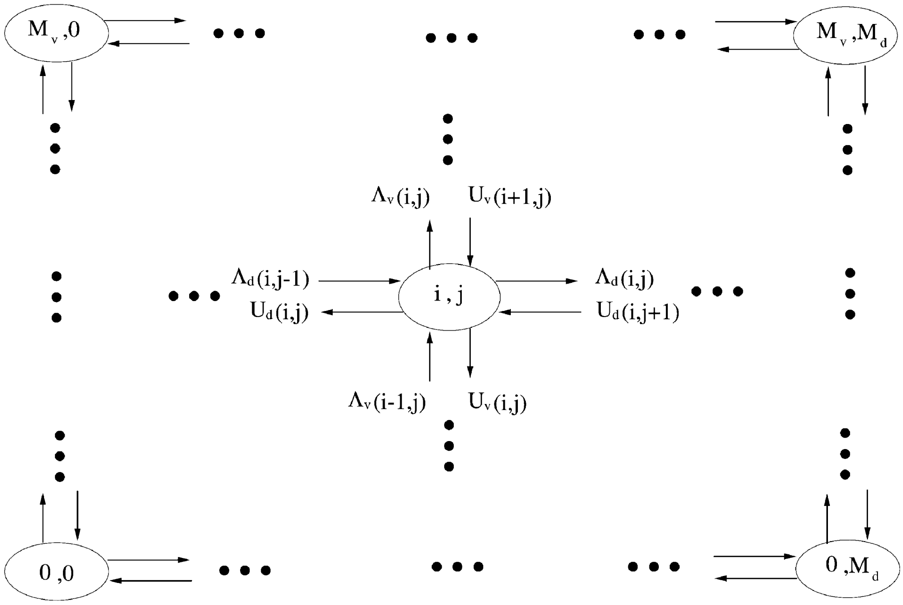


Fig. 1. State diagram of the 2-D traffic model.

and data calls, respectively. The overall policy can be described as follows: where i and j are the current numbers of admitted voice and data calls, respectively, in the cell:

> For voice calls—A new call is always accepted if $i < T_{\mathrm{v}}$, is accepted with probability β_{v} if $i = T_{\mathrm{v}}$, and is always rejected if $i > T_{\mathrm{v}}$; a handoff call is always accepted if $i < M_{\mathrm{v}}$, and rejected otherwise.
>
> For data calls—A new call is always accepted if $j < T_{\mathrm{d}}$, is accepted with probability β_{d} if $j = T_{\mathrm{d}}$, and is always rejected if $j > T_{\mathrm{d}}$; a handoff call is always accepted if $j < M_{\mathrm{d}}$, and is rejected otherwise.

The overall policy is represented by the six parameters $\{M_{\mathrm{v}}, T_{\mathrm{v}}, \beta_{\mathrm{v}}, M_{\mathrm{d}}, T_{\mathrm{d}}, \beta_{\mathrm{d}}\}$. Note that $M_{\mathrm{v}}, T_{\mathrm{v}}, M_{\mathrm{d}}$, and T_{d} represent the numbers of calls and not the numbers of channels. These parameters should be determined in such a way that given the traffic load, the system under the CAC policy can guarantee QoS to the users and achieve a high resource utilization efficiency. Here, we consider the following connection-level QoS measures:

> $B_{\mathrm{nv}}(D_{\mathrm{hv}})$—the new call blocking (handoff call dropping) probability for voice calls;
>
> $B_{\mathrm{nd}}, (D_{\mathrm{hd}})$—the new call blocking (handoff call dropping) probability for data calls;
>
> Π_{od}—the overload probability for data calls.

The QoS requirements are specified by the upper bounds of $\{B_{\mathrm{nv}}, D_{\mathrm{hv}}, B_{\mathrm{nd}}, D_{\mathrm{hd}}, \Pi_{\mathrm{od}}\}$, denoted by $\{Q_{\mathrm{nv}}, Q_{\mathrm{hv}}, Q_{\mathrm{nd}}, Q_{\mathrm{hd}}, Q_{\mathrm{od}}\}$. From the users' point of view, it is better to be blocked in the beginning rather than dropped in the middle of a connection. As a result, in addition to the resource reservation for handoff calls, the upper bounds for handoff calls (Q_{hv} and Q_{hd}) are given a lower (more restrictive) value than the corresponding upper bounds for the new calls (Q_{nv} and Q_{nd}). As QoS provisioning and high resource utilization are conflicting goals, to evaluate the performance of the CAC policy, the following measures related to resource utilization are used:

> $\Theta_{\mathrm{v}}(\Theta_{\mathrm{d}})$—the total throughput (i.e., average rate) at which all the admitted voice (data) calls are being served;
>
> Φ—the overall bandwidth utilization efficiency defined as the average percentage of the entire spectrum of the cell that is used by the admitted voice and data calls.

Since both voice and data calls share the total resources of the cell, the two LFGCPs are not independent. Their correlation is captured in the modeling and performance analysis discussed in the following and in the determination of the policy parameters to be discussed in Section IV. With Poisson arrivals, exponential service time for voice calls, and exponential length in packets for data calls, the traffic flows under the control of the proposed CAC policy can be modeled by a two-dimensional (2-D) continuous-time birth-death process with state (i, j), where i and j are the current numbers of admitted voice and data calls, respectively. The state diagram is shown in Fig. 1, where $\Lambda_{\mathrm{v}}(\Lambda_{\mathrm{d}})$ is the voice (data) call arrival rate, and $U_{\mathrm{v}}(U_{\mathrm{d}})$ is the voice (data) call service rate associated with each state. For state (i, j), the transition rates are shown in (2)–(5) at the bottom of the next page. Because voice calls have a higher priority in obtaining resources, $\Lambda_{\mathrm{v}}(i, j)$ and $U_{\mathrm{v}}(i, j)$ do not depend on j, the number of data calls. On the other hand, since data calls can occupy only the leftover bandwidth by voice calls, the determination of the service rate $U_{\mathrm{d}}(i, j)$ requires the knowledge of both i and j. The steady-state probability that there are i voice calls and j data calls in the cell $p_{\mathrm{vd}}(i, j)$ for $0 \leq i \leq M_{\mathrm{v}}$ and $0 \leq j \leq M_{\mathrm{d}}$ can be obtained numerically. In particular, the local balance equation for each state of the 2-D model and the normalization constraint $\sum_{0 \leq i \leq M_{\mathrm{v}}, 0 \leq j \leq M_{\mathrm{d}}} p_{\mathrm{vd}}(i, j) = 1$ constitute a set of $(M_{\mathrm{v}}+1)(M_{\mathrm{d}}+1)$ linearly independent equations [41]. These equations can be used to solve for the steady state probabilities, from which the performance of the policy can be evaluated.

Because $\Lambda_{\mathrm{v}}(i, j)$ and $U_{\mathrm{v}}(i, j)$ are independent of j, the 2-D model can be condensed to a one-dimensional (1-D) model in describing the resource occupancy of voice calls. The 1-D model is an $M/M/M_{\mathrm{v}}/M_{\mathrm{v}}$ queue, where each state represents the number of the active voice calls. The steady-state probability $p_{\mathrm{v}}(i)$ of the 1-D model $0 \le i \le M_{\mathrm{v}}$ is related to those of the 2-D model by

$$p_{\mathrm{v}}(i) = \begin{cases} \sum_{j=0}^{M_{\mathrm{d}}} p_{\mathrm{vd}}(i,j), & 0 \le i \le M_{\mathrm{v}} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

From the 1-D model, $p_{\mathrm{v}}(i)$ can also be explicitly calculated as follows:

$$p_{\mathrm{v}}(i) = \begin{cases} \frac{\rho_{\mathrm{v}}^i}{i!} p_{\mathrm{v}}(0), & \text{if } 0 \le i \le T_{\mathrm{v}} \\ \frac{\rho_{\mathrm{v}}^i[\alpha_{\mathrm{v}} + (1-\alpha_{\mathrm{v}})\beta_{\mathrm{v}}]}{i!} p_{\mathrm{v}}(0), & \text{if } i = T_{\mathrm{v}} + 1 \\ \frac{\rho_{\mathrm{v}}^i[\alpha_{\mathrm{v}} + (1-\alpha_{\mathrm{v}})\beta_{\mathrm{v}}]\alpha_{\mathrm{v}}^{i-T_{\mathrm{v}}-1}}{i!} p_{\mathrm{v}}(0), & \text{if } T_{\mathrm{v}} + 2 \le i \le M_{\mathrm{v}} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\rho_{\mathrm{v}} = (\lambda_{\mathrm{v}} + h_{\mathrm{v}})/\mu_{\mathrm{v}}^Z$ is the total voice traffic load in the cell, $\alpha_{\mathrm{v}} = h_{\mathrm{v}}/\lambda_{\mathrm{v}} + h_{\mathrm{v}})$ is the fraction of the voice traffic load composed of handoff traffic, and $p_{\mathrm{v}}(0)$ is a normalization constant given by

$$p_{\mathrm{v}}(0) = \left\{ \sum_{i=0}^{T_{\mathrm{v}}} \frac{\rho_{\mathrm{v}}^i}{i!} + \frac{\rho_{\mathrm{v}}^{T_{\mathrm{v}}+1}[\alpha_{\mathrm{v}} + (1-\alpha_{\mathrm{v}})\beta_{\mathrm{v}}]}{(T_{\mathrm{v}}+1)!} + \sum_{i=T_{\mathrm{v}}+2}^{M_{\mathrm{v}}} \frac{\rho_{\mathrm{v}}^i[\alpha_{\mathrm{v}} + (1-\alpha_{\mathrm{v}})\beta_{\mathrm{v}}]\alpha_{\mathrm{v}}^{i-T_{\mathrm{v}}-1}}{i!} \right\}^{-1}. \tag{8}$$

With the steady-state probabilities, the performance measures related to voice calls are given by

$$B_{\mathrm{nv}} = (1-\beta_{\mathrm{v}}) p_{\mathrm{v}}(T_{\mathrm{v}}) + \sum_{i=T_{\mathrm{v}}+1}^{M_{\mathrm{v}}} p_{\mathrm{v}}(i) \tag{9}$$

$$D_{\mathrm{hv}} = p_{\mathrm{v}}(M_{\mathrm{v}}) \tag{10}$$

$$\Theta_{\mathrm{v}} = \sum_{i=1}^{M_{\mathrm{v}}} i(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y) p_{\mathrm{v}}(i). \tag{11}$$

Similarly, the performance measures for data calls can be obtained from the 2-D model and are given by

$$B_{\mathrm{nd}} = (1-\beta_{\mathrm{d}}) \sum_{i=0}^{M_{\mathrm{v}}} p_{\mathrm{vd}}(i, T_{\mathrm{d}}) + \sum_{j=T_{\mathrm{d}}+1}^{M_{\mathrm{d}}} \sum_{i=0}^{M_{\mathrm{v}}} p_{\mathrm{vd}}(i,j) \tag{12}$$

$$D_{\mathrm{hd}} = \sum_{i=0}^{M_{\mathrm{v}}} p_{\mathrm{vd}}(i, M_{\mathrm{d}}) \tag{13}$$

$$\Theta_{\mathrm{d}} = \sum_{j=1}^{M_{\mathrm{d}}} \sum_{i=0}^{M_{\mathrm{v}}} j \left[\mu_{\mathrm{d}}^X + \frac{(C - i\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^L}{j} \right] p_{\mathrm{vd}}(i,j). \tag{14}$$

As the probability of a data user finding itself in the cell with k voice and $l-1$ other data users is proportional not only to $p_{\mathrm{vd}}(k,l)$ but also to l, the overload probability for data calls is

$$\Pi_{\mathrm{od}} = \sum_i \sum_j \frac{j p_{\mathrm{vd}}(i,j)}{\sum_{k=0}^{M_{\mathrm{v}}} \sum_{l=0}^{M_{\mathrm{d}}} l p_{\mathrm{vd}}(k,l)} \tag{15}$$

where i and j are chosen from the set $\{(i,j) : 0 \le i \le M_{\mathrm{v}}, 1 \le j \le M_{\mathrm{d}}, (C - i\gamma_{\mathrm{v}})/j < c_{\mathrm{d}}\}$, and the denominator is simply a normalization constant. As to the resource utilization efficiency, there exist unused resources only if there is no data user in the cell. The average amount of unused bandwidth is therefore $\sum_{i=0}^{M_{\mathrm{v}}} p_{\mathrm{vd}}(i,0)(C - i\gamma_{\mathrm{v}})$. Hence, the bandwidth utilization efficiency is given by

$$\Phi = \frac{C - \sum_{i=0}^{M_{\mathrm{v}}} p_{\mathrm{vd}}(i,0)(C - i\gamma_{\mathrm{v}})}{C}. \tag{16}$$

From the preceding analysis, the call-level QoS measures for voice traffic depend only on the CAC parameters $M_{\mathrm{v}}, T_{\mathrm{v}}$ and β_{v}, and are independent of the bandwidth assigned to a voice call (γ_{v}) and the presence of data calls; however, the QoS measures for data calls depend on γ_{v}, the arrival and service statistics of voice calls, due to their lower service priority.

Note that the preceding two-class Markovain modeling can be readily extended to a multiservice system supporting more than two traffic classes, where in user calls in different classes can have different bandwidth requirements. The modeling approach proposed in [16] can be adopted for such an extension. All the classes can be combined into two groups (real-time and nonreal-time), and the LFGCP can be applied to both groups for admission control. With heterogeneous call arrival processes,

$$\Lambda_{\mathrm{v}}(i,j) = \begin{cases} \lambda_{\mathrm{v}} + h_{\mathrm{v}} & \text{if } 0 \le i < T_{\mathrm{v}}; \\ \beta_{\mathrm{v}}\lambda_{\mathrm{v}} + h_{\mathrm{v}} & \text{if } i = T_{\mathrm{v}}; \\ h_{\mathrm{v}} & \text{if } T_{\mathrm{v}} < i < M_{\mathrm{v}}; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$U_{\mathrm{v}}(i,j) = \begin{cases} i(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y) & \text{if } 0 < i \le M_{\mathrm{v}}; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$\Lambda_{\mathrm{d}}(i,j) = \begin{cases} \lambda_{\mathrm{d}} + h_{\mathrm{d}} & \text{if } 0 \le j < T_{\mathrm{d}}; \\ \beta_{\mathrm{d}}\lambda_{\mathrm{d}} + h_{\mathrm{d}} & \text{if } j = T_{\mathrm{d}}; \\ h_{\mathrm{d}} & \text{if } T_{\mathrm{d}} < j < M_{\mathrm{d}}; \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$$U_{\mathrm{d}}(i,j) = \begin{cases} j \left[\mu_{\mathrm{d}}^X + \frac{(C - i\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^L}{j} \right] & \text{if } 0 \le i \le M_{\mathrm{v}} \quad \text{and} \quad 0 < j \le M_{\mathrm{d}}; \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

bandwidth requirements, call holding times, and cell residence times in different classes, the LFGCPs for voice and data can be modeled as a multidimensional Markov chain. The interaction between the two LFGCPs can still be characterized by the approach proposed in this section. Assume there are K_{RT} classes of real-time traffic and K_{NRT} classes of nonreal-time traffic. The K_{RT}-D Markov chain for the real-time LFGCP can be solved independent of the nonreal-time traffic due to its priority in resource access. Performance of the non-real-time traffic needs to be obtained by solving a $(K_{\mathrm{RT}} + K_{\mathrm{NRT}})$-D Markov chain to take the interaction between voice and data into account. In the following, without losing generality, we continue to focus on the voice/data two-class model when studying the optimal and adaptive CAC.

IV. Optimal CAC Parameters and Adaptive Policy

A. *Optimal CAC Parameters*

This section proposes a procedure to determine the optimal CAC policy parameters $\{M_\mathrm{v}, T_\mathrm{v}, \beta_\mathrm{v}, M_\mathrm{d}, T_\mathrm{d}, \text{ and } \beta_\mathrm{d}\}$. For voice calls, the algorithm `Min` proposed in [22] can be used, which is able to find the minimum required value of M_v and the corresponding values of T_v and β_v, such that the specified QoS upper bounds are satisfied. However, the parameters for data calls cannot be determined in a straightforward manner, because of their lower service priority and their nonconstant bandwidth allocation. Given the total resources of the cell, it may be impossible to satisfy all the QoS parameter upper bounds, especially when the traffic load is high. Among the QoS measures, the new call blocking probability for data traffic is the least important, since new data calls have a lower service priority than voice calls and handoff data calls. As a result, the proposed procedure is to determine the CAC parameters which guarantee only four upper bounds $\{Q_{\mathrm{nv}}, Q_{\mathrm{hv}}, Q_{\mathrm{hd}}, \text{ and } Q_{\mathrm{od}}\}$, given that the amount of the total resources C is large enough. Even though the upper bound for the new data call blocking probability Q_{nd} is not guaranteed, the proposed procedure takes necessary steps to minimize B_{nd} for any given environment. Given the total amount of capacity (C), the QoS requirements ($Q_{\mathrm{nv}}, Q_{\mathrm{hv}}, Q_{\mathrm{hd}}$, and Q_{od}), the bandwidth requirements ($\gamma_\mathrm{v}, \mu_\mathrm{d}^L$, and c_d), and the traffic conditions ($\lambda_\mathrm{v}, h_\mathrm{v}, \mu_\mathrm{v}^X, \mu_\mathrm{v}^Y, \lambda_\mathrm{d}, h_\mathrm{d}$, and μ_d^X), the procedure to determine the CAC parameters is summarized as follows:

Step 1) Determine the values of M_v, T_v and β_v by using algorithm `Min`. If there are not enough resources to support the demand from voice calls (i.e., $M_\mathrm{v}\gamma_\mathrm{v} > C$), stop.

Step 2) Set the intermediate value for β_d, denoted by β'_d, to 0.

Step 3) Determine the first intermediate value for M_d, denoted by M'_d. First, assume $g^{\beta'_\mathrm{d}}_{M'_\mathrm{d},M'_\mathrm{d}}$ is used for data calls, which makes no discrimination between new and handoff calls. Starting with $M'_\mathrm{d} = 1$, search sequentially for the largest M'_d such that the overload probability under the control of $g^{\beta_\mathrm{v}}_{T_\mathrm{v},M_\mathrm{v}}$ and $g^{\beta'_\mathrm{d}}_{M'_\mathrm{d},M'_\mathrm{d}}$ is not larger than Q_{od}. If there does not exist such an M'_d value; i.e., there are not enough resources to guarantee the overload probability requirement for data calls, stop.

Step 4) Determine the first intermediate value for T_d, denoted by T'_d. Starting from $T'_\mathrm{d} = M'_\mathrm{d}$, decrease the value of T'_d by 1 continuously until either a) the data handoff call dropping probability under the control of $g^{\beta_\mathrm{v}}_{T_\mathrm{v},M_\mathrm{v}}$ and $g^{\beta'_\mathrm{d}}_{T'_\mathrm{d},M'_\mathrm{d}}$ is not larger than Q_{hd}, or b) $T'_\mathrm{d} = 0$.

Step 5) Determine the second intermediate value for T_d, denoted by T''_d, and for M_d, denoted by M''_d. Starting with $T''_\mathrm{d} = T'_\mathrm{d}$ and $M''_\mathrm{d} = M'_\mathrm{d}$, search for the largest values of T''_d and M''_d by increasing the values of both by 1 simultaneously each time, such that the overload probability under the control of $g^{\beta_\mathrm{v}}_{T_\mathrm{v},M_\mathrm{v}}$ and $g^{\beta'_\mathrm{d}}_{T''_\mathrm{d},M''_\mathrm{d}}$ is not larger than Q_{od}.

Step 6) If the handoff call dropping probability is still violated, let $M'_\mathrm{d} = M''_\mathrm{d}$ and repeat Steps 4 and 5 until the Q_{hd} upper bound is satisfied. If T'_d reaches 0 at the end of Step 4 and M''_d cannot be raised further in Step 5 because of potential violation of the overload probability upper bound, i.e., there are not enough resources to guarantee the handoff call dropping probability for data calls, stop; otherwise, if suitable values for both T''_d and M''_d are obtained, go to Step 7.

Step 7) Starting with $M_\mathrm{d} = M''_\mathrm{d}$, search for the minimum value of M_d such that $M_\mathrm{d} \geq T''_\mathrm{d}$ and the Q_{hd} upper bound is still satisfied.

Step 8) Determine the final value of T_d. Starting with $T_\mathrm{d} = T''_\mathrm{d}$, search for the largest value for T_d such that a) $T_\mathrm{d} \leq M_\mathrm{d}$, and b) both Q_{hd} and Q_{od} upper bounds are satisfied under the control of $g^{\beta_\mathrm{v}}_{T_\mathrm{v},M_\mathrm{v}}$ and $g^{\beta'_\mathrm{d}}_{T_\mathrm{d},M_\mathrm{d}}$.

Step 9) Determine the final value of β_d. Do a bisection search over the interval [0, 1] to find the maximum value for β_d, such that both Q_{hd} and Q_{od} upper bounds are satisfied under the control of $g^{\beta_\mathrm{v}}_{T_\mathrm{v},M_\mathrm{v}}$ and $g^{\beta_\mathrm{d}}_{T_\mathrm{d},M_\mathrm{d}}$.

The minimal amount of resources is determined to satisfy the call blocking and dropping probabilities for voice calls in Step 1 due to their high priority. The minimum value for M_v is optimal in the sense that resource utilization is maximized for voice calls. At the same time, the optimal value for Γ is found to be $M_\mathrm{v}\gamma_\mathrm{v}$, which specifies the minimum resources necessary for voice calls and gives data calls as much leftover capacity as possible. The procedure to determine the CAC parameters for data calls (Steps 2–9) is first to guarantee the overload probability, then to guarantee the handoff call dropping probability, and finally to minimize the new call blocking probability. Since parameter M_d has a larger impact on the data call overload probability than parameter T_d, M'_d is first determined in Step 3 according to the overload probability upper bound. Note that any more restrictive LFGCP $g^{\beta'_\mathrm{d}}_{k,M'_\mathrm{d}}$ with any integer $k < M'_\mathrm{d}$ also satisfies the same overload probability upper bound under identical traffic conditions. Thus, the M'_d value can be used in Step 4 to determine T'_d. Given the M'_d value, one way to control the handoff call dropping probability is to vary the size of the guard bandwidth $(M'_\mathrm{d} - T'_\mathrm{d})$. In particular, the size of the guard bandwidth is a monotonically decreasing function of the call dropping probability D_{hd}. Therefore, in Step 4, the value of T'_d is reduced gradually to increase the size of the guard bandwidth, resulting in the intermediate policy $g^{\beta'_{\rm d}}_{T'_{\rm d},M'_{\rm d}}$. Compared to the policy $g^{\beta'_{\rm d}}_{M'_{\rm d},M'_{\rm d}}$ obtained in Step 3, the policy obtained in Step 4 is more restrictive because fewer new calls can be admitted. As a result, Step 5 is to check if there are any resources which can be used to reduce the new call blocking probability. The intermediate policy from Step 4 is relaxed by increasing both $T'_{\rm d}$ and $M'_{\rm d}$ simultaneously, under the constraint on the overload probability. When both $T'_{\rm d}$ and $M'_{\rm d}$ are increased simultaneously, the size of the guard bandwidth $(M'_{\rm d} - T'_{\rm d})$ remains the same as that determined in Step 4; hence, the relaxation of the intermediate policy has only beneficial but no adverse effect on the handoff call dropping probability. In Step 4, there is a possibility that the search is terminated prematurely (because $T'_{\rm d}$ reaches 0) without satisfying the upper bound $Q_{\rm hd}$. If the relaxation in Step 5 does not provide any benefit in this respect, it is necessary to repeat Steps 4 and 5 until the upper bound is satisfied, as specified in Step 6. Because of the repetitive relaxation of the policy in Step 5, the policy $g^{\beta'_{\rm d}}_{T''_{\rm d},M_{\rm d}}$ from Step 6 may not correspond to the minimal new call blocking probability. Steps 7 to 9 are to find the final values of $M_{\rm d}$, $T_{\rm d}$, and $\beta_{\rm d}$, which minimize the new call blocking probability. Note that in Step 9 a bisection search for the maximum $\beta_{\rm d}$ is possible because both $D_{\rm hd}$ and $\Pi_{\rm od}$ are monotonically increasing functions of $\beta_{\rm d}$. Provided that there is enough capacity in the cell, by using the preceding procedure, all the specified QoS requirements will be satisfied, and the QoS requirement for new data calls, though not guaranteed, will be maintained to the highest degree possible. Otherwise, the procedure stops before Step 9, in which case, additional resources are required and/or an adaptive CAC policy should be used.

Note that the previously described algorithm for determining CAC parameters incurs higher computation complexity compared to the single class `Min` algorithm. In the proposed algorithm, analysis of a 2-D Markov chain is involved in the iterative search for optimal CAC parameters. Also, the computation complexity increases when the number of priority classes increases. However, the efficient computation technique for multidimensional Markov chains has been widely studied, e.g., [14], [16], and the references therein, and can be adopted in the algorithm for CAC parameters. Furthermore, the computation complexity does not affect the online CAC, as the CAC parameters do not need to be calculated frequently. With the static policy, the offline computation can be applied; with the adaptive policies, to be discussed in the following, the CAC parameters are calculated over a timescale much larger than the call interarrival time. For real-time online admission control, only simple addition and comparison operations are required under the configured CAC parameters.

B. Adaptive Policy

The proposed CAC policy is designed for a target traffic condition. If the actual traffic condition deviates from the target condition, the policy may result in dissatisfactory service quality or under-utilized resources. One solution is to make the CAC adaptive to traffic load changes. To ensure the stability of the system and to preserve the validity of the analysis in Section III, the mean time between changes of traffic conditions is assumed to be much longer than the time required for the system to acquire stationary states. Consider changes in a) the voice traffic load $\lambda_{\rm v}$, with $\lambda_{\rm vi}$ being its initial values for which the CAC policy is originally designed for, and b) the mean message length of data calls in packets $l_{\rm d} = E(L_{\rm d}) = (\mu_{\rm d}^L)^{-1}$, with $l_{\rm di}$ being its initial value. Two phases of adaption are taken in the adaptive CAC, which is described as follows:

1) Phase 1—Load Adaption: Traffic conditions characterized by $\lambda_{\rm v}$ and $l_{\rm d}$ are periodically monitored (in a time scale much larger than the call interarrival time). If a significant change (i.e., the traffic load fluctuation exceeds a predefined threshold, for example $5\%\lambda_{\rm v}$ or $5\% l_{\rm d}$) is detected, the procedure described in Section 4-A is re-executed to find the new parameters of the CAC policy. This is to ensure that the policy is still able to maintain the same satisfactory service qualities to users under the new traffic condition. Due to the scarce wireless spectrum and expected large number of subscribers in the future wireless networks, load adaption alone may not be able to deliver satisfactory services simultaneously to both voice and data users. For example, when the traffic load of voice users becomes very large, the value of Γ starts to approach the cell capacity C. The resources left for data users starts to decrease and eventually, either delays suffered by data users become completely intolerable if the same number of data users are supported, or the new call blocking probability increases to a very large value if the number of data users allowed into the system is reduced. The second scenario also leads to lower throughput and reduced resource utilization for the system. As a result, load adaption should be combined with relaxed QoS to mitigate the problem, i.e., the second phase of adaption.
2) Phase 2—Bandwidth Allocation Adjustment: When traffic load increases, it is possible to take a small quantity of resources from the users already in the system and use the aggregate to compensate for the extra demand. This can be done as long as the degradation in service quality caused by the reduced resource allocation to each admitted user is tolerable. In particular, voice calls can tolerate a certain amount of reduction in transmission rate before the service quality drops to an unacceptable level. That is, the amount of resources allocated to each admitted voice user $\gamma_{\rm v}$ can be reduced to a certain extent as the traffic load increases. Let γ_{vi} be the standard transmission rate assigned to a voice call. When the traffic load, $\lambda_{\rm v}$ and/or $l_{\rm d}$, increases with respect to λ_{vi} and/or l_{di}, bandwidth allocation adjustment reallocates a new transmission rate (less than γ_{vi}) to each voice call. As an example, the relationship can be expressed as

$$\gamma_{\rm v} = \gamma_{vi} - \eta_{\rm v} \cdot u(\lambda_{\rm v} - \lambda_{vi}) - \eta_{\rm d} \cdot u(l_{\rm d} - l_{di}) \tag{17}$$

where the weights $\eta_{\rm v}$ and $\eta_{\rm d}$ are both positive constants, and $u(\,\cdot\,)$ is the unit step function. The actual values of the weights depend on the tolerance level of the voice users to service quality degradation. Using (17), the adaptability of the CAC policy to dynamic traffic loads can be described as follows: 1) When λ_{v} surges above λ_{vi}, γ_{v} is reduced. The reduction is used to compensate the increase in voice call arrivals. The purpose is to avoid the potential increase of the value of Γ, thereby avoiding the domination of the entire spectrum by voice traffic and stabilizing the throughput for data calls. 2) When λ_{v} drops below λ_{vi}, the value of γ_{v} remains the same. The value of Γ, however, will be reduced with the updated CAC parameters (for the new traffic conditions). Because of the available-rate characteristic of data calls, the extra leftover capacity will be automatically picked up by data calls, which increases the throughput of data traffic. 3) When l_{d} increases above l_{di}, γ_{v} is decreased so that the extra leftover capacity by voice calls can be used to compensate the increase in the data traffic load, thereby stabilizing the call blocking probability of new data calls. 4) When l_{d} drops below l_{di}, there is no change in the value of γ_{v}. The policy for voice calls will not be changed, and therefore, the leftover capacity remains the same. However, due to the decrease in the data traffic load, the service rate for data users is increased. Although the reduction of bandwidth has no adverse effect on the dropping and blocking probabilities of voice calls, it does degrade the transmission performance. The corresponding QoS measure can be given by $(\gamma_{vi} - \gamma_{\mathrm{v}})/\gamma_{vi}$, which is the normalized reduction in bandwidth allocation that each voice call experiences during heavy traffic conditions.

C. Determination of Handoff Call Arrival Rates

In the preceding sections, when we drive the CAC performance measures and the adaptive CAC parameters, it is assumed that both the voice and data handoff call arrival rates are known as h_{v} and h_{d}, respectively. However, since the overall system is assumed to be homogeneous in statistical equilibrium, the mean handoff arrival rate to a cell should be equal to the mean handoff departure rate toward neighboring cells. That is

$$h_{\mathrm{v}} = \sum_{i=1}^{M_{\mathrm{v}}} i\mu_{\mathrm{v}}^{X} p_{\mathrm{v}}(i) \tag{18}$$

$$h_{\mathrm{d}} = \sum_{j=1}^{M_{\mathrm{d}}} j\mu_{\mathrm{d}}^{X} \sum_{i=1}^{M_{\mathrm{v}}} p_{\mathrm{vd}}(i,j). \tag{19}$$

As seen in (18) and (19), the handoff arrival (departure) rates are dependent on state probabilities, while the state probabilities are derived using the handoff arrival rates. To solve this problem, we use the iterative algorithm presented in [34].

Assume that the total capacity (C), the QoS requirements $(Q_{\mathrm{nv}}, Q_{\mathrm{hv}}, Q_{\mathrm{hd}}, Q_{\mathrm{od}})$, the bandwidth requirements $(\gamma_{\mathrm{vi}}, \mu_{\mathrm{d}}^{L}, c_{\mathrm{d}})$, and the traffic conditions $(\lambda_{\mathrm{v}}, \mu_{\mathrm{v}}^{X}, \mu_{\mathrm{v}}^{Y}, \lambda_{\mathrm{d}}, \mu_{\mathrm{d}}^{X})$ are known. If nonadaptive CAC policy is used, the CAC parameters $(M_{\mathrm{v}}, T_{\mathrm{v}}, \beta_{\mathrm{v}}, M_{\mathrm{d}}, T_{\mathrm{d}}, \beta_{\mathrm{d}})$ for a pretargeted traffic condition can be obtained. The iterative algorithm for our case is implemented as follows:

Step 1) Set initial values for h_{v} and h_{d} as $(\mu_{\mathrm{v}}^{X}/\mu_{\mathrm{v}}^{Y})\lambda_{\mathrm{v}}$ and $(1/2)\lambda_{\mathrm{d}}$, respectively. The former is from the suggestion given in [34]; the latter is a heuristical estimation, as the average data call completion rate is not convenient to use. Both settings of the initial values lead the algorithm to convergence in numerical practice.

Step 2) If nonadaptive CAC policy is used, compute the steady state probabilities $p_{\mathrm{vd}}(i,j)$ according to Section III using the preset CAC parameters. Otherwise, if adaptive CAC policy is used, solve the CAC parameters and the steady state probabilities in a combined way according to the procedures described in Sections IV-A and B.

Step 3) Compute the mean departure rate of handoff voice and data calls, $h_{v,\mathrm{new}}$ and $h_{d,\mathrm{new}}$, according to (18) and (19).

Step 4) Let $\epsilon(>0)$ be a predefined small value. If $|h_{v,\mathrm{new}} - h_{\mathrm{v}}| > \epsilon$, $h_{\mathrm{v}} \leftarrow h_{v,\mathrm{new}}$ and go to Step 2 of the iteration algorithm for voice calls. The same ϵ rule is applied to the iteration algorithm for data calls.

Step 5) Compute the performance measures for voice and data calls using the obtained handoff rates.

It should be mentioned that 1) As the voice calls occupy resources with high priority, solving for handoff rates and CAC parameters of voice calls is executed first, independent of the data calls. With the obtained $h_{\mathrm{v}}, M_{\mathrm{v}}, T_{\mathrm{v}}$, and β_{v}, the iteration algorithm for the data calls is then executed. 2) With nonadaptive CAC, the CAC parameters for the pretargeted traffic condition are first solved according to the above algorithm. The CAC parameters are then fixed in the searching of handoff rates for other traffic load conditions. With adaptive CAC, the handoff rates and CAC parameters are solved jointly in the iteration algorithm. 3) The analysis of voice traffic can be condensed to a 1-D model, and the adaptive CAC policies select parameters to achieve the QoS as (or very close to) Q_{nv} and Q_{hv}. Therefore, the handoff rate of voice calls in a steady-state homogeneous cellular system can be derived and is given by

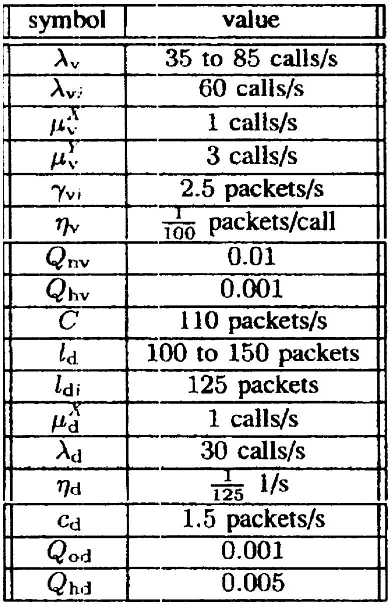
$$h_{\mathrm{v}} = \frac{\lambda_{\mathrm{v}}\mu_{\mathrm{v}}^{X}(1 - Q_{\mathrm{nv}})}{\mu_{\mathrm{v}}^{Y} + \mu_{\mathrm{v}}^{X}Q_{\mathrm{hv}}} \tag{20}$$

by setting the handoff arrival rate to the 1-D queue equal to the handoff departure rate.

V. Numerical Results and Discussion

This section presents numerical results based on the mathematical analysis in Sections III–IV to demonstrate the advantages of the adaptive CAC policy over the nonadaptive policy. Three policies are considered: Policy A is a nonadaptive policy under the engineered traffic condition $\lambda_{\mathrm{v}} = \lambda_{vi}$ and $l_{\mathrm{d}} = l_{di}$; Policy B is an adaptive policy with only load adaption; and Policy C is an adaptive policy with both load adaption and bandwidth allocation adjustment. For each policy, the optimal CAC parameters are solved according to Section IV. The performance of the policies is compared under the conditions specified in Table II. For simplicity, we consider the unit channel capacity as 1 packet/s. For Policy A, the handoff rates, the CAC parameters, and Γ are determined to be: $h_{\mathrm{v}} = 19.7934$ calls/s, $h_{\mathrm{d}} = 16.0645$ calls/s, $M_{\mathrm{v}} = 32$ calls, $T_{\mathrm{v}} = 30$ calls, $\beta_{\mathrm{v}} = 0.0$, $M_{\mathrm{d}} = 26$ calls, $T_{\mathrm{d}} = 10$ calls, $\beta_{\mathrm{d}} = 0.6519$, and $\Gamma = M_{\mathrm{v}}\gamma_{\mathrm{vi}} = 80$ packets/s. The determination of the three voice call parameters

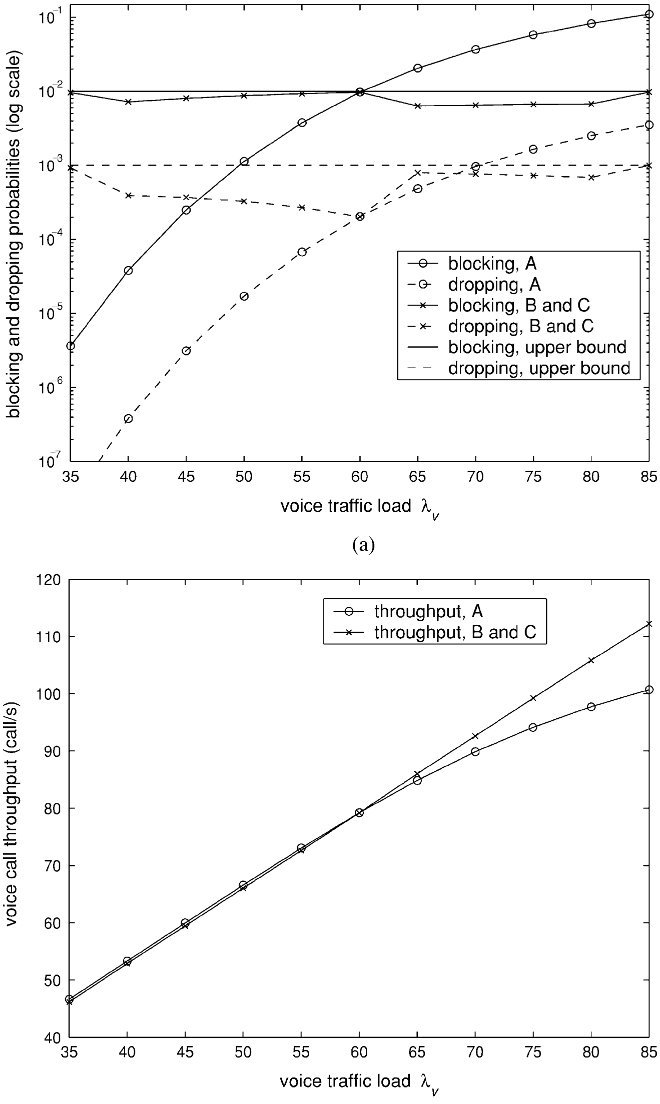
TABLE II
TRAFFIC CONDITIONS AND BANDWIDTH REQUIREMENTS USED IN THE NUMERICAL STUDY



symbol	value
λ_v	35 to 85 calls/s
λ_{vi}	60 calls/s
μ_v^X	1 calls/s
μ_v^Y	3 calls/s
γ_{vi}	2.5 packets/s
η_v	$\frac{1}{100}$ packets/call
Q_{nv}	0.01
Q_{hv}	0.001
C	110 packets/s
l_d	100 to 150 packets
l_{di}	125 packets
μ_d^X	1 calls/s
λ_d	30 calls/s
η_d	$\frac{1}{125}$ 1/s
c_d	1.5 packets/s
Q_{od}	0.001
Q_{hd}	0.005

($M_{\mathrm{v}}, T_{\mathrm{v}}$ and β_{v}) by Min does not require the knowledge of γ_{v}. This implies that the three values for both Policies B and C are the same. Furthermore, because voice traffic is given the preemptive priority over data traffic and their three CAC parameters are determined independently of data traffic, the parameters and therefore the QoS measures do not change with l_{d}.

The performance measures of the voice calls under the three CAC policies are given in Fig. 2. Fig. 2(a) shows the call blocking and dropping probabilities for voice calls as a function of λ_{v}. It is observed that both Policy B and C have the same performance, due to the assumption that a small variation in γ_{v} does not affect the call level QoS measures for voice calls. They can guarantee the QoS upper bounds in all the traffic conditions. However, Policy A is able to provide the required service quality only when $\lambda_{\mathrm{v}} \leq \lambda_{vi} = 60$ calls/s. For heavier traffic conditions, such as when $\lambda_{\mathrm{v}} = 75$ calls/s, B_{nv} and D_{hv} are 5.9 and 8.1 times larger than the respective upper bounds ($Q_{\mathrm{nv}} = 0.01$ and $Q_{\mathrm{hv}} = 0.001$). The small fluctuation of the probabilities below the bounds is due to the fact that M_{v} is an integer variable in the problem formulation. Fig. 2(b) shows the throughput for voice users as a function of λ_{v}. Note that the throughput for the cell includes both the completed calls and handoff departures. For Policy A, the resources are enough to guarantee the QoS, when $\lambda_{\mathrm{v}} \leq \lambda_{vi}$. The small B_{nv} and D_{hv} leads to a throughput $\Theta_{\mathrm{v}} \approx \lambda_{\mathrm{v}} + h_{\mathrm{v}}$ in that region. However, when $\lambda_{\mathrm{v}} > \lambda_{vi}$, Θ_{v} cannot increase linearly with λ_{v}, due to the rapid increase of B_{nv} and D_{hv}. For Policy B and Policy C, as B_{nv} and D_{hv} are always kept close to the upper bounds (Q_{nv} and Q_{hv}), $\Theta_{\mathrm{v}} \approx \lambda_{\mathrm{v}} + h_{\mathrm{v}}$ in all the traffic load conditions. Policy B and C have the same call level throughput. At the packet level, Policiy C executes bandwidth allocation adjustment, the growth in Γ is slower than that in Policy B, so that more capacity can be spared for data users. The constraint in the growth of Γ is essential to the overall performance of the system (especially for the data traffic), as discussed in the following.



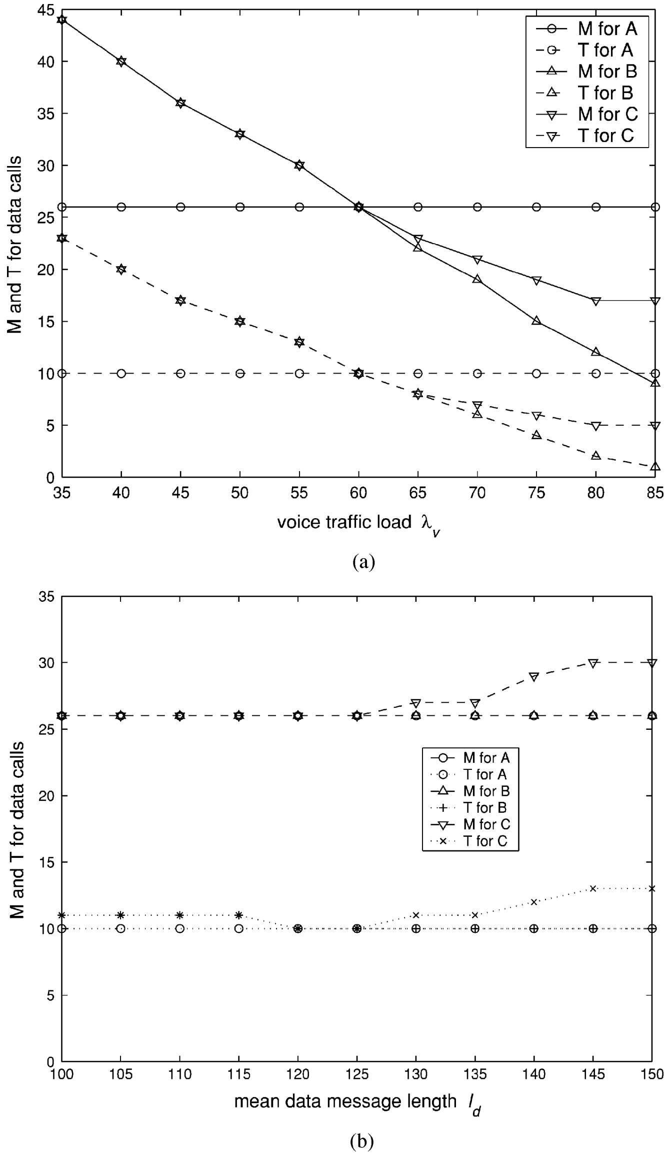
(a)

(b)

Fig. 2. Performance measures of the voice calls. (a) New call blocking and handoff call dropping probabilities. (b) Call level throughput.

Fig. 3 shows the values of M_{d} and T_{d} for the three policies. It is observed that 1) The values for Policy A are fixed for all the traffic conditions; 2) For Policy B, the values of M_{d} and T_{d} decrease rapidly as λ_{v} increases but do not change much with l_{d}. This observation reflects that the high priority voice traffic has a prominent impact on the resources available to the low priority data traffic. If bandwidth allocated to voice traffic is not limited, the data users may be completely blocked; 3) For Policy C, M_{d}, and T_{d} also decrease as λ_{v} increases but at a slower rate than that for Policy B. Because of the controlled growth of Γ, there is still reasonable leftover capacity for data users in the system at heavy traffic conditions. In particular, Policy C is able to admit more data users into the system when the data traffic load (l_{d}) is increased.

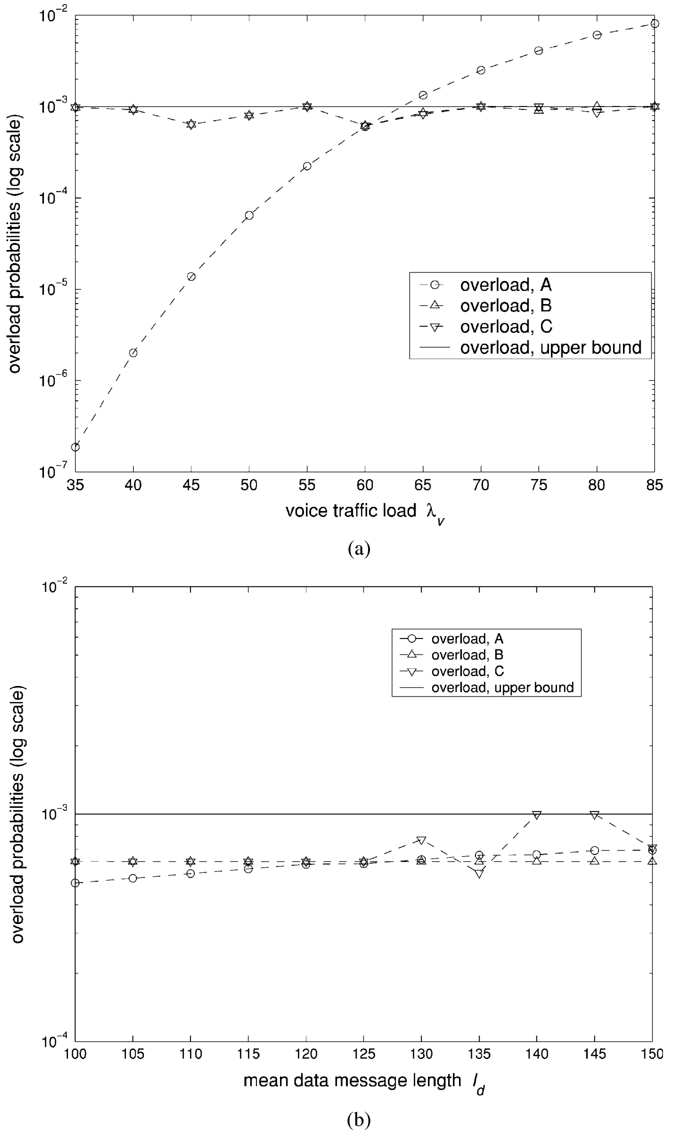
With the CAC parameters given in Fig. 3, the corresponding overload, handoff call dropping, and new call blocking probabilities for data calls are shown in Figs. 4–6. It is observed that a)



(a)

(b)

Fig. 3. Numbers M_{d} and T_{d} of data calls. (a) $l_{\mathrm{d}} = 150$ packets. (b) $\lambda_{\mathrm{v}} = 60$ calls/s.

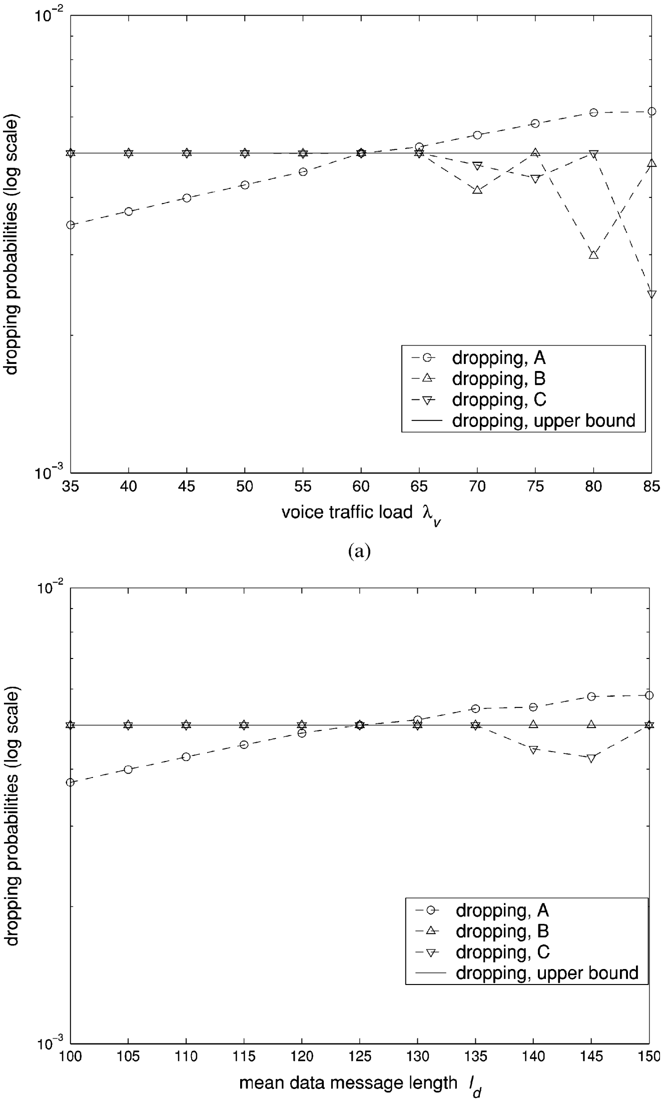


(a)

(b)

Fig. 4. Overload probability, Π_{od}, of data calls. (a) $l_{\mathrm{d}} = 150$ packets. (b) $\lambda_{\mathrm{v}} = 60$ calls/s.

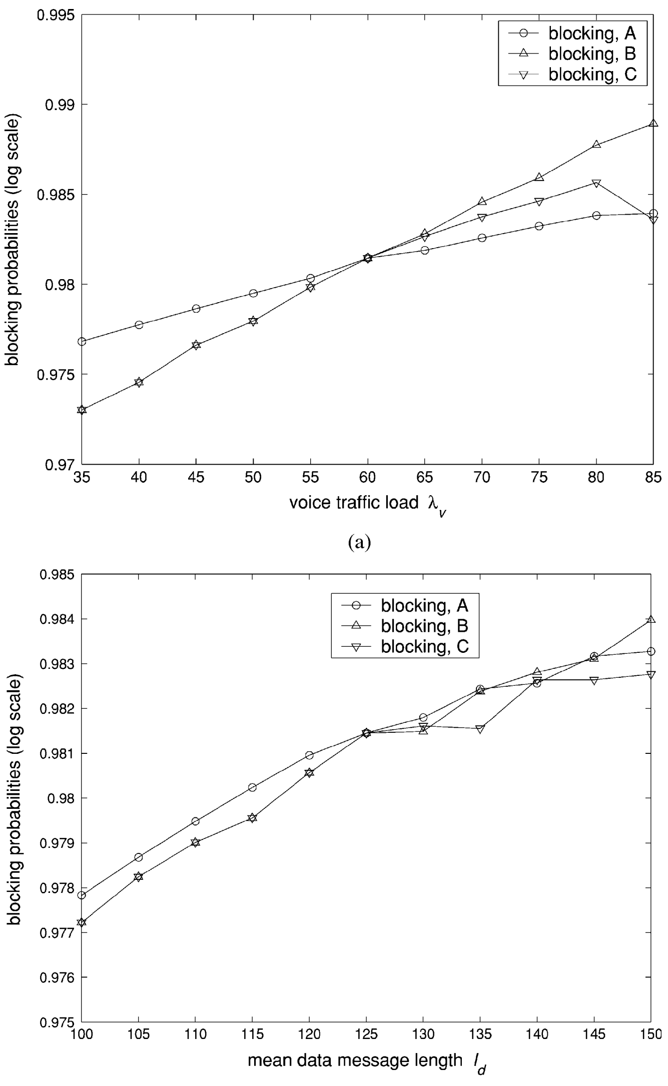
for Policy A, the overload probability increases sharply with the increase of voice traffic load λ_{v}. When λ_{v} becomes larger than the originally targeted load level, the leftover capacity becomes insufficient to guarantee the originally targeted QoS for data users. With the fixed CAC parameters, the data users are then aggressively admitted, leading to a rapid increase of the overload probability. The handoff call dropping probability requirement cannot be satisfied when either λ_{v} or l_{d} exceeds the originally targeted point. The new call blocking probability is close to 1 in all the traffic conditions, which shows that the leftover capacity is too small for the data traffic load $\lambda_{\mathrm{d}} = 30$ calls/s. On the other hand, the more than satisfactory QoS measure when $\lambda_{\mathrm{v}} < \lambda_{vi}$ translates into a low throughput in light traffic (to be demonstrated in Fig. 7). 2) Policies B and C can guarantee the overload probability and the handoff call dropping probability requirements under the different traffic load conditions. With the bandwidth allocation adjustment, Policy C can achieve a much larger throughput than Policy B in the heavy traffic load condition, as demonstrated in Fig. 7. Especially for the large l_{d} cases, Policy C performs obviously better than Policies A and B. Numerical results also show that the resource utilization efficiency for all the three policies is equal to or very close to 1, and is independent of the mean message length, due to the heavy data traffic load considered and the assumption that a single data user can use up the entire spectrum of the cell. Policy C has the highest utilization efficiency. In summary, Policy C is demonstrated to be superior to Policies A and B. The only disadvantage of Policy C is that the admitted voice users have to tolerate a reduction in the allocated resources when the traffic load is heavy, which is shown in Fig. 8. In light to medium traffic conditions ($\lambda_{\mathrm{v}} \leq 60$ Erlangs and $l_{\mathrm{d}} \leq 125$ packets), there is no bandwidth reduction. For medium to heavy traffic conditions, the reduction ranges from 0 to a maximum of 18%. As long as the reduction in bandwidth is kept below a certain level, voice quality will not degrade to an uncomfortable level. With its capability in maintaining call level QoS for both voice and data calls and, at the same time, achieving relatively high throughput and resource



(a)

(b)

Fig. 5. Handoff call dropping probability $D_{\rm hd}$ of data users. (a) $l_{\rm d} = 150$ packets. (b) $\lambda_{\rm v} = 60$ calls/s.



(a)

(b)

Fig. 6. New call blocking probability $B_{\rm nd}$ of data calls. (a) $l_{\rm d} = 150$ packets. (b) $\lambda_{\rm v} = 60$ calls/s.

utilization, Policy C is more suitable than both Policies A and B for the dynamic environment of the future cellular networks.

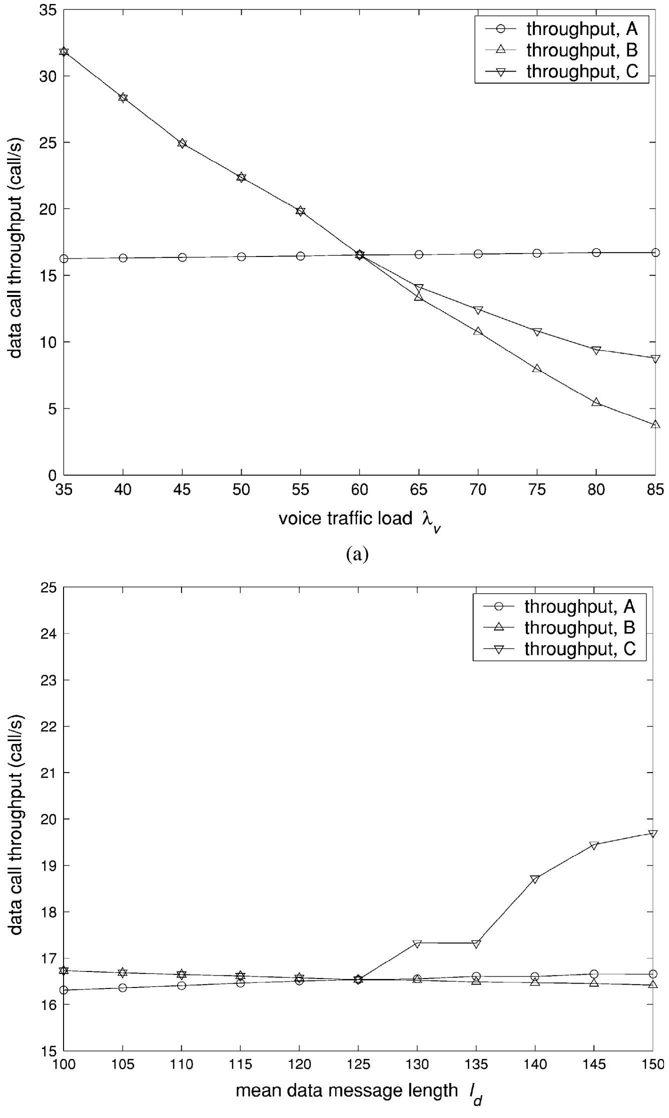
Note that in the preceding results, when Policy B or Policy C is involved, the relationship between the QoS performance and the traffic load is not monotonic. This is due to the integer CAC parameters $(M_{\rm v}, M_{\rm d})$ and the simultaneous adjustments of the admission region, the rate of voice calls, and the traffic load. However, our method of determining the CAC parameters can always satisfy the QoS specifications, and fluctuations of the QoS measures are only observed below the QoS upper bounds.

VI. Simulation Results

In the preceding, we demonstrate using Markovian analysis that the adaptive CAC is able to simultaneously provide satisfactory QoS to both voice and data users and maintain a relatively high resource utilization in a dynamic traffic load environment, by load adaption and bandwidth allocation adjustment. For mathematical tractability, the data call length (data file size) in packets is assumed to follow an exponential distribution. However, recent measurement-based modeling shows that the Internet data file size follows a lognormal distribution [39]. In this section, we use computer simulations to examine the impact of lognormal distribution. Two cases of CAC are simulated, with the lognormally and exponentially distributed data call length, respectively, and the performance measures are compared.

We simulate the CAC for a cell cluster of 19 cells, as shown in Fig. 9. The initial users existing in the system are uniformly distributed in all the cells.[3] In a cell, when a handoff happens, one of the six possible directions is randomly selected as the handoff direction. At the boundary cells when a handoff call moves out of the cluster, a handoff arrival is randomly generated to keep the handoff arrival rate to the cluster equal to the handoff departure rate from the cluster. The traffic conditions, the bandwidth and QoS requirements, and the CAC policy parameters are set the

[3] In fact, we run simulations with both uniform and nonuniform initial user distributions, and the QoS performance in the stationary state are the same.



(a)

(b)

Fig. 7. Throughput $\Theta_{\rm d}$ of data calls. (a) $l_{\rm d} = 150$ packets. (b) $\lambda_{\rm v} = 60$ calls/s.

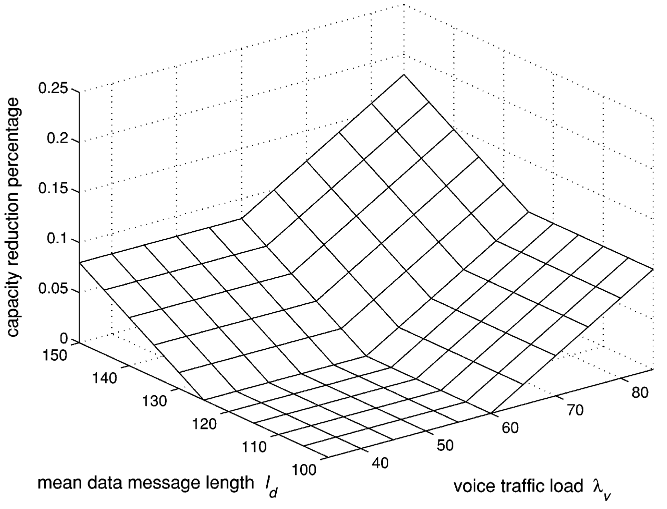


Fig. 8. Normalized reduction in the allocated bandwidth suffered by voice calls.

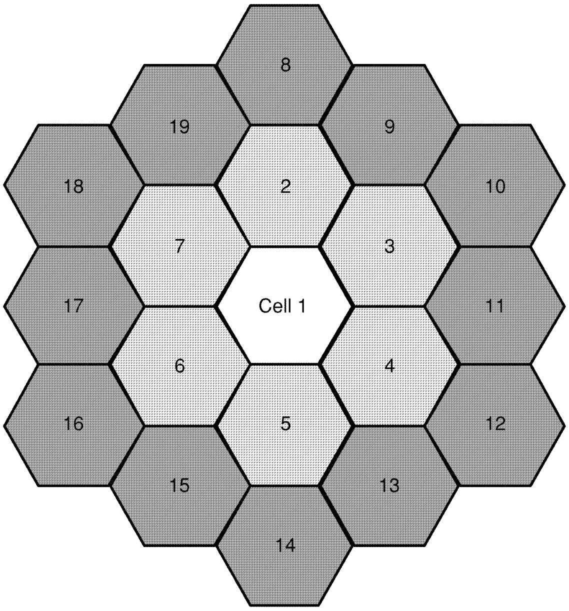
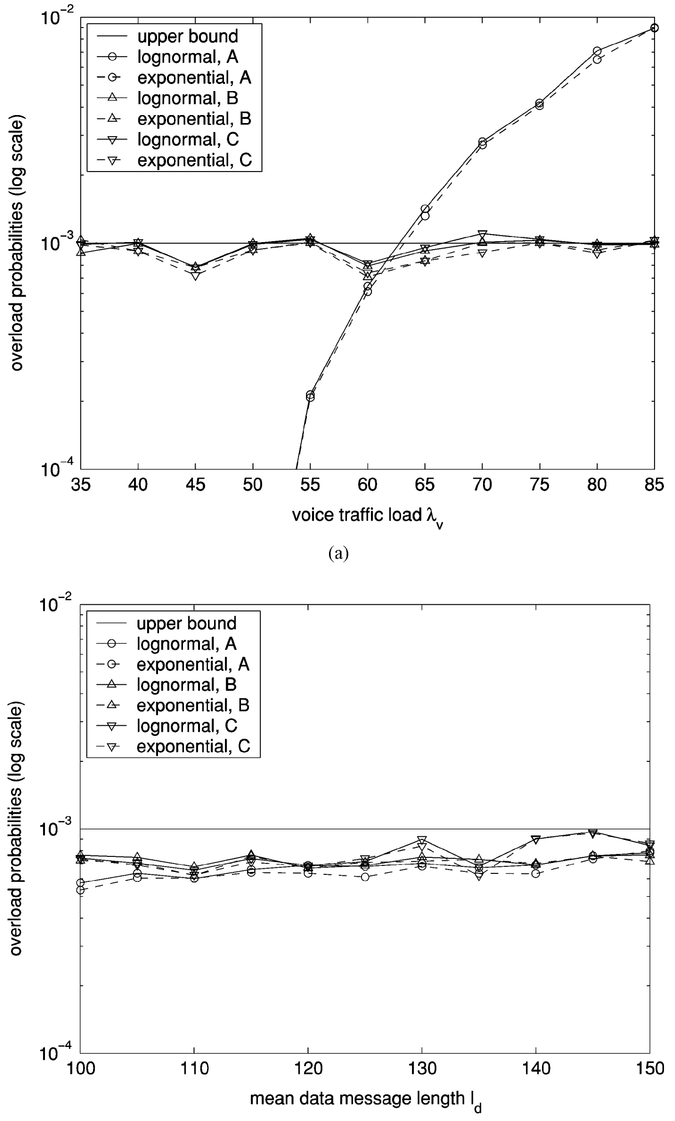


Fig. 9. Cell cluster used for simulation.

same as those used in the numerical analysis for both cases, except that the data call length distributions are lognormal and exponential, respectively. The conventional Monte Carlo approach is used, where 2×10^8 new arrivals (including both the voice and data calls) are simulated to get accurate estimation of the performance measures. As the CAC for voice calls is independent of the data traffic, the data call length distribution only affects the data CAC performance.

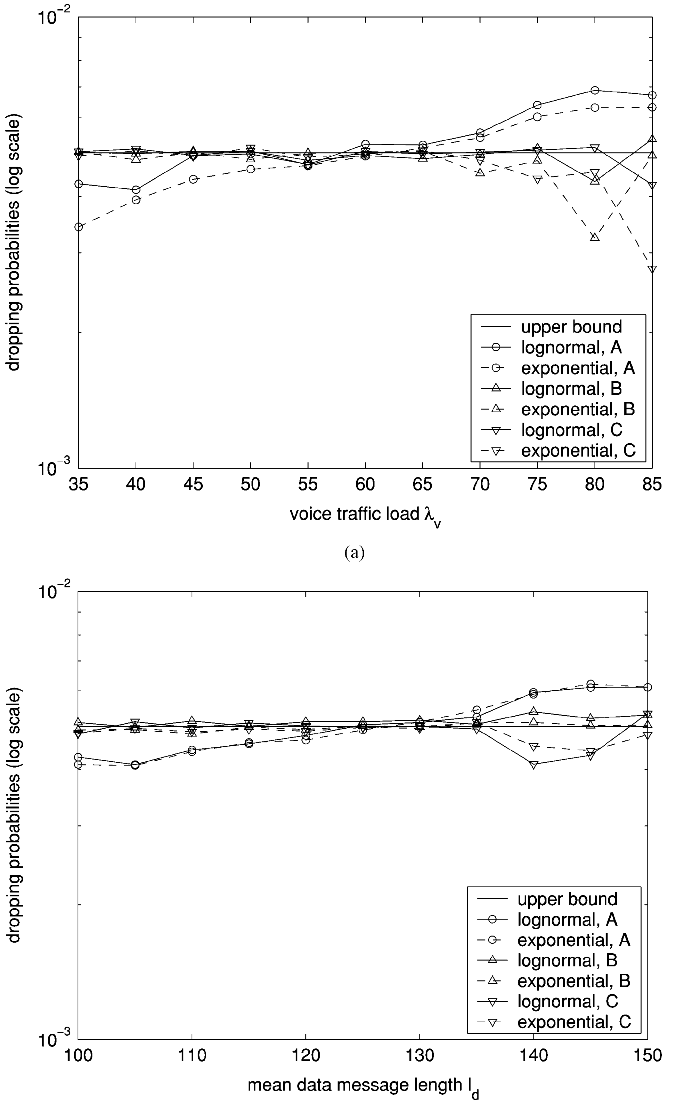
The estimated overload probabilities and handoff call dropping probabilities are presented in Figs. 10 and 11, respectively. The simulation results with an exponential data call length are almost exactly the same as the Markovian analysis results. Considering the estimation deviation due to the limited number of arrival samples, it can be concluded that the Markovian analysis is accurate for the exponential case, even though some simplified assumptions have to be made in the analysis. The Markovian analysis results are not included in Figs. 10 and 11 for clarity. It has been found from the simulations that: 1) The relationships among the performance curves from the different CAC policies, under the exponential distribution assumption, also apply to the case of lognormal distribution. Policy C again performs the best in QoS and resource utilization; 2) the performance measures in the lognormal case degrade to some degree as compared to those in the exponential case; 3) the deviations introduced by different data call length distributions are almost constant with respect to $\lambda_{\rm v}$ or $l_{\rm d}$; the lognormal curves seem to match the up-moved exponential curves. In other words, the deviation brought by the format of distribution is independent of its mean value (the traffic load) and the CAC policy. We do not show those untreated lognormal curves to not clutter the figures, and instead we show the compensated lognormal curves, to be discussed next. The three observations obtained from the simulations suggest that the impact of lognormal distribution can be



(a)

(b)

Fig. 10. Simulation results of the overload probability Π_{od} of data calls. (a) $l_{\mathrm{d}} = 150$ packets. (b) $\lambda_{\mathrm{v}} = 60$ calls/s.



(a)

(b)

Fig. 11. Simulation results of the handoff call dropping probability D_{hd} of data calls. (a) $l_{\mathrm{d}} = 150$ packets. (b) $\lambda_{\mathrm{v}} = 60$ calls/s.

compensated by conservatively applying the Markovian analysis results. Intuitively, the compensation may be implemented as follows. We can choose a compensation factor, termed as $\alpha(<1)$, and use the cell capacity αC instead of C in the Markovian analysis to calculate CAC parameters for the targeted traffic condition. The obtained CAC parameters and capacity C are then used in simulation, where the data call length distribution is set as lognormal with the same mean and variance. Compare the simulated performance measures with the target QoS, and adjust α correspondingly to reduce the difference between them. The adjustment of α is executed iteratively until all the performance measures obtained via simulation are close to (but smaller than) the QoS specifications. Moreover, as the QoS degradation due to the lognormal distribution does not depend much on the CAC policy and the traffic load condition, the compensation factor can be searched offline under the engineered load, and then directly applied to adaptive CAC policies. The lognormal curves plotted in Figs. 10 and 11 are obtained with the previously described compensation procedure. The compensation factor $\alpha = 0.92$ is determined under the engineered load $\lambda_{\mathrm{v}} = 60$ and $l_{\mathrm{d}} = 125$, and applied to all the other traffic conditions. In the figures, we can see that the compensated lognormal curves (obtained with capacity C) are a close match to the exponential curves (obtained with capacity αC) under all the CAC policies.

VII. Concluding Remarks

This paper proposes a priority-based resource sharing scheme and the optimal LFGCP CAC policies for voice/data integrated cellular networks. Handoffs due to user mobility, the heterogeneous traffic characteristics, and dynamic traffic load are taken into account. We develop a mathematical model that is able to describe the complex interaction between the voice and data traffic sharing the total resources. Based on such a model, the optimal CAC parameters for both voice and data traffic are determined for maximal resource utilization. When traffic conditions change dynamically, adaptive CAC policies with load adaption and rate adaption are proposed to enhance the system performance. Numerical results demonstrate that the proposed adaptive CAC policies not only have a better capability in QoS provisioning but also achieve a higher resource utilization efficiency, compared with the static CAC. Such benefits are achieved with degraded packet-level QoS provided to voice users who may experience a tolerable reduction in bandwidth allocation in a heavy traffic load condition. In the mathematical analysis, the data call length in packets is assumed to be exponentially distributed for tractability, but in practice, the length has been shown to be lognormally distributed. We use computer simulations to demonstrate that the impact of lognormal distribution can be compensated by conservatively applying the Markovian analysis results.

Note that voice service implies a symmetric communication mode. However, many Internet data services have the nature of asymmetric communications. In an asymmetric environment, data broadcast (e.g., the broadcast disk scheme [42] and the references therein) may be more efficient in wireless resource utilization. For future multiservice cellular networks, it will be an interesting and important research topic to investigate efficient resource allocation schemes that support both connection based and broadcast based multimedia applications.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments, which improved the presentation of this paper.

REFERENCES

[1] W. Mohr and W. Konhauser, “Access network evolution beyond third generation mobile communications,” *IEEE Commun. Mag.*, vol. 38, no. 12, pp. 122–133, Dec. 2000.

[2] K. D. Wong and V. K. Varma, “Supporting real-time IP multimedia services in UMTS,” *IEEE Commun. Mag.*, vol. 41, no. 11, pp. 148–155, Nov. 2003.

[3] S. Dixit, Y. Guo, and Z. Antoniou, “Resource management and quality of service in third generation wireless networks,” *IEEE Commun. Mag.*, vol. 39, no. 12, pp. 125–133, Feb. 2001.

[4] S. I. Maniatis, E. G. Nikolouzou, and I. S. Venieris, “QoS issues in the coverged 3G wireless and wired networks,” *IEEE Commun. Mag.*, vol. 40, pp. 44–53, Aug. 2002.

[5] O. Sallent, J. Perez-Romero, R. Agusti, and F. Casadevall, “Provisioning multimedia wireless networks for better QoS: RRM strategies for 3G W-CDMA,” *IEEE Commun. Mag.*, vol. 41, pp. 100–106, Feb. 2003.

[6] S.-C. Lo, G. Lee, W.-T. Chen, and J.-C. Liu, “Architecture for mobility and QoS support in all-IP wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 22, pp. 691–705, May 2004.

[7] A. Elwalid and D. Mitra, “Effective bandwidth of general Markovian traffic sources and admission control of high speed networks,” in *IEEE/ACM Trans. Netw.*, vol. 1, Jun. 1993, pp. 329–343.

[8] F. P. Kelly, “Notes on effective bandwidth,” in *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins, Eds. London, U.K.: Oxford Univ. Press, 1996, pp. 141–168.

[9] B. Epstein and M. Schwartz, “Reservation strategies for multimedia traffic in a wireless environment,” in *Proc. IEEE 45th Vehicular Technology Conf.*, vol. 1, Chicago, IL, Jul. 1995, pp. 165–169.

[10] M. Naghshineh and A. S. Acampora, “QoS provisioning in micro-cellular networks supporting multimedia traffic,” in *Proc. IEEE INFOCOM’95*, Boston, MA, 1995, pp. 1075–1084.

[11] F. S. Lai, J. Misic, and S. T. Chanson, “Complete sharing versus partitioning: Quality of service management for wireless multimedia networks,” in *Proc. 7th Int. Conf. Comput. Comm. Net.*, Lafayette, LA, Oct. 1998, pp. 584–593.

[12] Y. Fang and Y. Zhang, “Call admission control schemes and performance analysis in wireless mobile networks,” *IEEE Trans. Veh. Technol.*, vol. 51, no. 2, pp. 371–382, Mar. 2002.

[13] Y.-R. Huang, Y.-B. Lin, and J. M. Ho, “Performance analysis for voice/data integration on a finite mobile systems,” *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 367–378, Mar. 2000.

[14] S. C. Borst and D. Mitra, “Virtual partitioning for robust resource sharing: Computational techniques for heterogeneous traffic,” *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 668–678, Jun. 1998.

[15] R. Garg and H. Saran, “Fair bandwidth sharing among virtual networks: A capacity resizing approach,” *Proc. INFOCOM’00*, vol. 1, pp. 255–264, 2000.

[16] J. Yao, J. W. Mark, T. C. Wong, Y. H. Chew, K. M. L, and K.-C. Chua, “Virtual partitioning resource allocatin for multiclass traffic in cullular systems with QoS constraints,” *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 847–864, May 2004.

[17] P. B. Key, “Optimal control and trunk reservation in loss networks,” *Prob. Eng. Info. Sci*, vol. 4, pp. 203–242, 1990.

[18] R. J. Gibbens and F. P. Kelly, “Network programming methods for loss networks,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1189–1198, Sep. 1995.

[19] E. Bouillet, D. Mitra, and K. G. Ramakrishnan, “The structure and management of service level agreements in networks,” *IEEE J. Sel. Areas Commun.*, vol. 20, no. 4, pp. 691–699, May 2002.

[20] B. Li, L. Li, B. Li, K. M. Sivalingam, and X.-R. Cao, “Call admission control for voice/data integrated cellular networks: Performance analysis and comparative study,” *IEEE J. Sel. Areas Commun.*, vol. 22, no. 4, pp. 706–718, May 2004.

[21] C. W. Leong, “CAC for Voice and Data Services in Wireless Communications,” M.A.Sc. thesis, Univ. Waterloo, Waterloo, ON, Canada, 2001.

[22] R. Ramjee, D. Towsely, and R. Nagarajan, “On optimal call admission control in cellular networks,” *Wireless Netw.*, vol. 3, no. 1, pp. 29–41, 1997.

[23] C. Ho and C. Lea, “Improving call admission policies in wireless networks,” *Wireless Netw.*, vol. 5, no. 4, pp. 257–265, 1999.

[24] B. Li, C. Liu, and S. Chanson, “Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks,” *ACM/Baltzer J. Wireless Netw.*, vol. 4, pp. 279–290, Aug. 1998.

[25] O. Yu and V. Leung, “Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN,” *IEEE J. Sel. Areas Commun.*, vol. 15, no. 7, pp. 1208–1225, Sep. 1997.

[26] D. Levine, I. Akyildiz, and M. Naghshineh, “A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept,” in *IEEE/ACM Trans. Netw.*, vol. 5, Feb. 1997, pp. 1–12.

[27] B. M. Epstein and M. Schwartz, “Predictive QoS-based admission control for multiclass traffic in cellular wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 523–534, Mar. 2000.

[28] B. Li, L. Yin, K. Y. M. Wong, and S. Wu, “An efficient and adaptive bandwidth allocation scheme for mobile wireless networks based on on-line local parameter estimations,” *ACM/Baltzer J. Wireless Netw.*, vol. 7, pp. 107–116, Mar./Apr. 2001.

[29] Y. Cheng and W. Zhuang, “DiffServ resource allocation for fast handoff in wireless mobile Internet,” *IEEE Commun. Mag.*, vol. 40, pp. 130–136, May 2002.

[30] S. K. Das, S. K. Sen, K. Basu, and H. Lin, “A framework for bandwidth degradation and call admission control schemes for multiclass traffic in next-generation wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 21, no. 10, pp. 1790–1802, Dec. 2003.

[31] C. W. Ahn and R. S. Ramakrishna, “QoS provisioning dynamic connection-admission control for multimedia wireless networks using a hopfield neural network,” *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 106–117, Jan. 2004.

[32] D. Wu, Y. T. Hou, and Y.-Q. Zhang, “Scalable video coding and transport over broadband wireless networks,” *Proc. IEEE*, vol. 89, pp. 6–20, Jan. 2001.

[33] J. Liu, B. Li, Y. T. Hou, and I. Chlamtac, “On optimal layering and bandwidth allocation for multisession video broadcasting,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 2, pp. 656–667, Mar. 2004.

[34] W. S. Jeon and D. G. Jeong, “Call admission control for mobile multimedia communications with traffic asymmetry between uplink and downlink,” *IEEE J. Sel. Areas Commun.*, vol. 50, no. 1, pp. 59–66, Jan. 2001.

[35] G. Haring, R. Marie, R. Puigjaner, and K. Trivedi, “Loss formulas and their application to optimization for cullular networks,” *IEEE Trans. Veh. Technol.*, vol. 50, no. 3, pp. 664–673, May 2001.

[36] 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. Veh. Technol.*, vol. VT-35, no. 3, pp. 77–92, Aug. 1986.

[37] E. Del Re, R. Fantacci, and G. Giambene, "Handover and dynamic channel allocation techniques in mobile cellular networks," *IEEE Trans. Veh. Techol.*, vol. 44, no. 2, pp. 229–237, May 1995.

[38] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226–244, Jun. 1995.

[39] A. B. Downey, "The structural cause of file size distributions," *ACM SIGMETRICS Performance Evaluation Review*, vol. 29, pp. 328–329, Jun. 2001.

[40] C. W. Leong and W. Zhuang, "Soft QoS in call admission control for wireless personal communications," *Wireless Personal Commun.*, vol. 20, no. 2, pp. 127–144, Feb. 2002.

[41] S. M. Ross, *Introduction to Probability Models*, 8th ed. London, U.K.: Academic, 2003.

[42] S. Acharya, R. Alonso, M. Franklin, and S. Zdonik, "Broadcast disks: Data management for asymmetric communication environments," in *Proc. ACM SIGMOD Conf.*, San Jose, CA, May 1995.

Chi Wa Leong received the B.Sc. degree in computer engineering and the M.Sc. degree in electrical engineering, both from the University of Manitoba, Winnipeg, MB, Canada, and the M.A.Sc. degree in computer engineering from the University of Waterloo, Waterloo, ON, Canada.

Since graduation in 2001, he has been with the Information Technology Department of Banco Comercial de Macau. His research interests include call admission control, mobile networks, computer systems, E-commerce, and Internet applications.

Weihua Zhuang (M'93–SM'01) received the B.Sc. and M.Sc. degrees from Dalian Maritime University, Dalian, China, and the Ph.D. degree from the University of New Brunswick, Fredericton, NB, Canada, all in electrical engineering.

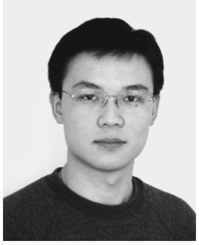
Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, where she is a Professor. She is a co-author of the textbook *Wireless Communications and Networking* (Englewood Cliffs, NJ; Prentice-Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning.

Dr. Zhuang is a licensed Professional Engineer in the Province of Ontario, Canada. She received the Outstanding Performance Award in 2005 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is an Associate Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *EURASIP Journal on Wireless Communications and Networking*.

Yu Cheng (S'01–M'04) received the B.E. and M.E. degrees from Tsinghua University, Beijing, China, in 1995 and 1998, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2003, all in electrical engineering.

From September 2003 to August 2004, he was a Postdoctoral Fellow in the Department of Electrical and Computer Engineering at the University of Waterloo. Since September 2004, he has been a Postdoctoral Fellow in the Department of Electrical and Computer Engineering at the University of Toronto, Toronto, ON, Canada. His research interests include QoS provisioning in IP networks, resource management, traffic engineering, and wireless/wireline interworking.

Dr. Cheng received a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC) in 2004.



Lei Wang (S'02) received the B.S.E. and M.S.E. degrees in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, in 1994 and 1997, respectively. He is currently working toward the Ph.D. degree in electrical engineering at the University of Waterloo, Waterloo, ON, Canada.

His research interests include admission control, resource management, and QoS provisioning in wireless networks.