Call Admission Control for Integrated On/Off Voice and Best-Effort Data Services in Mobile Cellular Communications

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

***Abstract*—This paper proposes a call admission control (CAC) policy for a cellular system supporting voice and data services, and providing a higher priority to handoff calls than to new calls. A procedure for searching the optimal admission region is given. The traffic flow is characterized by a three-dimensional (3-D) birth-death model, which captures the complex interaction between the on/off voice and best-effort data traffic sharing the total resources without partition. To reduce complexity, the 3-D model is simplified to an exact (approximate) 2-D model for voice (data). The mathematical expressions are then derived for the performance measures and for the minimal amount of resources required for quality-of-service (QoS) provisioning. Numerical results demonstrate that: 1) the proposed CAC policy performs well in terms of QoS satisfaction and resource utilization; 2) the approximate 2-D model for data traffic can achieve a high accuracy in the traffic flow characterization; and 3) the admission regions obtained by the proposed search method agree very well with those obtained by numerically solving the mathematical equations. Furthermore, computer simulation results demonstrate that the impact of lognormal distributed data file size is not significant, and may be compensated by conservatively applying the Markovian analysis results.**

***Index Terms*—Call admission control (CAC), quality-of-service (QoS), resource utilization, statistical multiplexing, voice/data.**

I. Introduction

QUALITY-OF-SERVICE (QoS) provisioning is one of the most significant features of the future multimedia wireless communications. By limiting the number of concurrent users in a system, call admission control (CAC) is critical in maintaining satisfactory QoS to the admitted users. For each call request, CAC makes the decision of whether or not to accept the user. The decision is based on the admission region in terms of the number of users and/or the amount of allocated resources to users. The objective is to simultaneously guarantee QoS and achieve high resource utilization. Employing cellular structure for frequency reuse results in handoff calls. Handoff users from adjacent cells are users already in service, and their service quality can be maintained by reservation of resources. The admission region for potential handoff users must be determined carefully to avoid resources being underused. Limited radio spectrum, user mobility, and heterogeneous nature of multimedia traffic constitute the major challenges to CAC. A solution for CAC in a mobile environment is, therefore, not simple to reach.

CAC with multiple call types and priorities has been addressed for quite a long time in wireline networks. The main approach is to assign a constant bandwidth, the peak rate, or an effective bandwidth, to a class of calls to guarantee the packet-level QoS [1], [2], and apply a certain resource partition paradigm among different classes to satisfy the call-level QoS [3], [4]. CAC in multimedia wireless networks can be viewed as an extension of the CAC problem in multimedia wireline networks to take into account handoff calls due to user mobility. However, the CAC based on constant bandwidth request and resource partitioning can not fully use the channel capacity. While the linear admission control ignores the packet-level statistical multiplexing effect [5], the resource partitioning among classes leads to the waste of bandwidth in those underloaded classes [4]. As best-effort data services (such as web browsing, e-mail, and file transfer) become more and more popular, we investigate CAC for a cellular system accommodating both on/off voice and best-effort data traffic in this paper. CAC for constant-rate voice and best-effort data has been considered in [6], under the assumption that the service time distribution of a data user is independent of its allocated resources. Studies related to admission control for integrated voice and data services in a code-division multiple-access (CDMA) system have been carried out in [7]–[9]. A queueing model for traffic flows is proposed in [7] to describe voice calls and data calls, under the assumption of constant-rate voice traffic. In [8] and [9], optimal policies are obtained to achieve target call-blocking probability and throughput for a single base station, without consideration of handoff traffic arrivals from adjacent cells.

In this paper, the proposed CAC policy uses the limited fractional guard channel policy (LFGCP) [10] to reserve resources (called guard bandwidth) exclusively for potential handoff calls to maintain service quality for admitted users. The major contributions lie in the development of a CAC policy for the cellular system integrating on/off voice and best-effort data services, and the development of a traffic model to characterize the interaction of the voice and data traffic flows. In comparison with the previous work on CAC, the proposed CAC can improve resource utilization while guaranteeing the packet-level and call-level QoS for both voice and data services in a multicell environment,

Paper approved by K. K. Leung, the Editor for Wireless Network Access and Performance of the IEEE Communications Society. Manuscript received July 27, 2000; revised August 7, 2001 and June 17, 2003. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Research Grant 155131. This paper was presented in part at IEEE Globecom, San Francisco, CA, November 2000.

The authors 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; ycheng@bbcr.uwaterloo.ca; l29wang@bbcr.uwaterloo.ca).

Digital Object Identifier 10.1109/TCOMM.2004.826247

where handoff calls have a higher priority than new calls: 1) the voice and data users share the total capacity C in a cell. Voice users have priority access to the cell capacity and can use the capacity up to $\Gamma(< C)$; 2) the time-variant residue capacity is for data services, which is considered as available-rate (best-effort) services, and therefore, the channel holding time distribution for data users depends on the allocated bandwidth; and 3) the proposed CAC exploits the statistical multiplexing among on/off voice calls and between voice and data traffic for high resource utilization. With both voice and data calls, the estimation of total capacity requirement is more complex than the previous work for constant-rate services only. Two methods are proposed to determine the admission region and allocated resources, based on a search procedure and on mathematical analysis. The proposed traffic model accurately captures the characteristics of on/off voice calls and the dependency of the best-effort data call service time on the arrival and service statistics of both voice and data calls. To reduce the computational cost, the traffic model is simplified, respectively, to an exact two-dimensional (2-D) model for voice traffic and to an approximate 2-D model with high accuracy for data traffic. Numerical results demonstrate that the proposed CAC policy can achieve both QoS satisfaction and high resource utilization.

II. System Model and the Proposed CAC Policy

A. System Model

Consider a cellular system supporting both voice and data services. We assume that the overall system is homogeneous in statistical equilibrium. 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. Consider a single test cell with a total bandwidth (capacity) denoted by C. Each voice call consists of an on/off packet stream, carrying real-time information. The on and off periods are exponentially distributed with mean b^{-1} and a^{-1}, respectively. The probability of a call in the on state is thus $p_{\mathrm{on}} = a/(a+b)$. During an on period, each voice user generates the packet stream at a constant rate, requiring a bandwidth of r_{v} for transmission. Using on/off detectors for voice calls, resources not used by silent voice calls can be used by other users to enhance the resource utilization. The total bandwidth allocated to voice traffic is restricted to $\Gamma(< C)$. With statistical multiplexing, $M_{\mathrm{v}}(> (\Gamma/r_{\mathrm{v}}))$ voice users can be served. However, when the instantaneous number of on voice users j is larger than Γ/r_{v}, the resources allocated to the admitted voice users are not sufficient. The situation is called voice overload, and we should limit the service disruption due to it. One solution is to temporarily reduce the amount of bandwidth allocated to each voice user to $\Gamma/j(< r_{\mathrm{v}})$. With the use of multilayered speech-encoding techniques [5], voice information can be separated into most and least significant portions, and anything in between. As a result, during an overload period, voice packets carrying less significant information can be dropped without severely degrading the voice quality at the receiver. This solution also prevents call disconnections, thereby maintaining the connection-level QoS promised to the users at the cost of degraded packet-level QoS. For high resource utilization, statistical multiplexing among on/off voice users is considered subject to the upper bound (maximal allowed value) of the voice overload probability. Given $j(>0)$, the amount of bandwidth allocated to each voice user, $\gamma_{\mathrm{v}}(j)$, is

$$\gamma_{\mathrm{v}}(j) = \begin{cases} r_{\mathrm{v}}, & j \le \Gamma/r_{\mathrm{v}} \\ \Gamma/j, & \text{otherwise.} \end{cases} \tag{1}$$

The information carried by a data call is non-real-time, and can be characterized by medium tolerance in transmission delay, but low tolerance in packet loss. Each data call consists of a single burst of length L_{d} in packets. For the transmission of several documents in a network session, each document constitutes a call. At any time, the system gives all the data calls the same priority, and therefore, allocates an equal amount of the leftover bandwidth by voice calls to each admitted data call [6]; as a result, the data traffic is considered as best-effort traffic, and the channel holding times for different message lengths will be different. This method of allocating resources to the different types of traffic is first proposed in [6]. The scheme guarantees a certain amount of bandwidth $(C - \Gamma)$ always available to data calls by not allowing the high-priority (voice) traffic to occupy the total resources. When there are i voice and k data users in the test cell and, out of the i voice users, j users are on, the share of the leftover bandwidth for each data user is

$$\gamma_{\mathrm{d}}(j,k) = \frac{C - j\gamma_{\mathrm{v}}(j)}{k} \ge \frac{C - \Gamma}{k}. \tag{2}$$

As the bandwidth allocated to each data user depends on the instantaneous j and k values, there are chances that the bandwidth drops below a critical threshold, denoted by c_{d}. The threshold is the minimum resources required to maintain each data link efficiently with the minimum service quality. If $\gamma_{\mathrm{d}}(j,k) < c_{\mathrm{d}}$, the service quality of the admitted data users is not satisfactory. This situation is called data overload, and the probability of its occurrence should be kept low by restricting the number of admitted data users. In a practical system, the resource reallocation (as the number of on voice users changes) is implemented via medium-access control, which requires both time and resource overhead.

In the cell, there are four types of call arrivals: new voice and data calls originating within the cell; and handoff voice and data calls coming from adjacent cells. These arrivals are assumed to be independent of each other. After the admission process, these arrivals will either be 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. For voice calls, we assume that the cell residence time, X_{v}, and the call duration, Y_{v}, are both exponentially distributed with mean $(\mu_{\mathrm{v}}^X)^{-1}$ and $(\mu_{\mathrm{v}}^Y)^{-1}$, respectively. The channel holding time of a voice call, $Z_{\mathrm{v}} = \min(X_{\mathrm{v}}, Y_{\mathrm{v}})$, also has an exponential distribution with mean $(\mu_{\mathrm{v}}^Z)^{-1} = (\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y)^{-1}$. New voice call arrivals are assumed to be Poisson with rate λ_{v}. The exponential channel holding time and Poisson new 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, X_{d}, and the data call length in packets (data file size) L_{d}, are assumed to be exponentially distributed with mean $(\mu_{\mathrm{d}}^{X})^{-1}$ and $(\mu_{\mathrm{d}}^{L})^{-1}$, respectively. With the bandwidth of a data channel, $\gamma_{\mathrm{d}}(j,k)$ from (2) given j and k, the call duration of a data user with message length L_d is $Y_{\mathrm{d}} = L_{\mathrm{d}}[\sum_{j,k}\gamma_{\mathrm{d}}(j,k)p_{\mathrm{vd}}(j,k)]^{-1}$, with mean $(\mu_{\mathrm{d}}^{Y})^{-1} = (\mu_{\mathrm{d}}^{L})^{-1}[\sum_{j,k}p_{\mathrm{vd}}(j,k)(C-j\gamma_{\mathrm{v}}(j))/k]^{-1}$, where $p_{\mathrm{vd}}(j,k)$ is the probability of having j on voice users and k data users in the cell. The channel holding time of a data user in the cell is then $Z_{\mathrm{d}} = \min(X_{\mathrm{d}}, Y_{\mathrm{d}})$. In general, there is no closed-form expression for the distribution of the call duration Y_{d}. To facilitate further analysis, we make the assumption that Y_{d} follows an exponential distribution with the mean $(\mu_{\mathrm{d}}^{Y})^{-1}$. As a result, the channel holding time Z_{d} is exponentially distributed with mean $(\mu_{\mathrm{d}}^{Z})^{-1} = (\mu_{\mathrm{d}}^{X} + \mu_{\mathrm{d}}^{Y})^{-1}$. New data call arrivals are assumed to be Poisson with the rate λ_{d}. The exponential channel holding time and Poisson new arrival process lead to a Poisson handoff arrival process of data calls, where the handoff rate is denoted as h_{d}.

The exponential distribution is considered here for tractability in the analysis, to provide some insight on CAC for integrated voice and data services to a certain degree. In today's high-speed networks, both the cell residence time and call duration time may deviate from the exponential distribution [11]. The impact of a nonexponential data file size is examined by computer simulation in Section V.

B. Proposed CAC Policy

The policy consists of two LFGCPs: one for handling the admission of voice calls; and the other for data calls. 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 and data calls, respectively. The overall policy can be described as follows, where i and k are the current numbers of admitted voice (including both on and off) and data users, respectively, in the cell. *A new voice (data) call is always accepted if* $i < T_{\mathrm{v}}(k < T_{\mathrm{d}})$, *is accepted with probability* $\beta_{\mathrm{v}}(\beta_{\mathrm{d}})$ *if* $i = T_{\mathrm{v}}(k = T_{\mathrm{d}})$, *and is always rejected if* $i > T_{\mathrm{v}}(k > T_{\mathrm{d}})$. *A handoff call is always accepted if* $i < M_{\mathrm{v}}(k < M_{\mathrm{d}})$, *and rejected, otherwise*. The LFGCP for voice is independent of data traffic, but the LFGCP for data is affected by voice traffic due to the low service priority of data traffic. The correlation between voice and data calls is captured in the admission region determination of Section II-C and the performance analysis of Section III. The admission region is represented by the six parameters $\{M_{\mathrm{v}}, T_{\mathrm{v}}, \beta_{\mathrm{v}}, M_{\mathrm{d}}, T_{\mathrm{d}}, \beta_{\mathrm{d}}\}$. The parameters $M_{\mathrm{v}}, T_{\mathrm{v}}, M_{\mathrm{d}}$, and T_{d} represent the numbers of the users which are determined in such a way that the resource utilization is maximized, while guaranteeing both the call-level and packet-level QoS requirements. As there is a control/signaling resource overhead associated with each of the available-rate data connections, when the available bandwidth for each data connection is very low, the resources for the connections are not used efficiently, due to the relative large overhead. Also, the number of base station transceivers for the data services is practically limited. Hence, it is necessary to limit the number of the non-real-time data users, even though the best-effort data sources can adapt to whatever bandwidth is available. The QoS measures consist of the new call blocking, handoff call dropping, and overload probabilities for voice (data) calls and packet dropping rate for voice calls, which are denoted as $B_{\mathrm{nv}}(B_{\mathrm{nd}}), D_{\mathrm{hv}}(D_{\mathrm{hd}}), \Pi_{\mathrm{ov}}(\Pi_{\mathrm{od}}), \Delta_{\mathrm{ov}}$, respectively. The QoS requirements are specified by the upper bounds of $\{B_{\mathrm{nv}}, D_{\mathrm{hv}}, \Pi_{\mathrm{ov}}, B_{\mathrm{nd}}, D_{\mathrm{hd}}, \Pi_{\mathrm{od}}\}$, denoted by $\{Q_{\mathrm{nv}}, Q_{\mathrm{hv}}, Q_{\mathrm{ov}}, Q_{\mathrm{nd}}, Q_{\mathrm{hd}}, Q_{\mathrm{od}}\}$. In practice, dropped data calls may actually enter a queue and be served at a later time when more resources for data calls are available. In this case, the proposed policy can be extended to have finite buffers for data handoff and/or new calls [12].

C. Determination of the Admission Regions

First consider, given C, how to determine the CAC parameters together with Γ. For voice calls, the algorithm `Min M` proposed in [10] is used to find the minimum required value of M_{v} and the corresponding values of T_{v} and β_{v}, such that two of the specified QoS upper bounds (Q_{nv} and Q_{hv}) are satisfied. Based on the monotonicity properties of the call blocking and dropping probabilities, the algorithm searches for the minimum M_{v} iteratively. The minimum M_{v} implies that the least amount of resources can be used to satisfy the voice-overload probability upper bound. The corresponding amount of resources assigned to voice calls, Γ, is therefore optimized to the minimum required value. The consequence is the maximum resource utilization for voice calls, and maximum leftover capacity for data users. For data calls, the parameters cannot be determined in a straightforward manner. This is due to their lower service priority and their nonconstant bandwidth allocation. With the given C, it may be impossible to satisfy all the QoS requirements, especially when the traffic load is high. Among the QoS measures, the new call blocking probability for data traffic is the least important. Hence, the CAC parameters are to be determined to guarantee only the five upper bounds $\{Q_{\mathrm{nv}}, Q_{\mathrm{hv}}, Q_{\mathrm{ov}}, Q_{\mathrm{hd}}, Q_{\mathrm{od}}\}$, and to minimize B_{nd}. Given the QoS requirements, the capacity 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 in Table I. The parameters for voice calls are determined in Step 1 to satisfy the call blocking and dropping probabilities due to the higher priority of voice calls. The minimum value of Γ is determined in Step 2 to satisfy the voice overload probability. The procedure (Steps 3–9) is then devoted to finding the three parameters for data calls, first to guarantee the overload probability (Step 3), then to guarantee the handoff call dropping probability (Steps 4–7), and finally, to minimize the new call blocking probability (Steps 8–9). If C is not sufficiently large, the procedure stops before Step 9.

Next, consider how to determine the minimum C value (together with Γ) and the six CAC parameters, so that the QoS requirements can be satisfied. Even though this is a resource planning issue, we address it because of its close relation to the CAC [10] and its usefulness in resource reallocation when the traffic load changes on a large scale. As the mean service time for data is upper bounded by the mean cell residence time, for a conservative measure, we use the cell residence time distribution to approximately represent the data service time distribution, which is exponential and independent of voice traffic. Hence, we can

TABLE I
PROCEDURE TO DETERMINE Γ AND ADMISSION REGION WHEN C IS GIVEN

1	determine (M_{v}, T_{v}, β_{v}) by MinM
	such that $B_{\mathrm{nv}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}) \le Q_{\mathrm{nv}}$ & $D_{\mathrm{hv}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}) \le Q_{\mathrm{hv}}$;
2a	$\Gamma := 0$;
b	while $\Pi_{\mathrm{ov}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}) > Q_{\mathrm{ov}}$
	$\Gamma := \Gamma + r_{\mathrm{v}}$;
c	if $\Gamma > C$ then terminate the procedure;
3 a	$\beta_{\mathrm{d}} := 0$; $M_{\mathrm{d}}(1) := 1$;
b	while $\Pi_{\mathrm{od}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{M_{\mathrm{d}}(1),M_{\mathrm{d}}(1)}) \le Q_{\mathrm{od}}$
	$M_{\mathrm{d}}(1) := M_{\mathrm{d}}(1) + 1$;
c	if $M_{\mathrm{d}}(1) = 1$ then terminate the procedure;
d	$T_{\mathrm{d}}(1) := M_{\mathrm{d}}(1) - 1$; $M_{\mathrm{d}}(1) := M_{\mathrm{d}}(1) - 1$; $m := 2$; $n := 1$;
4 a	$T_{\mathrm{d}}(m) = T_{\mathrm{d}}(m-1)$;
b	while $D_{\mathrm{hd}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}}(m),M_{\mathrm{d}}(n)}) > Q_{\mathrm{hd}}$ & $T_{\mathrm{d}}(m) > 0$
	$T_{\mathrm{d}}(m) := T_{\mathrm{d}}(m) - 1$;
c	$m := m + 1$; $n := n + 1$;
5 a	$T_{\mathrm{d}}(m) = T_{\mathrm{d}}(m-1)$; $M_{\mathrm{d}}(n) = M_{\mathrm{d}}(n-1)$;
b	while $\Pi_{\mathrm{od}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}}(m),M_{\mathrm{d}}(n)}) \le Q_{\mathrm{od}}$
	$T_{\mathrm{d}}(m) := T_{\mathrm{d}}(m) + 1$; $M_{\mathrm{d}}(n) := M_{\mathrm{d}}(n) + 1$;
c	$T_{\mathrm{d}}(m) := T_{\mathrm{d}}(m) - 1$; $M_{\mathrm{d}}(n) := M_{\mathrm{d}}(n) - 1$;
6	if $D_{\mathrm{hd}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}}(m),M_{\mathrm{d}}(n)}) > Q_{\mathrm{hd}}$ then
	if $T_{\mathrm{d}}(m) = T_{\mathrm{d}}(m-1) = 0$ then terminate the procedure;
	else $m := m + 1$; goto Step **4;**
	else goto Step 7;
7 a	$M_{\mathrm{d}} := M_{\mathrm{d}}(n)$;
b	while $D_{\mathrm{hd}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}}(m),M_{\mathrm{d}}}) \le Q_{\mathrm{hd}}$ & $M_{\mathrm{d}} \ge T_{\mathrm{d}}(m)$
	$M_{\mathrm{d}} := M_{\mathrm{d}} - 1$;
c	$M_{\mathrm{d}} := M_{\mathrm{d}} + 1$;
8 a	$T_{\mathrm{d}} := T_{\mathrm{d}}(m)$;
b	while $D_{\mathrm{hd}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}},M_{\mathrm{d}}}) \le Q_{\mathrm{hd}}$ &
	$\Pi_{\mathrm{od}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}},M_{\mathrm{d}}}) \le Q_{\mathrm{od}}$ & $T_{\mathrm{d}} \le M_{\mathrm{d}}$
	$T_{\mathrm{d}} := T_{\mathrm{d}} + 1$;
c	$T_{\mathrm{d}} := T_{\mathrm{d}} - 1$;
9	do a bisection search for β_{d} within $[0, 1]$:
	search for the largest value of β_{d} such that
	$D_{\mathrm{hd}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}},M_{\mathrm{d}}}) \le Q_{\mathrm{hd}}$ & $\Pi_{\mathrm{od}}(g^{\beta_{\mathrm{v}}}_{T_{\mathrm{v}},M_{\mathrm{v}}}, g^{\beta_{\mathrm{d}}}_{T_{\mathrm{d}},M_{\mathrm{d}}}) \le Q_{\mathrm{od}}$;

TABLE II
TRANSITION RATES FOR 3-D AND 2-D TRAFFIC MODELS

Transition rates for the original 3-D (i, j, k) traffic model		
transitions	rates	valid ranges
$U_i(i,j,k)$	$i(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y)$	$0 < i \le M_{\mathrm{v}}, 0 \le j < i, 0 \le k \le M_{\mathrm{d}}$
$\Lambda_j(i,j,k)$	$(i-j)a$	$0 < i \le M_{\mathrm{v}}, 0 \le j < i, 0 \le k \le M_{\mathrm{d}}$
$U_j(i,j,k)$	jb	$0 < i \le M_{\mathrm{v}}, 0 < j \le i, 0 \le k \le M_{\mathrm{d}}$
$\Lambda_k(i,j,k)$	$\lambda_{\mathrm{d}} + h_{\mathrm{d}}$	$0 \le i \le M_{\mathrm{v}}, 0 \le j \le i, 0 \le k < T_{\mathrm{d}}$
	$\beta_{\mathrm{d}}\lambda_{\mathrm{d}} + h_{\mathrm{d}}$	$0 \le i \le M_{\mathrm{v}}, 0 \le j \le i, k = T_{\mathrm{d}}$
	h_{d}	$0 \le i \le M_{\mathrm{v}}, 0 \le j \le i, T_{\mathrm{d}} < k < M_{\mathrm{d}}$
$U_k(i,j,k)$	$k\left[\mu_{\mathrm{d}}^X + \frac{(C - j\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^L}{k}\right]$	$0 \le i \le M_{\mathrm{v}}, 0 \le j \le i, 0 < k \le M_{\mathrm{d}}$
$\Lambda_{ij}(i,j,k)$	$\lambda_{\mathrm{v}} + h_{\mathrm{v}}$	$0 \le i < T_{\mathrm{v}}, 0 \le j \le i, 0 \le k \le M_{\mathrm{d}}$
	$\beta_{\mathrm{v}}\lambda_{\mathrm{v}} + h_{\mathrm{v}}$	$i = T_{\mathrm{v}}, 0 \le j \le i, 0 \le k \le M_{\mathrm{d}}$
	h_{v}	$T_{\mathrm{v}} < i < M_{\mathrm{v}}, 0 \le j \le i, 0 \le k \le M_{\mathrm{d}}$
$U_{ij}(i,j,k)$	$i(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y)$	$0 < i \le M_{\mathrm{v}}, i = j, 0 \le k \le M_{\mathrm{d}}$
Transition rates for the exact 2-D (i, j) traffic model		
$U_i(i,j)$	$i(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y)$	$0 < i \le M_{\mathrm{v}}, 0 \le j < i$
$\Lambda_j(i,j)$	$(i-j)a$	$0 < i \le M_{\mathrm{v}}, 0 \le j < i$
$U_j(i,j)$	jb	$0 < i \le M_{\mathrm{v}}, 0 < j \le i$
$\Lambda_{ij}(i,j)$	$\lambda_{\mathrm{v}} + h_{\mathrm{v}}$	$0 \le i < T_{\mathrm{v}}, 0 \le j \le i$
	$\beta_{\mathrm{v}}\lambda_{\mathrm{v}} + h_{\mathrm{v}}$	$i = T_{\mathrm{v}}, 0 \le j \le i$
	h_{v}	$T_{\mathrm{v}} < i < M_{\mathrm{v}}, 0 \le j \le i$
$U_{ij}(i,j)$	$i(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y)$	$0 < i \le M_{\mathrm{v}}$,$i = j$
Transition rates for the approximate 2-D (i, k) traffic model		
$\tilde{\Lambda}_i(i,k)$	$\lambda_{\mathrm{v}} + h_{\mathrm{v}}$	$0 \le i < T_{\mathrm{v}}, 0 \le k \le M_{\mathrm{d}}$
	$\beta_{\mathrm{v}}\lambda_{\mathrm{v}} + h_{\mathrm{v}}$	$i = T_{\mathrm{v}}, 0 \le k \le M_{\mathrm{d}}$
	h_{v}	$T_{\mathrm{v}} < i < M_{\mathrm{v}}, 0 \le k \le M_{\mathrm{d}}$
$\tilde{U}_i(i,k)$	$i(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y)$	$0 < i \le M_{\mathrm{v}}, 0 \le k \le M_{\mathrm{d}}$
$\tilde{\Lambda}_k(i,k)$	$\lambda_{\mathrm{d}} + h_{\mathrm{d}}$	$0 \le i \le M_{\mathrm{v}}, 0 \le k < T_{\mathrm{d}}$
	$\beta_{\mathrm{d}}\lambda_{\mathrm{d}} + h_{\mathrm{d}}$	$0 \le i \le M_{\mathrm{v}}, k = T_{\mathrm{d}}$
	h_{d}	$0 \le i \le M_{\mathrm{v}}, T_{\mathrm{d}} < k < M_{\mathrm{d}}$
$\tilde{U}_k(i,k)$	$k\left[\mu_{\mathrm{d}}^X + \frac{(C - \mu_{\mathrm{on}}(i))\mu_{\mathrm{d}}^L}{k}\right]$	$0 \le i \le M_{\mathrm{v}}, 0 < k \le M_{\mathrm{d}}$

use `Min M` to find the three data policy parameters to satisfy the call blocking and dropping requirements for data users. With the CAC parameters for data traffic, the minimum value of C can be determined by a bisection search over the interval from Γ to $\Gamma + M_{\mathrm{d}} \cdot c_{\mathrm{d}}$, to satisfy the data overload probability. The procedure can be summarized into three steps, given the six performance upper bounds, the capacity requirements, and the traffic conditions. Step 1 is to determine the parameters for voice calls. In Step 2, by assuming $\mu_{\mathrm{d}}^L = 0$, the data service time and call performance become independent of those of voice traffic. As a result, $\rho_{\mathrm{d}} = (\lambda_{\mathrm{d}} + h_{\mathrm{d}})/\mu_{\mathrm{d}}^X$ is constant, and `Min M` can be applied to data calls. In Step 3, the required value of C is found through a bisection search, letting μ_{d}^L assume its original value.

III. PERFORMANCE ANALYSIS

A. *Traffic Flow Model*

The voice and data traffic flows with Poisson arrivals and exponential service times can be modeled by a 3-D continuous-time birth-death process with state (i, j, k), where i is the number of admitted voice calls, of which j are in the on state, and k is the number of admitted data calls, with $0 \le j \le i \le M_{\mathrm{v}}$ and $0 \le k \le M_{\mathrm{d}}$. For tractability, the following assumptions are made: 1) there is no distinction between a new call and a handoff call after the call is admitted; 2) an admitted voice user starts from the on state; 3) a voice user is in the off state during handoff and/or service completion.

The transition rates are defined in Table II, where invalid transitions are assigned a rate of zero. Λ and U are used to represent a transition for a birth/arrival event and for a death/departure event, respectively. The subscripts for Λ and U are used to indicate whether any of i, j, and k will change after the transition occurs. The subscripts are for notational purposes only, and are not variables. Because of the preemptive priority for voice calls, in Table II, the transition rates related to voice calls do not depend on k for all valid states. On the other hand, as data calls use only leftover resources, the service rate U_k depends on both j and k. The transitions in Table II can further be explained as follows. Along the i direction, only death is allowed, because we assume that when a new or handoff voice user is admitted, this user starts from the on state. The admission of a voice user is represented by those diagonal transitions that cause both i

and j to be increased by one. Along the j direction, both birth and death are allowed to represent the on/off phenomenon of the voice traffic. Along the k direction, both birth and death are allowed, because data calls can come and leave as long as the actions are permitted by the policy.

The state-space cardinality of the 3-D model is approximately $(1/2)M_{\rm v}^2 M_{\rm d}$. There are no product-form solutions for the steady-state probabilities $p(i,j,k)$, which is defined in the region $\{0 \le i \le M_{\rm v}, 0 \le j \le i, 0 \le k \le M_{\rm d}\}$, and zero otherwise. For large $M_{\rm v}$ and $M_{\rm d}$, the computation required to solve all the local balance equations is not trivial. To reduce the computational complexity, the 3-D model can be reduced to a 2-D model in describing the resource occupancy of voice calls, as the transition rates related to voice calls are not dependent on k within the valid range. The reduced model is referred to as the 2-D (i,j) model with state (i,j), with transition rates also given in Table II, where $i (0 \le i \le M_{\rm v})$ is the number of voice users in the cell, and $j (0 \le j \le i)$ is the number of on voice users. Each state of the 2-D (i,j) model is a megastate consisting of $M_{\rm d}$ states of the original 3-D model. The transitions among the states of this 2-D model are governed by the same rules and assumptions made for any (i,j)-plane of the original 3-D model. The steady-state probabilities of the 2-D model, defined as $p_{\rm vv}(i,j)$, can be related to those of the 3-D model by $p_{\rm vv}(i,j) = \sum_{k=0}^{M_{\rm d}} p(i,j,k)$ for $\{0 \le i \le M_{\rm v}, 0 \le j \le i\}$, and zero otherwise. The simplification of the 3-D model to the 2-D model is an exact conversion, due to the access priority given to the voice service. Furthermore, because the transition rates $U_i(\cdot,\cdot), \Lambda_{ij}(\cdot,\cdot)$ and $U_{ij}(\cdot,\cdot)$ do not depend on j for $0 \le j \le i$, the 2-D model can be reduced to a 1-D model for admitted voice users, which is basically an $\mathrm{M/M}/M_{\rm v}/M_{\rm v}$ queue. This simplification is an exact conversion. The states of the 1-D model represent the number of all the admitted voice calls in the cell. The steady-state probabilities of the 1-D model, $p_{\rm v}(i)$, are related to those of the 2-D (i,j) model by $p_{\rm v}(i) = \sum_{j=0}^{i} p_{\rm vv}(i,j)$ for $0 \le i \le M_{\rm v}$, and zero otherwise.

B. QoS Measures for Voice Calls

From the 1-D traffic flow model, it can be derived that

$$p_{\rm v}(i) = \begin{cases} \frac{\rho_{\rm v}^i}{i!} p_{\rm v}(0), & 0 \le i \le T_{\rm v} \\ \frac{\rho_{\rm v}^i [\alpha_{\rm v} + (1-\alpha_{\rm v})\beta_{\rm v}] \alpha_{\rm v}^{i-T_{\rm v}-1}}{i!} p_{\rm v}(0), & T_{\rm v}+1 \le i \le M_{\rm v} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\rho_{\rm v} = (\lambda_{\rm v} + h_{\rm v})/\mu_{\rm v}^Z$ is the total voice traffic load in the cell, $\alpha_{\rm v} = h_{\rm v}/(\lambda_{\rm v} + h_{\rm v})$ is the fraction of the voice traffic load composed of handoff traffic, and $p_{\rm v}(0) = \{\sum_{i=0}^{T_{\rm v}} (\rho_{\rm v}^i / i!) + \sum_{i=T_{\rm v}+1}^{M_{\rm v}} (\rho_{\rm v}^i [\alpha_{\rm v} + (1-\alpha_{\rm v})\beta_{\rm v}] \alpha_{\rm v}^{i-T_{\rm v}-1}/i!)\}^{-1}$. With $p_{\rm v}(i)$, the call blocking and dropping probabilities can be calculated by

$$\begin{aligned} B_{\rm nv} &= (1-\beta_{\rm v}) p_{\rm v}(T_{\rm v}) + \sum_{i=T_{\rm v}+1}^{M_{\rm v}} p_{\rm v}(i) \\ D_{\rm hv} &= p_{\rm v}(M_{\rm v}). \end{aligned} \tag{4}$$

The overload probability and the packet dropping rate can be calculated in terms of the steady-state probabilities of the 2-D (i,j) model, which is $p_{\rm vv}(i,j) = \binom{i}{j} p_{\rm on}^j (1-p_{\rm on})^{i-j} p_{\rm v}(i)$. As the chances of an on voice user finding itself in the cell with $(j-1)$ other on voice users are proportional not only to $\sum_{i=j}^{M_{\rm v}} p_{\rm vv}(i,j)$ but also to j, the voice overload probability is

$$\Pi_{\rm ov} = \sum_{\lfloor \frac{\Gamma}{r_{\rm v}} \rfloor < j \le M_{\rm v}} \frac{j \sum_{i=j}^{M_{\rm v}} p_{\rm vv}(i,j)}{\sum_{l=1}^{M_{\rm v}} l \sum_{i=l}^{M_{\rm v}} p_{\rm vv}(i,l)} \tag{5}$$

where $\lfloor \cdot \rfloor$ is the floor function, $\sum_{i=j}^{M_{\rm v}} p_{\rm vv}(i,j)$ is the probability of having j on voice users, and the denominator is a normalization constant. The packet dropping rate is then

$$\Delta_{\rm v} = \frac{1}{r_{\rm v}} \sum_{\lfloor \frac{\Gamma}{r_{\rm v}} \rfloor < j \le M_{\rm v}} \left(r_{\rm v} - \frac{\Gamma}{j} \right) \frac{j \sum_{i=j}^{M_{\rm v}} p_{\rm vv}(i,j)}{\sum_{l=1}^{M_{\rm v}} l \sum_{i=l}^{M_{\rm v}} p_{\rm vv}(i,l)}. \tag{6}$$

C. QoS Measures for Data Calls

To calculate the performance measures, we need to know both the steady-state probabilities of k data calls $p_{\rm d}(k)$, and the steady-state probabilities of the 3-D model $p(i,j,k)$. To reduce computational complexity for the 3-D model with large $M_{\rm v}$ and $M_{\rm d}, p_{\rm d}(k)$ and $p(i,j,k)$ are calculated in the following based on an approximately condensed 2-D (i,k) model. There is a natural time-scale decomposition in our system that arises due to the large disparity between the duration of an on/off period, the call interarrival duration, and the channel holding time of a call. That is, normally $\{a,b\} \gg \{\lambda_{\rm v}, \mu_{\rm v}^Z\}$. Typically, each on or off period is in the neighborhood of 0.5 s [5], while the voice call interarrival duration and channel holding time are in the neighborhood of 5 s and 50 s, respectively [13]. As the interarrival duration and channel holding time of data calls are normally in the same magnitude as (or longer than) those of voice calls, it is reasonable to assume that $\{a,b\} \gg \{\lambda_{\rm d}, \mu_{\rm d}^Z\}$. Consequently, the 3-D state transitions of the j dimension due to on/off switching will reach equilibrium much faster than the transitions of the i and k dimensions due to call arrivals/departures, and the notion of nearly completely decomposable (NCD) Markov chains [14], [15] can be applied to our system. The 3-D model can be approximated by a 2-D model in which the on/off statistics of the j dimension for voice calls are averaged out [15], [16]. In other words, the transitions of the j dimension are ignored, as long as for each i, we know how many voice users are, on average, in the on state. The transition rates of the 2-D (i,k) model are summarized in the bottom part of Table II, where $\mu_{\rm on}(i) = \sum_{j=0}^{i} j \gamma_{\rm v}(j) \binom{i}{j} p_{\rm on}^j (1-p_{\rm on})^{i-j}$, representing the average amount of resources occupied by i admitted voice users. Although there are still no product-form solutions of the 2-D (i,k) model, the state space is reduced by a factor of $(M_{\rm v}/2)$. The steady-state probability of the 2-D (i,k) model $\tilde{p}_{\rm vd}(i,k)$ can then be solved numerically from the balance equations. From $\tilde{p}_{\rm vd}(i,k)$, the steady-state probability of the number of data users in the cell can be obtained approximately by

$$p_{\rm d}(k) \approx \sum_{i=0}^{M_{\rm v}} \tilde{p}_{\rm vd}(i,k) \tag{7}$$

for $0 \le k \le M_{\rm d}$, and zero otherwise. The values of $p_{\rm d}(k)$ can then be used to calculate the data call blocking and dropping probabilities, as described in (3) and (4) with all subscripts "v" replaced by "d."

To calculate the overload probability for data calls, we need to know the probability of the individual states $p(i,j,k)$ where the overload occurs, i.e., where $(C - j \cdot \gamma_{\mathrm{v}}(j))/k < c_{\mathrm{d}}$. For relatively large values of M_{v} and M_{d}, knowing the number of admitted data calls does not provide much information about the number of admitted voice calls, and vice versa; therefore, we assume that the number of admitted voice calls is independent of the number of admitted data calls, i.e.,

$$p(i,j,k) \approx p_{\mathrm{vv}}(i,j) p_{\mathrm{d}}(k). \tag{8}$$

With $p(i,j,k)$, the data overload probability can be calculated by (9), shown at the bottom of the page.

Note that in the condensed 2-D (i,k) model, both $\tilde{p}_{\mathrm{vd}}(i,k)$ and $p_{\mathrm{vv}}(i,j)$, and therefore, the obtained QoS measures for data calls are affected by voice traffic only through ρ_{v}. But in the accurate 3-D model, $\lambda_{\mathrm{v}}, h_{\mathrm{v}}, \mu_{\mathrm{v}}$ may impact the QoS of data traffic separately.

D. Calculation of the Minimum Γ *and* C

In addition to the search method in Section II-C, the required Γ and C values can be determined mathematically. To find Γ to satisfy the voice-overload probability upper bound Q_{ov}, we need to know the required admission region for voice calls. This can be obtained from Step 1 of the search method in Table I. We can then obtain the steady-state probability $p_{\mathrm{v}}(i)$ from (3). Let

$$\begin{aligned} f_\Gamma(\Gamma) = Q_{\mathrm{ov}} - \frac{1}{G_1} & \sum_{\lfloor \frac{\Gamma}{r_{\mathrm{v}}} \rfloor < i \le M_{\mathrm{v}}} i p_{\mathrm{on}} p_{\mathrm{v}}(i) \\ & \times \left[Q\left(\frac{\lfloor \frac{\Gamma}{r_{\mathrm{v}}} \rfloor - 0.5 - \xi_{i-1}}{\sigma_{i-1}} \right) - Q\left(\frac{i - 0.5 - \xi_{i-1}}{\sigma_{i-1}} \right) \right] \end{aligned} \tag{10}$$

where $G_1 = \sum_{j=1}^{M_{\mathrm{v}}} j \sum_{i=j}^{M_{\mathrm{v}}} p_{\mathrm{vv}}(i,j), \xi_l = l p_{\mathrm{on}}$ and $\sigma_l = \sqrt{l p_{\mathrm{on}}(1 - p_{\mathrm{on}})}$. It is shown in Appendix that the equation $f_\Gamma(\Gamma) = 0$, with the initial estimate for Γ given in (A5), can be solved numerically to obtain the minimum Γ value. Similarly, to find the minimum value of C in order to satisfy the data-overload probability upper bound Q_{od}, we need to use the `Min M` algorithm to obtain the required admission regions for both voice and data calls. Let

$$\begin{aligned} f_C(C) = Q_{\mathrm{od}} - & \sum_{i=0}^{M_{\mathrm{v}}} \sum_{j=0}^{i} p_{\mathrm{vv}}(i,j) \\ & \times \left[Q\left(\frac{\lfloor \frac{C - j \cdot \gamma_{\mathrm{v}}(j)}{c_{\mathrm{d}}} \rfloor - 0.5 - \rho_{\mathrm{d}}}{\sqrt{\rho_{\mathrm{d}}}} \right) \right. \\ & \left. - Q\left(\frac{M_{\mathrm{d}} - 0.5 - \rho_{\mathrm{d}}}{\sqrt{\rho_{\mathrm{d}}}} \right) \right]. \end{aligned} \tag{11}$$

As derived in Appendix, with the initial estimate for C given in (A10), the equation $f_C(C) = 0$ can be solved numerically for the required C value for a long message length of data calls.

IV. Numerical Results

Example 1. Accuracy of the Approximate 2-D (i,k) *Model:* Consider the system parameters given in columns 1 and 2 of Table III(a). The steady-state probability $p(k)$ and data call performance obtained from the original 3-D model and the approximate 2-D (i,k) model are given in Table III(a) (columns 4–7) and (b), with the precision of six decimal places. For the fixed voice traffic load $\rho_{\mathrm{v}} = 10$, two sets of $\{\lambda_{\mathrm{v}}, h_{\mathrm{v}}, \mu_{\mathrm{v}}^Z\}$ are examined. It is observed that the approximate 2-D (i,k) model and (7) and (8) of Section III-C are very accurate in describing the behaviors of data users in the case where $\{\lambda_{\mathrm{v}}, h_{\mathrm{v}}, \mu_{\mathrm{v}}^Z\} = (0.15, 0.05, 0.02)$ calls/s, i.e., the on/off jumping rate is around 10 times the call-level transition rate $\lambda_{\mathrm{v}} + h_{\mathrm{v}} + i\mu_{\mathrm{v}}^Z$, where i is the number of admitted voice calls in the cell. For the other case, where the call-level transition rate is in the same magnitude as on/off switching rate, numerical comparison shows that (7) remains accurate; however, (9) based on (8) starts to degenerate, where the 2-D approximate model underestimates the overload probability with a 50% error. Equation (7) is accurate because only one type of error is incurred when the channel holding time approaches the on/off period, which is the error due to the reduction of the 3-D model to the 2-D model. For Π_{od} based on (8) and (9), the error results from both the 3-D to 2-D model reduction and the assumption that the numbers of the admitted data and voice calls are independent. The error due to the independence assumption increases as the voice channel holding time approaches the on/off period, which does not represent a normal traffic load condition. The numerical results show that the the approximate 2-D (i,k) model for data calls works well when the on/off switching rate is over 10 times as large as the call-level transition rate.

Example 2. Admission Regions for a Given C: Consider the system parameters specified in column 2 of Table IV. For each value of $\{\rho_{\mathrm{v}}, \lambda_{\mathrm{d}}\}$, the admission region $\{M_{\mathrm{v}}, T_{\mathrm{v}}, \beta_{\mathrm{v}}, M_{\mathrm{d}}, T_{\mathrm{d}}, \beta_{\mathrm{d}}\}$ and the minimum value of Γ are found using the procedure proposed in Section II-C. We then calculate the performance to examine whether or not the QoS requirements are indeed satisfied. Since voice users have preemptive priority over data users, the admission region for voice users in terms of M_{v} and Γ has to be increased for increasing load ρ_{v} in order to satisfy the QoS requirements (see Steps 1 and 2). Because of the simultaneous increase in the admission region, the value of Γ, and the traffic load, and because of the fact that M_{v} is an integer variable, the relationship between the call-level performance and the traffic

$$\Pi_{\mathrm{od}} = \sum_{i=0}^{M_{\mathrm{v}}} \sum_{j=0}^{i} \sum_{\{k: 1 \le k \le M_{\mathrm{d}}, \frac{C - j \cdot \gamma_{\mathrm{v}}(j)}{k} < c_{\mathrm{d}}\}} \frac{k p(i,j,k)}{\sum_{i=0}^{M_{\mathrm{v}}} \sum_{j=0}^{i} \sum_{k=0}^{M_{\mathrm{d}}} k p(i,j,k)} \tag{9}$$

TABLE III
ACCURACY OF APPROXIMATE 2-D (i,k) MODEL: SYSTEM PARAMETERS AND $p_{\mathrm{d}}(k)$

system parameters		k	$(\lambda_{\mathrm{v}}, h_{\mathrm{v}}, \mu_{\mathrm{v}}^{Z})$ $= (0.15, 0.05, 0.02)$ packets/s		$(\lambda_{\mathrm{v}}, h_{\mathrm{v}}, \mu_{\mathrm{v}}^{Z})$ $= (1.5, 0.5, 0.2)$ packets/s	
			exact $p_{\mathrm{d}}(k)$	approximate $p_{\mathrm{d}}(k)$	exact $p_{\mathrm{d}}(k)$	approximate $p_{\mathrm{d}}(k)$
M_{v}	10 users	0	0.000000	0.000000	0.000000	0.000000
T_{v}	7 users	1	0.000000	0.000000	0.000001	0.000000
β_{v}	0.35 calls/s	2	0.000006	0.000006	0.000006	0.000006
ρ_{v}	10 Erlangs	3	0.000039	0.000039	0.000039	0.000039
r_{v}	2.5 packets/s	4	0.000203	0.000204	0.000202	0.000204
$\Gamma(C)$	20(40) (packets/s)	5	0.000858	0.000858	0.000854	0.000858
M_{d}	14 users	6	0.003048	0.003049	0.003036	0.003049
T_{d}	11 users	7	0.009350	0.009354	0.009322	0.009354
β_{d}	0.35	8	0.025244	0.025251	0.025189	0.025251
λ_{d}	0.15 calls/s	9	0.060853	0.060863	0.060765	0.060863
h_{d}	0.075 calls/s	10	0.132496	0.132508	0.132391	0.132508
μ_{d}^{X}	0.01 calls/s	11	0.263031	0.263036	0.262982	0.263036
μ_{d}^{L}	10^{-4} (packets/user)$^{-1}$	12	0.271905	0.271894	0.272005	0.271894
c_{d}	1.5 packets/s	13	0.152941	0.152926	0.153075	0.152926
a	1.66 changes/s	14	0.080024	0.080012	0.080133	0.080012
b	2.5 changes/s					

(a)

ACCURACY OF APPROXIMATE 2-D (i,k) MODEL: QOS MEASURES

$(\lambda_{\mathrm{v}}, h_{\mathrm{v}}, \mu_{\mathrm{v}}^{Z})$	performance	exact value	approximate value
$(0.15, 0.05, 0.02)$ packets/s	B_{nd}	0.675841	0.675805
	D_{hd}	0.080024	0.080012
	Π_{od}	5.22033×10^{-5}	4.86920×10^{-5}
$(1.5, 0.5, 0.2)$ packets/s	B_{nd}	0.676152	0.675805
	D_{hd}	0.080133	0.080012
	Π_{od}	9.433753×10^{-5}	4.86920×10^{-5}

(b)

load is not monotonic; fluctuations of the call-level performance below the upper bounds are observed in the results. As all the performance measures for voice calls are independent of the change in data traffic load, we show the performance for voice calls as functions of the voice traffic load ρ_{v} only. Fig. 1(a) shows the call blocking, dropping, and overload probabilities. It is clear that the CAC policy guarantees all the QoS requirements to voice users. The average packet dropping rate suffered by a voice user, Δ_{v}, is shown in Fig. 1(b). With Δ_{v} in the neighborhood of 0.5×10^{-3}, packet dropping due to the statistical multiplexing among voice users is expected to have a negligible effect on the audibility of voice calls. Our numerical results also demonstrate that using on/off statistical multiplexing achieves a significant improvement in resource utilization, where Γ with statistical multiplexing is obtained by the search method, and Γ with peak rate allocation is equal to $M_{\mathrm{V}} r_{\mathrm{V}}$. For example, at $\rho_{\mathrm{v}} = 20$, in serving the same number of voice users, the required bandwidth is reduced almost to one half (50%) by the use of multiplexing. Consequently, the CAC policy is not only able to deliver satisfactory service quality to voice users, but also able to achieve high resource utilization. For data calls, their performance depends on both the voice traffic load ρ_{v} and the data users' arrival rate λ_{d}. For clarity of presentation, however, Fig. 2 shows the performance of data calls as functions of λ_{d} only, with $\rho_{\mathrm{v}} = 15$ Erlangs. It is observed that the system guarantees the requirements of the call dropping and overload probabilities, and with the given C, the call-blocking probability increases with the traffic load λ_{d}.

Example 3. Determination of the Admission Region Together With Γ and C: Consider the system parameters given in column 3 of Table IV. The admission region, together with the required Γ and C values, is determined for each λ_{d} value, using the procedure described in Section II-C. The system performance is then evaluated according to the analysis presented in Section III. As *Example 2* demonstrates that the CAC policy works well for voice calls, this example focuses on the performance for data calls as functions of data users' arrival rate λ_{d} with voice traffic load $\rho_{\mathrm{v}} = 10$ Erlangs. The optimal values of Γ and C are shown in Fig. 3, for a large mean data message length, $(\mu_{\mathrm{d}}^{L})^{-1} = 25\,000$ packets. For a small value of λ_{d}, C is found to be equal to Γ. This is because the leftover capacity from within Γ is enough to cover the small amount of bandwidth required for data calls. As λ_{d} increases, C grows linearly away from Γ. Γ does not change, because voice traffic conditions remain unchanged. Since both C and λ_{d} are allowed to be increased simultaneously, the monotonic relationship between the

TABLE IV
TRAFFIC CONDITIONS AND QoS REQUIREMENTS FOR NUMERICAL EXAMPLES AND SIMULATIONS

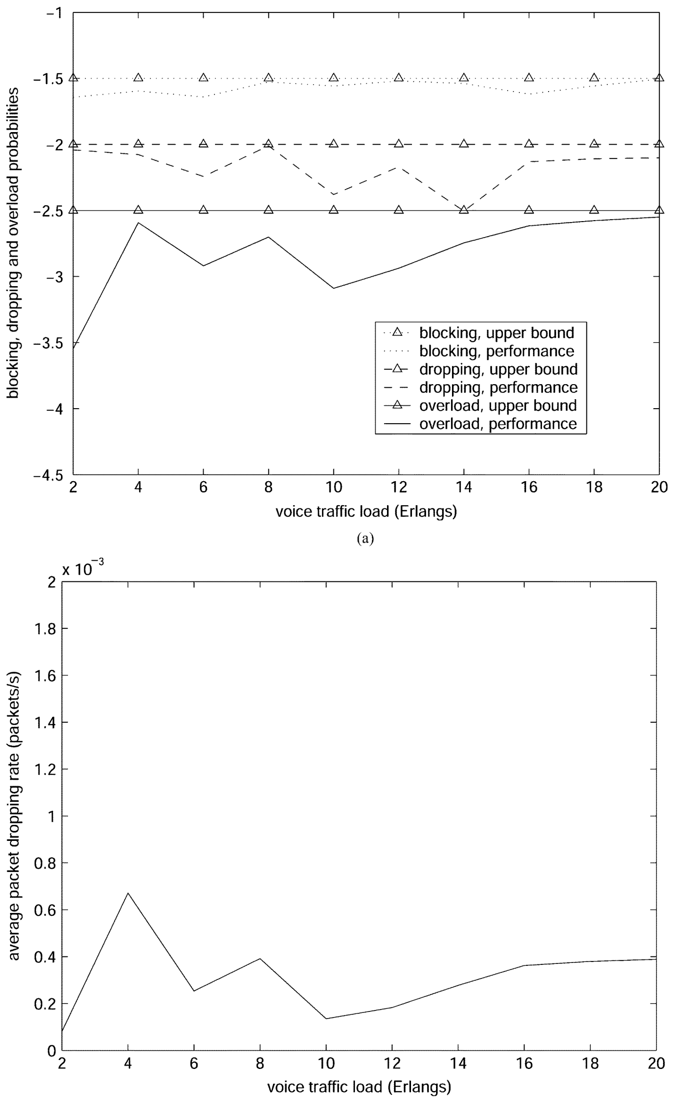
symbol	example 2	example 3	example 4(a)	example 4(b)	simulation
ρ_{v} (Erlangs)	2 to 20	10	20	20	
λ_{v} (calls/s)	0.4	0.4	0.4	0.4	0.4
h_{v} (calls/s)	0.2	0.2	0.2	0.2	
r_{v} (packets/s)	1.7	1.7	1.7	1.7	1.7
Q_{nv}	$10^{-1.5}$	$10^{-1.5}$	10^{-2}	10^{-2}	$10^{-1.5}$
Q_{hv}	10^{-2}	10^{-2}	10^{-3}	10^{-3}	10^{-2}
Q_{ov}	$10^{-2.5}$	10^{-3}	10^{-6} to 10^{-1}	10^{-3}	10^{-3}
λ_{d} (calls/s)	0.15 to 0.35	0.03 to 0.33		0.3	0.2 to 0.36
h_{d} (calls/s)	$0.5\lambda_{\mathrm{d}}$	$0.5\lambda_{\mathrm{d}}$		0.15	
μ_{d}^{X} (calls/s)	0.01	0.02		0.02	0.01
μ_{d}^{L} (packets/user)$^{-1}$	$\frac{1}{10000}$	$\frac{1}{25000}$		$\frac{1}{25000}$	$\frac{1}{75}$
c_{d} (packets/s)	0.5	0.5		0.5	1
Q_{nd}		10^{-1}		10^{-1}	$10^{-1.5}$
Q_{hd}	$10^{-1.5}$	$10^{-1.5}$		$10^{-1.5}$	10^{-2}
Q_{od}	10^{-2}	10^{-2}		10^{-6} to 10^{-1}	10^{-2}
μ_{v}^{X} (calls/s)					0.02
μ_{v}^{Y} (calls/s)					0.04
a (changes/s)					1.66
b (changes/s)					2.5

call-level performance and λ_{d} does not exist. Because of this and the fact that M_{d} is an integer variable, fluctuations can be observed from the call-level performance of data calls. Fig. 4 shows the call blocking, dropping, and overload probabilities for the same mean message length. Each of the three performance measures is below and very close to the corresponding required upper bound, indicating that the admission region for data calls is very close to the optimal values. As the procedure described in Section II-C is based on the assumption of a long message length, how the procedure performs for short data calls is studied. Fig. 4 also shows the results for $(\mu_{\mathrm{d}}^{L})^{-1} = 5000$ and 500 packets, respectively, with all other system parameters being the same. As mentioned in Section II-C, without taking μ_{d}^{L} into account, the proposed procedure gives the same admission region for all three mean message lengths. The effect of the message length to the optimality of the admission region can be observed from Fig. 4. For long and medium message lengths, the QoS requirements are satisfied with a small margin, translating into high resource utilization; however, for a short message length, the proposed procedure is conservative in QoS provisioning, resulting in the QoS measures much lower than the specified upper bounds. The over-satisfactory service provisioning to short data calls is achieved at the cost of resource utilization.

Example 4. Search of the Γ *and* C *Values:* This example is to verify the proposed search method for the required Γ and C values presented in Section II-C, by using the mathematical expressions given in Section III-D. For the required Γ value, the system parameters related to voice calls are specified in column 4 of Table IV. The required Γ values obtained by the search procedure and by numerically solving $f_{\Gamma}(\Gamma) = 0$ are shown in Fig. 5, as a function of the required voice-overload probability upper bound Q_{ov}. It is observed that the results of the two methods agree well with each other. Also shown is the initial estimate, Γ^{*}, as obtained from (A5). Note that for the initial estimate, two trends of values are observed, one below and one above the breakdown point at $Q_{\mathrm{ov}} \approx 10^{-3.5}$. This is shown in (A5) as a condition on I_{Γ}. The initial estimate is used in numerically solving $f_{\Gamma}(\Gamma) = 0$. The accuracy of the initial estimate, as compared directly with the optimal values of Γ, increases as Q_{ov} runs away from the breakdown point. For the required C value, the system parameters are specified in column 5 of Table IV, where the mean message length, $(\mu_{\mathrm{d}}^{L})^{-1}$, is purposely set at a large value for a fair comparison between the values obtained by the search and by solving $f_{C}(C) = 0$. The required C values, together with the initial estimate obtained from (A10), are also shown in Fig. 5, as a function of the data-overload probability upper bound Q_{od}. Again, the values of C obtained by the two different methods are very close to each other. There are also two trends of values for the initial estimate, due to the conditions related to I_{C} in (A10). The initial estimate is used in numerically solving $f_{C}(C) = 0$, and the breakdown point for the initial estimate is around $Q_{\mathrm{od}} = 10^{-2.75}$.

V. SIMULATION RESULTS

While the data call length in packets (data file size) is assumed an exponential distribution in the numerical analysis, the recent measurement-based modeling shows that the Internet data file size follows a lognormal distribution [17]. The impact of lognormal distribution is examined by computer simulations. We simulate the CAC for a cell cluster of 19 cells. 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



(a)

(b)

Fig. 1. QoS measures of voice calls under the CAC. (a) Blocking B_{nv}, dropping D_{hv} and overload Π_{ov} probabilities (in log scale) versus voice traffic load ρ_{v}. (b) Average packet dropping rate Δ_{v} versus voice traffic load ρ_{v}.

the cluster equal to the handoff departure rate from the cluster. Furthermore, both the voice and data handoff call arrival rates (h_{v} and h_{d}) are assumed to be known when deriving the the CAC parameters and CAC performance measures. 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, $h_{\mathrm{v}} = \sum_{i=1}^{M_{\mathrm{v}}} i\mu_{\mathrm{v}}^{X} p_{\mathrm{v}}(i)$ and $h_{\mathrm{d}} = \sum_{k=1}^{M_{\mathrm{d}}} k\mu_{\mathrm{d}}^{X} p_{\mathrm{d}}(k)$. The handoff arrival (departure) rates are dependent on state probabilities (and implicitly on the CAC parameters), while the state probabilities (and CAC parameters) are derived using the handoff arrival rates. To solve this problem, we use the iterative algorithm presented in [18]. For each point of simulation results, to be presented in the following, such an iterative algorithm is evoked to find $\{h_{\mathrm{v}}, M_{\mathrm{v}}, T_{\mathrm{v}}, \beta_{\mathrm{v}}, h_{\mathrm{d}}, M_{\mathrm{d}}, T_{\mathrm{d}}, \beta_{\mathrm{d}}\}$ in the equilibrium point.

Consider the system parameters given in column 6 of Table IV. As the Markovian analysis for voice traffic is accurate, the simulations focus on the data traffic. The admission region, together with the required Γ and C values to guarantee the QoS requirements, is determined for each λ_{d} value, using the procedure described in Section II-C and the iterative algorithm presented in [18]. The Γ and C values are then used in the simulations to estimate the QoS of data traffic. The conventional Monte Carlo approach is used, where 4×10^{5} new data call arrivals are simulated for each λ_{d} value to get a reasonably accurate estimation of the QoS measures within a reasonable simulation time. The QoS measures are the average values over the 19 cells within the cluster. Two cases of CAC are simulated, with the lognormally and exponentially distributed data call lengths, respectively. The data file size in both cases has the same mean and variance. The simulation results and the numerical analysis results are compared in Fig. 6, with an average data file size of 100 packets.

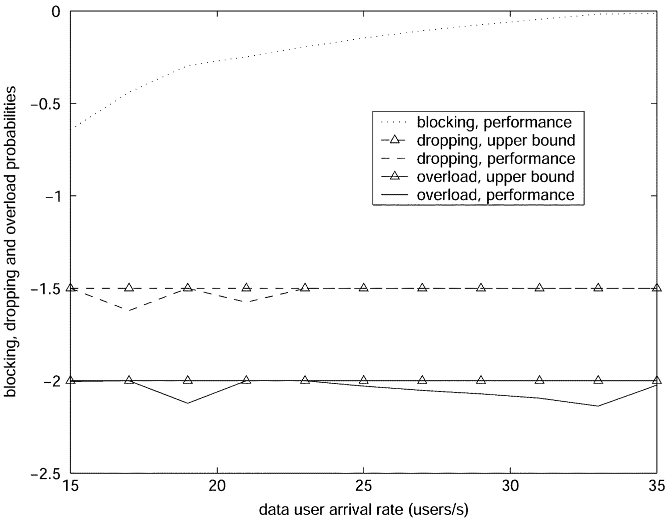


Fig. 2. QoS measures of data calls under the CAC. Blocking B_{nd}, dropping D_{hd} and overload Π_{od} probabilities (in log scale) versus data user arrival rate λ_{d}, with $\rho_{\mathrm{v}} = 15$ Erlangs.

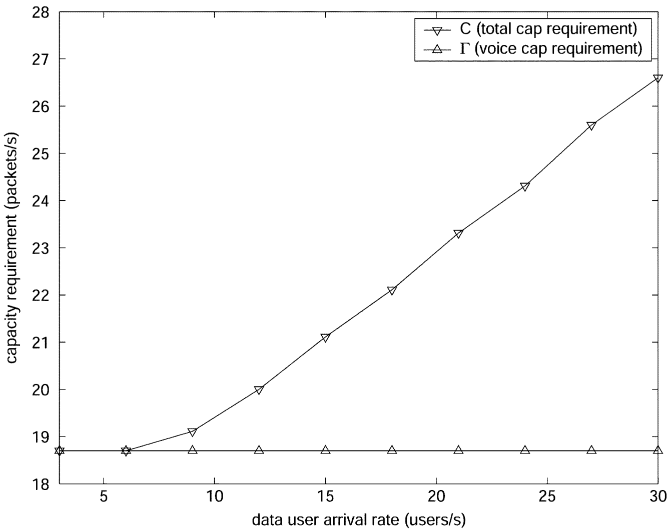


Fig. 3. Bandwidth requirements in terms of Γ and C versus data user arrival rate λ_{d}, with $\rho_{\mathrm{v}} = 10$ Erlangs.

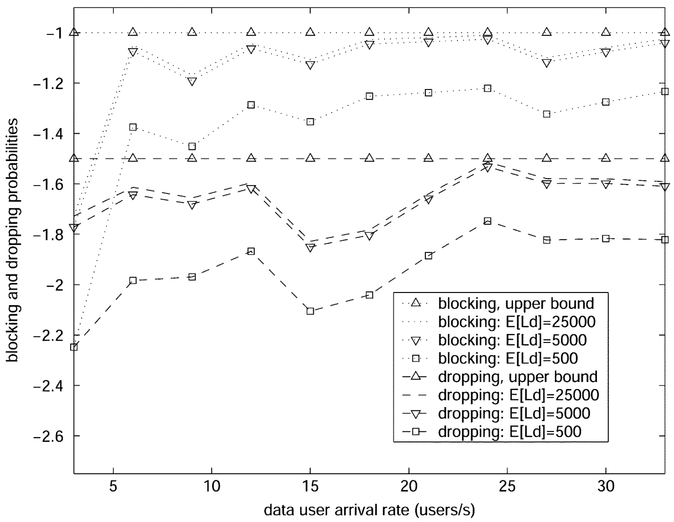


Fig. 4. New call blocking B_{nd} and handoff call dropping D_{hd} probabilities for data calls versus each data user arrival rate λ_{d}, with $\rho_{\mathrm{v}} = 10$ Erlangs, $(\mu_{\mathrm{d}}^{L})^{-1} = 500, 5000$, and $25\,000$ packets/s, respectively.

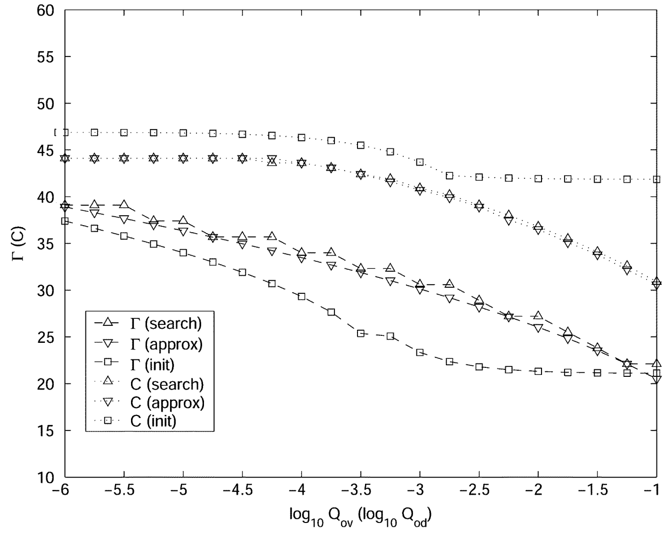
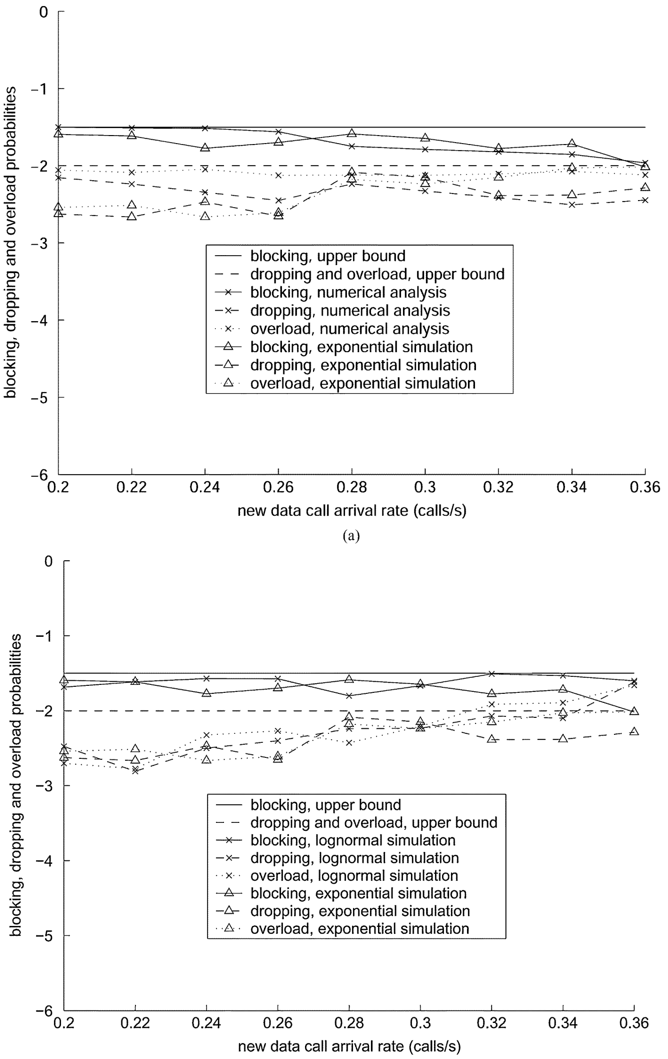


Fig. 5. Minimum value of $\Gamma(C)$ in packets/s versus the voice-overload probability $Q_{\mathrm{ov}}(Q_{\mathrm{od}})$.

From Fig. 6(a), we can see that the simulation results with exponential data call length are very close to the Markovian analysis results, as expected. The difference between the simulation results and the numerical analysis are due to two reasons. 1) When we iteratively search the handoff rate in the equilibrium point, the stop criteria is set as (|handoff departure rate−handoff arrival rate| /handoff arrival rate) < 10%. Therefore, the searched handoff rate is not exactly the handoff rate in the equilibrium point, which results in the single cell numerical analysis results deviating from the cluster simulation results. 2) The Monte Carlo estimation from the limited number of arrival samples also leads to some deviation. The interesting, but not surprising, observations are that the simulation results for the exponential case and those for the lognormal case are also very similar, as shown in Fig. 6(b). The observations are in agreement with the recent simulation results in [19] and [20]. However, a



(a)

(b)

Fig. 6. Comparison of analytical and simulation results for exponential data file size, and comparison of simulation results for exponential and lognormal data file sizes. (a) Numerical analysis results compared with exponential simulation results. (b) Lognormal simulation results compared with exponential simulation results.

theoretical explanation for the closeness between exponential results and lognormal results needs further research. From Fig. 6(b) and other simulation results with different mean data file sizes, we do observe that the exponential assumption leads to underestimation of the new call blocking, handoff call dropping, and overload probabilities, when the new data call arrival rate is high and the average data file size is large. When new data call arrival rate is larger than 0.3 calls/s, the QoS requirements on handoff call dropping and overload probabilities may not be satisfied in the lognormal case. These observations suggest that the impact of lognormal distribution can be compensated by conservatively applying the Markovian analysis results.

VI. Conclusion

CAC is an important function for future personal communication systems, but the design is challenging. In this paper, we proposed a CAC policy for the wireless system where both on/off voice and best-effort data services share the total available resources without partition. The proposed policy gives a higher priority to handoff calls than to new calls, and takes into account the statistical multiplexing among on/off voice calls and among voice and data calls for high resource utilization. The system performance under the proposed CAC policy, in terms of QoS provisioning, was evaluated, and the mathematical expressions for the required resources were derived. Numerical results demonstrate that, given the traffic load condition, the proposed CAC policy can guarantee the connection-level QoS requirements to mobile users with the use of a minimal amount of resources. In the numerical analysis, the data call length in packets is assumed to be exponentially distributed for tractability, but in practice, the length is shown to be lognormally distributed. We use computer simulations to demonstrate that the impact of lognormal distribution is not significant and may be compensated by conservatively applying the Markovian analysis results.

Appendix
Calculation of Γ and C

First, we want to find a minimum Γ value which can satisfy the voice-overload probability requirement. Letting $p'_{\mathrm{v}}(j) = \sum_{i=j}^{M_{\mathrm{v}}} p_{\mathrm{vv}}(i,j), G_1 = \sum_{l=1}^{M_{\mathrm{v}}} l p'_{\mathrm{v}}(l)$, and $A = \lfloor (\Gamma/r_{\mathrm{v}})\rfloor$, (5) becomes

$$\begin{aligned}\frac{1}{G_1}&\sum_{A<j\le M_{\mathrm{v}}} j p'_{\mathrm{v}}(j)\\ &= \frac{1}{G_1}\sum_{A<j\le M_{\mathrm{v}}} j \sum_{i=j}^{M_{\mathrm{v}}} \binom{i}{j} p_{\mathrm{on}}^{j}(1-p_{\mathrm{on}})^{i-j} p_{\mathrm{v}}(i)\\ &= \frac{1}{G_1}\sum_{A<i\le M_{\mathrm{v}}} p_{\mathrm{v}}(i)\cdot i\cdot p_{\mathrm{on}}\\ &\quad\times \sum_{A-1<j\le i-1} \binom{i-1}{j} p_{\mathrm{on}}^{j}(1-p_{\mathrm{on}})^{i-1-j}\\ &\approx \frac{1}{G_1}\sum_{A<i\le M_{\mathrm{v}}} i p_{\mathrm{on}} p_{\mathrm{v}}(i)\int_{A-1+0.5}^{i-1+0.5}\frac{1}{\sqrt{2\pi}\sigma_{i-1}}\\ &\quad\times \exp\left[-\frac{(s-\xi_{i-1})^2}{2\sigma_{i-1}^2}\right] ds\end{aligned}$$

where we have made use of the fact that the binomial distribution within the inner summation can be approximated by a Gaussian distribution with mean $\xi_{i-1} = (i-1)p_{\mathrm{on}}$ and standard deviation $\sigma_{i-1} = \sqrt{(i-1)p_{\mathrm{on}}(1-p_{\mathrm{on}})}$. The accuracy of the approximation improves as $i-1$ increases, and as p_{on} approaches $(1/2)$ [21]. By the use of the standard Q-function $Q(x) = (1/\sqrt{2\pi})\int_x^\infty \exp(-(s^2/2))\, ds$, the voice-overload probability can be approximately represented as

$$\Pi_{\mathrm{ov}} \approx \frac{1}{G_1}\sum_{A<i\le M_{\mathrm{v}}} i p_{\mathrm{on}} p_{\mathrm{v}}(i)\left\{Q\left(\frac{A-0.5-\xi_{i-1}}{\sigma_{i-1}}\right) - Q\left(\frac{i-0.5-\xi_{i-1}}{\sigma_{i-1}}\right)\right\}. \quad \text{(A1)}$$

Similarly, we have

$$G_1 \approx \sum_{1\le i\le M_{\mathrm{v}}} i p_{\mathrm{on}} p_{\mathrm{v}}(i)\left[Q\left(\frac{-0.5-\xi_{i-1}}{\sigma_{i-1}}\right) - Q\left(\frac{i-0.5-\xi_{i-1}}{\sigma_{i-1}}\right)\right]. \quad \text{(A2)}$$

Let

$$f_\Gamma(\Gamma) = Q_{\mathrm{ov}} - \frac{1}{G_1}\sum_{\lfloor \frac{\Gamma}{r_{\mathrm{v}}}\rfloor < i \le M_{\mathrm{v}}} i p_{\mathrm{on}} p_{\mathrm{v}}(i) \times \left[Q\left(\frac{\lfloor \frac{\Gamma}{r_{\mathrm{v}}}\rfloor - 0.5-\xi_{i-1}}{\sigma_{i-1}}\right) - Q\left(\frac{i-0.5-\xi_{i-1}}{\sigma_{i-1}}\right)\right]$$

where we have replaced A with $\lfloor(\Gamma/r_{\mathrm{v}})\rfloor$. The equation $f_\Gamma(\Gamma) = 0$ can then be solved to obtain an approximate value for Γ. Solving equations numerically usually requires an initial estimate of the result, which can be obtained as follows:

1) apply the approximation $Q(x) \approx (1/\sqrt{2\pi}x) \exp(-(x^2/2))$ to the Q-function involving A;
2) make use of the fact that the most dominant term in the summation of (A1) occurs when $i = M_{\mathrm{v}}$ and retain only this term;
3) replace Π_{ov} with Q_{ov};
4) ignore the correction factor 0.5;
5) move all terms which do not involve A to the left-hand side (LHS) and let the entire LHS equal to I_Γ.

We can then transform both sides of the equation as follows:

$$I_\Gamma = \mathrm{LHS} = \sqrt{2\pi}\left\{\frac{G_1\Pi_{\mathrm{ov}}}{M_{\mathrm{v}} p_{\mathrm{on}} p_{\mathrm{v}}(M_{\mathrm{v}})} + Q\left(\frac{M_{\mathrm{v}} - \xi_{M_{\mathrm{v}}-1}}{\sigma_{M_{\mathrm{v}}-1}}\right)\right\} \quad \text{(A3)}$$

$$I_\Gamma \approx \mathrm{RHS} = \frac{\sigma_{M_{\mathrm{v}}-1}}{A-\xi_{M_{\mathrm{v}}-1}}\exp\left[-\frac{1}{2}\left(\frac{A-\xi_{M_{\mathrm{v}}-1}}{\sigma_{M_{\mathrm{v}}-1}}\right)^2\right]. \quad \text{(A4)}$$

In (A4), when $I_\Gamma > 1$, the first term on the right-hand side (RHS) dominates because the exponential term is always smaller than or equal to one; otherwise, when $I_\Gamma < 1$, the exponential term dominates. As a result, for $I_\Gamma > 1, I_\Gamma \approx \sigma_{M_{\mathrm{v}}-1}/(A-\xi_{M_{\mathrm{v}}-1})$. Rearranging the terms, we have $A = \lfloor(\Gamma/r_{\mathrm{v}})\rfloor \approx (\sigma_{M_{\mathrm{v}}-1}/I_\Gamma) + \xi_{M_{\mathrm{v}}-1}$, and thus $\Gamma \approx [(\sigma_{M_{\mathrm{v}}-1}/I_\Gamma) + \xi_{M_{\mathrm{v}}-1}]r_{\mathrm{v}}$. Similarly, for $I_\Gamma < 1$, we have $I_\Gamma \approx \exp[-(1/2)((A-\xi_{M_{\mathrm{v}}-1})/\sigma_{M_{\mathrm{v}}-1})^2]$, or $\Gamma \approx [\sigma_{M_{\mathrm{v}}-1}\sqrt{-2\ln I_\Gamma} + \xi_{M_{\mathrm{v}}-1}]r_{\mathrm{v}}$. Let Γ^* be the initial estimate for Γ. The expression for Γ^* can be summarized as follows:

$$\Gamma^* \approx \begin{cases}\left[\frac{\sigma_{M_{\mathrm{v}}-1}}{I_\Gamma} + \xi_{M_{\mathrm{v}}-1}\right] r_{\mathrm{v}}, & I_\Gamma \ge 1\\ [\sigma_{M_{\mathrm{v}}-1}\sqrt{-2\ln I_\Gamma} + \xi_{M_{\mathrm{v}}-1}]r_{\mathrm{v}}, & I_\Gamma < 1\end{cases} \quad \text{(A5)}$$

where I_Γ can be obtained from (A3).

Similar to the case for Γ, next we can derive an expression to numerically obtain the value of C to guarantee the required data overload probability. In (9), by letting $G_2 = \sum_{i=0}^{M_{\mathrm{v}}}\sum_{j=0}^{i}\sum_{k=0}^{M_{\mathrm{d}}} k p(i,j,k) = \sum_{k=0}^{M_{\mathrm{d}}} k p_{\mathrm{d}}(k)$, splitting the three summations, using the approximate expression for $p(i,j,k)$ from (8), and $A(j) = \lfloor(C - j\cdot\gamma_{\mathrm{v}}(j))/c_{\mathrm{d}}\rfloor$, (9) becomes $\Pi_{\mathrm{od}} \approx$

$(1/G_2)\sum_{i=0}^{M_\mathrm{v}}\sum_{j=0}^{i} p_\mathrm{vv}(i,j)\sum_{A(j)<k\le M_\mathrm{d}} kp_\mathrm{d}(k)$. Since there is no closed-form expression for $p_\mathrm{d}(k)$, we need to make some assumptions before proceeding any further. Assume that $g^0_{M_\mathrm{d},M_\mathrm{d}}$ is used for data calls. This assumption is reasonable, as long as the difference between M_d and T_d is small, as in most of the cases. Nonetheless, a C value obtained based on $g^0_{M_\mathrm{d},M_\mathrm{d}}$ is always an upper bound of the actually required C for the more restrictive policy $g^{\beta_\mathrm{d}}_{T_\mathrm{d},M_\mathrm{d}}$. For long data calls, the mean channel holding time approaches the constant $(\mu^X_\mathrm{d})^{-1}$. With the use of $g^0_{M_\mathrm{d},M_\mathrm{d}}$, the probability distribution for data calls can, therefore, be approximated by $p_\mathrm{d}(k) \approx (1/G_3)(\rho^k_\mathrm{d}/k!)$, where $\rho_\mathrm{d} = (\lambda_\mathrm{d}+h_\mathrm{d})/\mu^X_\mathrm{d}$ and $G_3 = \sum_{l=0}^{M_\mathrm{d}} \rho^l_\mathrm{d}/l!$. With this approximate closed-form expression for $p_\mathrm{d}(k)$, we have

$$\begin{aligned}\Pi_\mathrm{od} &\approx \frac{1}{G_2G_3}\sum_{i=0}^{M_\mathrm{v}}\sum_{j=0}^{i} p_\mathrm{vv}(i,j)\sum_{A(j)<k\le M_\mathrm{d}} \frac{k\rho^k_\mathrm{d}}{k!}\\ &= \frac{1}{G_2G_3}\sum_{i=0}^{M_\mathrm{v}}\sum_{j=0}^{i} p_\mathrm{vv}(i,j)\cdot\rho_\mathrm{d}\sum_{A(j)-1<k\le M_\mathrm{d}-1} \frac{\rho^k_\mathrm{d}}{k!}. \quad \text{(A6)}\end{aligned}$$

Consider a discrete Poisson random variable W with mean (and variance) equal to ρ_d. The probability mass function of W at k can be approximated by a Gaussian density function at $s = k$, i.e., $\exp(-\rho_\mathrm{d})\rho^k_\mathrm{d}/k! \approx (1/\sqrt{2\pi\rho_\mathrm{d}})\exp[-(1/2)((s-\rho_\mathrm{d})^2/\rho_\mathrm{d})]\Delta s$. The accuracy of the approximation improves as ρ_d becomes large [21]. From this, we have

$$\begin{aligned}\Pi_\mathrm{od} &\approx \rho_\mathrm{d}\frac{\exp(\rho_\mathrm{d})}{G_2G_3}\sum_{i=0}^{M_\mathrm{v}}\sum_{j=0}^{i} p_\mathrm{vv}(i,j)\\ &\quad\times\int_{A(j)-1+0.5}^{M_\mathrm{d}-1+0.5}\frac{1}{\sqrt{2\pi\rho_\mathrm{d}}}\exp\left[-\frac{1}{2}\frac{(s-\rho_\mathrm{d})^2}{\rho_\mathrm{d}}\right]ds\\ &\approx \sum_{i=0}^{M_\mathrm{v}}\sum_{j=0}^{i} p_\mathrm{vv}(i,j)\left[Q\left(\frac{A(j)-0.5-\rho_\mathrm{d}}{\sqrt{\rho_\mathrm{d}}}\right)\right.\\ &\quad\left.-\,Q\left(\frac{M_\mathrm{d}-0.5-\rho_\mathrm{d}}{\sqrt{\rho_\mathrm{d}}}\right)\right] \quad \text{(A7)}\end{aligned}$$

where we have used the following approximation:

$$\begin{aligned}\frac{\exp(\rho_\mathrm{d})}{G_2G_3} &= \frac{\exp(\rho_\mathrm{d})}{G_3\sum_{k=0}^{M_\mathrm{d}}\frac{1}{G_3}k\rho^k_\mathrm{d}/k!}\\ &\approx \frac{\exp(\rho_\mathrm{d})}{\exp(\rho_\mathrm{d})\int_0^\infty\frac{s}{\sqrt{2\pi\rho_\mathrm{d}}}\exp\left[-\frac{1}{2}\frac{(s-\rho_\mathrm{d})^2}{\rho_\mathrm{d}}\right]ds}\\ &\approx \left\{\sqrt{\frac{\rho_\mathrm{d}}{2\pi}}\exp\left(-\frac{\rho_\mathrm{d}}{2}\right)\right.\\ &\quad\left.+\,\rho_\mathrm{d}\left[1-\frac{1}{\sqrt{2\pi\rho_\mathrm{d}}}\exp\left(-\frac{\rho_\mathrm{d}}{2}\right)\right]\right\}^{-1} = \rho^{-1}_\mathrm{d}.\end{aligned}$$

Let

$$\begin{aligned}f_C(C) = Q_\mathrm{od} - \sum_{i=0}^{M_\mathrm{v}}\sum_{j=0}^{i} p_\mathrm{vv}(i,j)&\left[Q\left(\frac{A(j)-0.5-\rho_\mathrm{d}}{\sqrt{\rho_\mathrm{d}}}\right)\right.\\ &\left.-\,Q\left(\frac{M_\mathrm{d}-0.5-\rho_\mathrm{d}}{\sqrt{\rho_\mathrm{d}}}\right)\right].\end{aligned}$$

The equation $f_C(C) = 0$ can then be solved numerically to obtain an approximate value for C. An initial estimate for C can also be obtained from (A7). First, we realize that the most significant terms in the double summation occur when $A(j)$ is smallest, or when $j\cdot\gamma_\mathrm{v}(j)$ is largest (which is Γ). Let $A_\Gamma = \lfloor(C-\Gamma)/c_\mathrm{d}\rfloor$ and $G_4 = \sum_{i=0}^{M_\mathrm{v}}\sum_{\{j:0\le j\le i, j\gamma_\mathrm{v}(j)\ge\Gamma\}} p_\mathrm{vv}(i,j)$. We then proceed as follows: 1) retain the most significant terms of the double summation; 2) use the approximation $Q(x) \approx (1/\sqrt{2\pi}x)\exp(-x^2/2)$ for the Q-function involving A_Γ; 3) ignore the correction factors 0.5; 4) replace Π_od with Q_od; and 5) move all the terms not involving A_Γ to the LHS and let this equal I_C. We can then transform both sides of the equation into

$$\begin{aligned}I_C &= \text{LHS}\\ &= \frac{1}{G_4}\sqrt{\frac{2\pi}{\rho_\mathrm{d}}}\left[Q_\mathrm{od} + \sum_{i=0}^{M_\mathrm{v}}\sum_{\{j:0\le j\le i, j\gamma_\mathrm{v}\ge\Gamma\}} p_\mathrm{vv}(i,j)\right.\\ &\quad\left.\times\,Q\left(\frac{M_\mathrm{d}-\rho_\mathrm{d}}{\sqrt{\rho_\mathrm{d}}}\right)\right] \quad \text{(A8)}\end{aligned}$$

$$I_C \approx \text{RHS} = \frac{1}{A_\Gamma-\rho_\mathrm{d}}\exp\left[-\frac{1}{2}\left(\frac{A_\Gamma-\rho_\mathrm{d}}{\sqrt{\rho_\mathrm{d}}}\right)^2\right]. \quad \text{(A9)}$$

Let C^* be the initial estimate for C. From (A9), we can arrive at the expression for C^* as

$$C^* \approx \begin{cases}\left(I_C^{-1}+\rho_\mathrm{d}\right)c_\mathrm{d}+\Gamma, & I_C\ge 1\\ \left[\sqrt{\rho_\mathrm{d}}\sqrt{-2\ln I_C}+\rho_\mathrm{d}\right]c_\mathrm{d}+\Gamma, & I_C<1\end{cases} \quad \text{(A10)}$$

where I_C can be obtained from (A8).

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions which improved the presentation of this paper.

REFERENCES

[1] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, June 1993.

[2] A. W. Berger and W. Whitt, "Effective bandwidth with priorities," *IEEE/ACM Trans. Networking*, vol. 6, pp. 447–460, Aug. 1998.

[3] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. New York: Springer-Verlag, 1995.

[4] S. C. Borst and D. Mitra, "Virtual partitioning for robust resource sharing: Computational techniques for heterogeneous traffic," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 668–678, June 1998.

[5] M. Schwartz, *Broadband Integrated Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[6] M. Naghshineh and A. Acampora, "QoS provisioning in micro-cellular networks supporting multiple classes of traffic," *Wireless Networks*, vol. 2, no. 3, pp. 195–203, Aug. 1996.

[7] C. Wu, Y. Tsai, and J. Chang, "A quality-based birth-and-death queueing model for evaluating the performance of an integrated voice/data CDMA cellular system," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 83–89, Jan. 1999.

[8] T. Liu and J. Silvester, "Joint admission/congestion control for wireless CDMA systems supporting integrated services," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 845–857, Aug. 1998.

[9] W. Yang and E. Geraniotis, "Admission policies for integrated voice and data traffic in CDMA packet radio networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 654–664, May 1994.

[10] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," *Wireless Networks*, vol. 3, no. 1, pp. 29–41, Mar. 1997.

[11] Y. Fang, I. Chlamtac, and Y. Lin, "Channel occupancy times and handoff rate for mobile computing and PCS networks," *IEEE Trans. Comput.*, vol. 47, pp. 679–692, June 1998.

[12] Y.-B. Lin and W. Chen, "Call request buffering in a PCS network," in *Proc. IEEE INFOCOM*, June 1994, pp. 585–592.

[13] C. Jedrzycki and V.C.M. Leung, "Probability distribution of channel holding time in cellular telephone systems," in *Proc. IEEE Vehicular Technology Conf.*, Apr.–May 1996, pp. 247–251.

[14] E. Altman and V. A. Gaitsogory, "Stability and singular perturbations in constrained Markov decision problems," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 971–975, June 1993.

[15] D. Mitra, M. I. Reiman, and J. Wang, "Robust dynamic admission control for unified cell and call QoS in statistical multiplexing," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 692–707, June 1998.

[16] M. Cheung and J. W. Mark, "Resource allocation in wireless networks based on joint packet/call levels QoS constraints," in *Proc. IEEE GLOBECOM*, Nov. 2000, pp. 271–275.

[17] A. B. Downey, "The structural cause of file size distributions," *ACM SIGMETRICS Performance Eval. Rev.*, vol. 29, pp. 328–329, June 2001.

[18] W. S. Jeon and D. G. Jeong, "Call admission control for mobile multimedia communications with traffic asymmetry between uplink and downlink," *IEEE J. Select. Areas Commun.*, vol. 50, pp. 59–66, Jan. 2001.

[19] M. May, J.-C. Bolot, A. Jean-Marie, and C. Diot, "Simple performance models of differentiated services schemes for the Internet," in *Proc. IEEE INFOCOM*, vol. 3, Mar. 1999, pp. 1385–1394.

[20] E. Chan and X. Hong, "Analytical model for an assured forwarding differentiated service over wireless links," *Proc. IEE Commun.*, vol. 148, no. 1, pp. 19–23, Feb. 2001.

[21] I. F. Blake, *An Introduction to Applied Probability*. Melbourne, FL: Krieger, 1987.

Chi Wa Leong received the B.Sc. degree in computer engineering in 1995 and the M.Sc. degree in electrical engineering in 1998, both from the University of Manitoba, Winnipeg, MB, Canada. He also received the M.A.Sc. degree in electrical engineering in 2001 from the University of Waterloo, Waterloo, ON, Canada.

He was a Research Assistant at the University of Waterloo, where he carried out research on call admission control for mobile networks.

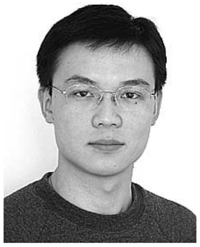
Weihua Zhuang (M'93–SM'01) received the B.Sc. and M.Sc. degrees in 1982 and 1985, respectively, from Dalian Maritime University, Liaoning, China, and the Ph.D. degree in 1993 from the University of New Brunswick, Fredericton, NB, Canada, all in electrical engineering.

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 coauthor 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 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 the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

Yu Cheng 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.

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



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.