Stochastic Admission Control for Quality of Service in Wireless Packet Networks

Majid Ghaderi[†] and Raouf Boutaba
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1, Canada
{mghaderi,rboutaba}@uwaterloo.ca

Gary W. Kenward
Nortel Networks
PO Box 3511 Station C
Ottawa, ON K1Y 4H7, Canada
gkenward@nortelnetworks.com

Abstract

Call admission control is a key element for providing quality of service (QoS) in mobile wireless networks. Traditional admission control schemes only address call-level QoS guarantee because of the underlying circuit-based network architecture. In contrast, emerging wireless technologies such as 3G and 4G tend to be packet-switched rather than circuit-switched because the packet-based architecture allows better sharing of the scarce wireless resources. This paper introduces a novel distributed call admission control scheme called PFG, which maximizes the wireless channel utilization subject to a predetermined bound on the call dropping and packet loss probabilities for variable-bit-rate traffic in a packet-switched wireless cellular network. In particular, we show that in wireless packet networks, the undesired event of dropping an ongoing call can be completely eliminated without sacrificing the bandwidth utilization. The proposed control algorithm is stochastic and dynamic, hence it is able to adapt to a wide range of traffic fluctuations and mobility patterns. Extensive simulation results show that our scheme satisfies the hard constraint on call dropping and packet loss probabilities while maintaining a high bandwidth utilization.

Index Terms

Call admission control, wireless packet networks, quality of service.

[†]corresponding author

Stochastic Admission Control for Quality of Service in Wireless Packet Networks

I. Introduction

Emerging wireless technologies such as 3G and 4G [1], [2] tend to be packet-switched rather than circuit-switched because the packet-based architecture allows better sharing of limited wireless resources. In a packet network, calls do not require dedicated circuits for the entire duration of connection. Unfortunately, this enhanced flexibility makes it more difficult to effectively control the admission of connections into the network [3], [4].

In wireless packet networks there exist two levels of quality of service, namely, call-level and packet-level [5]. At call-level, two important parameters which determine the quality of service are *call blocking probability* and *call dropping probability*. Since dropping a call in progress has more negative impact from the user perspective, handoff calls are given higher priority than new calls in access to wireless resources. This preferential treatment of handoffs increases the blocking of new calls and hence degrades the utilization of wireless bandwidth. At packet-level, *packet loss probability*, delay and jitter are the most important QoS parameters. There is always a trade-off between the network utilization and the QoS perceived by users. It is desired to have a resource allocation scheme which can satisfy the prespecified QoS constraints while maximizing the utilization of the network resources.

Most of the researchers in wireless networking field have focused only on call-level quality of service parameters for admission control and resource allocation [5]–[17] because the primary concern has been voice traffic support in a circuit-switched wireless network. Therefore, there is no packet transmission and consequently no packet-level quality of service. Our approach consists of taking into consideration a combination of both call-level and packet-level QoS parameters in making the admission decision. The main idea is to model the bandwidth requirement in each cell based on two factors: 1) mobility patterns of users, and 2) packet generation characteristics of individual calls. Based on this model, the time-dependent packet loss probability is calculated and used to find the appropriate acceptance ratio for new calls requesting access to the network. The motivation behind this study is to support variable-bit-rate (VBR) multimedia traffic in emerging wireless packet cellular networks.

The rest of the paper is organized as follows. Section II reviews the related work in the area. Our system model, assumptions and notations are described in section III. Section IV describes the high-level operation of the proposed admission control algorithm while in section V, detailed analysis of the algorithm is presented. Some simulation results are presented in section VI and finally, section VII concludes this paper.

II. Background and Motivation

Call admission control (CAC) has been extensively studied in circuit-switched (voice) wireless cellular networks (see [5], [11], [15] and references there in). Hong and Rappaport [6] are the first who systematically analyzed the famous *guard channel* (GC) scheme, which is currently deployed in cellular networks supporting voice calls. Ramjee et al. [7] have formally defined and categorized the admission control problem in cellular networks. They showed that the guard channel scheme is optimal for minimizing a linear objective function of call blocking and dropping probabilities while the *fractional guard channel* scheme (FG) is optimal for minimizing call blocking probability subject to a hard constraint on call dropping probability. Instead of explicit bandwidth reservation as in the GC, the FG accepts new calls according to a randomization parameter called the *acceptance ratio*.

Because of user mobility, it is impossible to describe the state of the system by using only local information, unless we assume that the network is uniform and approximate the overall state of the system by the state of a single cell in isolation. To include the global effect of mobility, collaborative or distributed admission control schemes have been proposed [8]–[10], [12], [16]. Information exchange among a cluster of neighboring cells is the approach adopted by all distributed schemes.

In particular, Naghshineh and Schwartz [8] proposed a collaborative admission control known as distributed call admission control (DCA). DCA periodically gathers some information, namely the number of active calls, from the adjacent cells to make, in combination with the local information, the admission decision. It has been shown that DCA is not stable and violates the required dropping probability as the load increases [16]. Levin et al. [9] proposed a more sophisticated version of the original DCA based on the shadow cluster concept, which uses dynamic clusters for each user based on its mobility pattern instead of restricting itself (as

DCA) to direct neighbors only. A practical limitation of the shadow cluster scheme in addition to its complexity and inherent overhead is that it requires a precise knowledge of the mobile's trajectory. Recently, Wu et al. [16] proposed a stable distributed scheme (SDCA) based on the classical fractional guard channel scheme which can precisely achieve the target call dropping probability. A key feature of SDCA is the formulation of the time-dependent call dropping probability which can be computed by the diffusion approximation of the channel occupancy.

None of these papers has considered a wireless packet-switched network. There is no packet-level consideration in these works. Call dropping and blocking probabilities are the only QoS parameters considered. In circuit-switched networks, when a handoff call arrives while there is no idle circuit (wireless channel), the handoff fails and hence the call is dropped. In contrast, in a packet-switched network it is still possible to accept the handoff call at the expense of probably increasing the number of dropped packets. While this approach completely eliminates the call dropping event, we will show that its impact on packet loss can be effectively controlled.

We introduce a *packetized fractional guard channel* (PFG) call admission control mechanism for cellular packet networks that achieves a high bandwidth utilization while satisfying a target packet loss probability without dropping any ongoing call. In particular, we consider a packetized version of the so-called MINB problem [7] in packet-switched cellular networks. We define the packetized MINB as follow:

> *for a given cell capacity, maximize the bandwidth utilization while achieving zero percent call dropping probability subject to a hard constraint on the packet loss probability.*

To the best of our knowledge, PFG is the first to address the packetized MINB problem. The main features of PFG are as follows:

1) PFG achieves zero percent call dropping.
2) PFG is dynamic, therefore, adapts to a wide range of system parameters and traffic conditions.
3) PFG is distributed and takes into consideration the information from direct neighboring cells in making admission decisions.
4) The control mechanism is stochastic and periodic to reduce the overhead associated with distributed control schemes.

III. System Model

A packet-switched cellular network is considered in this paper. We assume that there is one type of calls in the system. As mentioned earlier, in the system under consideration, no handoff call is dropped instead overflow packets are dropped. The considered system is bufferless, hence packet delay and jitter due to packet queueing process are zero. Therefore, packet loss probability is the only packet-level QoS parameter. Furthermore, we assume that cell overflow (receiving more traffic than what can be actually transmitted) is the only source of packet loss, i.e. no packet loss due to wireless channel effects.

Below is the notation which will be used throughout this paper.

- $\mathcal{B}$: total number of cells in the network
- $\mathcal{A}_i$: the set of adjacent cells of cell i
- c_i: capacity of cell i, which is equal to the packet transmission rate of base station i
- $R_i(t)$: packet arrival rate at time t in cell i
- $L_i(t)$: packet loss probability at time t in cell i
- $N_i(t)$: number of active calls at time t in cell i
- a_i: new call acceptance ratio in cell i
- λ_i: new call arrival rate into cell i
- $1/\mu$: mean call duration
- $1/h$: mean cell residency time
- T: length of the control period
- P_L: target packet loss probability
- r_{ji}: routing probability from cell $j \in \mathcal{A}_i$ to cell i
- P_{ji}: handoff probability from cell $j \in \mathcal{A}_i$ to cell i
- $E[z]$: the mean of random variable z
- $V[z]$: the variance of random variable z
- $\tilde{z}$: the time-averaged value of random variable z

The considered system is not required to be uniform. Each cell can experience a different load, e.g. some cells can be over-utilized while others are under-utilized. Also cells may have different capacities. Moreover, we consider that

1) The cell capacity is fixed over time. However, the approach that we propose next can be

extended to include cases in which c_i varies over time.

2) New call arrivals to a cell are independent and Poisson distributed.
3) Cell residency times are independent and exponentially distributed. However, we show that the proposed algorithm is insensitive to this assumption.
4) Call durations are independent and exponentially distributed.

The exponential call durations and cell residency times are widely used in literature [5]–[8], [10]–[13], [16]. In the real word, the cell residence time distribution may not be exponential but exponential distributions provide the mean value analysis, which indicates the performance trend of the system. Furthermore, our proposed admission control algorithm involves a periodic control where the length of the control period is set to much less than the average cell residency time of a call to make the algorithm insensitive to this assumption.

A. Maximum Occupancy in a Cell

Let M_i denote the maximum occupancy, i.e. maximum number of calls, in cell i under the so-called *average bandwidth assignment* scheme. This scheme allocates to each VBR call a share of bandwidth equal to the call's average bandwidth requirement. Let m denote the average bandwidth requirement of a call, then

$$M_i = \frac{c_i}{m}\,. \tag{1}$$

Although this scheme achieves a high bandwidth utilization, it leads to a high rate of packet loss [18]. If there are more than M_i calls in cell i, then we say that the cell is in *overloaded state*. In the overloaded state, probability of packet loss is very high. Our scheme, PFG, rejects new call requests when a cell is in overloaded state.

B. Multiple Handoffs Probability

As mentioned earlier, in order to make the optimal admission decision, distributed schemes regularly exchange some information with other cells in the network. Those cells involved in the information exchange form a *cluster*. Due to the intercell information exchange, base station interconnection network incurs a high signalling overhead. Moreover, as the cluster size increases the operational complexity of the control algorithm increases too. In particular, two major factors affect the overhead and complexity of distributed CAC schemes; (1) frequency of information exchange, and, (2) depth of information exchange, i.e. how many cells away information is exchanged.

To reduce the overhead, distributed CAC schemes typically have a periodic structure in which only at the beginning of control periods information exchange is triggered. Moreover, information exchange is typically restricted to a cluster of neighboring cells. Note that, if the control interval is too small then frequent communications increases the signalling overhead. On the other hand, if the control period is too long then the state information stored locally may become stale. Similarly, if the cluster is too small then the exchanged information will poorly reflects the state of the network. On the other hand, a big cluster will lead to higher overhead. An efficient CAC scheme must compromise between the frequency and depth of information exchange.

In this paper, we set the control interval in such a way that the probability of having multiple handoffs in one control period becomes negligible. Therefore, we can effectively assume that only those cells directly connected to a cell can influence the number of calls in that cell during a control period. In a sense, we reduce the control interval in favor of a smaller cluster size. We claim that using this technique, the signalling overhead will not increase, while the collected information on the network status will be sufficiently accurate for the purpose of a stochastic admission control. The reason is that: first, by decreasing the control interval, the probability of multiple handoffs decays to zero exponentially (see section V-E); second, a cluster shrinks quadratically with decreasing the depth of information exchange (see below).

Without loss of generality, consider a symmetric network where each cell has exactly $\mathcal{A}$ neighbors. Consider cell i and all the cells around it forming circular layers as shown in Fig. 1. From cell i, all the cells up to layer n are accessible with n handoffs assuming that cell i forms layer 0. The number of cells reachable by n handoffs from cell i denoted by $M(n)$ is given by

$$\begin{aligned} M(n) &= 1 + \mathcal{A} + \cdots + n\mathcal{A} \\ &= 1 + \frac{1}{2}n(n+1)\mathcal{A}\,. \end{aligned} \tag{2}$$

Therefore, by slightly reducing the control interval, we essentially achieve the same control accuracy but with reduced signalling overhead. The problem of choosing the proper control interval will be further addressed in section V-E.

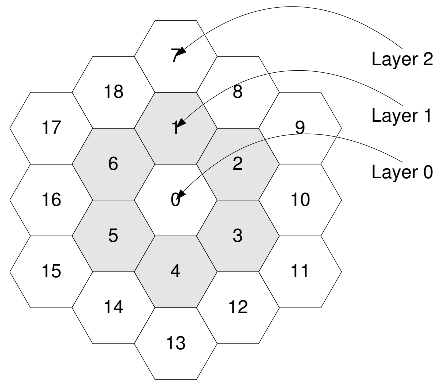


Fig. 1. A cellular system with 3 layers.

C. Time-Dependent Handoff Probability

Let random variables t_d and t_r denote the call duration and cell residency time of a typical call, respectively. We compute here some useful probabilities required for the rest of our discussion.

Let $P_h(t)$ denote the probability that a call hands off to another cell by time t and remains active until t, given that it has been active at time 0. Also, let $P_s(t)$ denote the probability that a call remains active in its home cell until time t, given that it has been active at time 0. Then,

$$\begin{aligned} P_h(t) &= \Pr(t_r \le t)\ \Pr(t_d > t) \\ &= (1 - e^{-ht})\, e^{-\mu t}, \end{aligned} \tag{3}$$

and,

$$\begin{aligned} P_s(t) &= \Pr(t_r > t)\ \Pr(t_d > t) \\ &= e^{-(\mu+h)t}\,. \end{aligned} \tag{4}$$

These equations are valid as far as the memoryless property of call duration and cell residency is satisfied. On average, for any call which arrives at time $t' \in (0, t]$, the average handoff and stay probabilities $\tilde{P}_h$ and $\tilde{P}_s$ are expressed as

$$\tilde{P}_h(t) = \frac{1}{t}\int_0^t P_h(t - t')\, dt', \tag{5}$$

$$\tilde{P}_s(t) = \frac{1}{t}\int_0^t P_s(t - t')\, dt'\,. \tag{6}$$

These integrals can be easily computed with respect to (3) and (4).

Similar to [8] and [10], we assume that during a control period each call experiences at most one handoff. This assumption is justified by setting the length of the control period T reasonably shorter than the average cell residency time. Since cell residency is exponentially distributed, the number of cell crossings (handoff events) that an active call experiences during its lifetime has a Poisson distribution with the mean rate h. Therefore, Th is the expected number of handoffs during an interval of length T for an arbitrary call. By setting $T \ll 1/h$, we make sure that the probability of having more than one handoff during the interval of length T is negligible. This will be further addressed in section V-E.

Finally, let $P_{ji}(t)$ denote the time-dependent handoff probability that an active call in cell j at time 0 will be in cell i at time t, where $j \in \mathcal{A}_i$. Since each call experiences at most one handoff during the control period, it is obtained that

$$P_{ji}(t) = P_h(t)\, r_{ji}\, . \tag{7}$$

Similarly, the average handoff probability $\tilde{P}_{ji}(t)$ for a call which arrives at any time $t' \in (0, t]$ is given by

$$\tilde{P}_{ji}(t) = \tilde{P}_h(t)\, r_{ji}\, . \tag{8}$$

In next section, we will use the computed probabilities to find the maximum acceptance ratio for a given cell i with respect to the prespecified packet loss probability P_L.

IV. Admission Control Algorithm

The proposed distributed algorithm, PFG, consists of two components. The first component is responsible for retrieving the required information from the neighboring cells and computing the acceptance ratio. Since each call experiences at most one handoff during the control period, the immediate neighbors of cell i, i.e. $\mathcal{A}_i$, are the ones that will affect the number of calls, and consequently the packet arrival process in cell i during a control period. Hence, in our distributed admission control algorithm, information exchange is limited to direct neighboring cells. Using the computed acceptance ratio, the second component enforces the admission control locally in each cell. The following sections describe these two components in detail.

A. Distributed Control Algorithm

As mentioned earlier, to reduce the signalling overhead, PFG has a periodic structure. All the information exchange and acceptance ratio computations happen only once at the beginning of

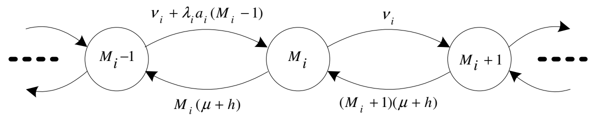


Fig. 2. Packetized fractional guard channel transition diagram.

each control period of length T. Several steps involved in PFG distributed control are described below:

1) At the beginning of a control period, each cell i sends the following information to its adjacent cells:
   a) the number of active calls in the cell at the beginning of the control period denoted by $N_i(0)$.
   b) the number of new calls, N_i, which were admitted in the last control period.
2) Each cell i receives $N_j(0)$ and N_j from every adjacent cell $j \in \mathcal{A}_i$.
3) Now, cell i uses the received information and those available locally to compute the acceptance ratio a_i using the technique described in section V.
4) Finally, the computed acceptance ratio a_i is used to admit call requests into cell i using the algorithm presented in section IV-B.

B. Local Control Algorithm

In PFG, handoffs are always accepted with probability 1 even when the destination cell is overloaded. In this situation, accepting incoming handoffs may increase the number of dropped packets. There is no buffer in the system and packets are served/dropped according to a FIFO scheduling.

Let s_i denote the state of cell i, where there are s calls active in the cell. Let $a_i(s)$ denote the acceptance ratio where the cell state is s_i. Fig. 2 shows the state transition diagram of the PFG scheme in cell i. In this diagram, ν_i is the handoff arrival rate into cell i, and M_i is the maximum occupancy given by (1).

For an accurate control, the call blocking probability in each period is given by complementing the acceptance ratio. Therefore, by averaging acceptance ratios over a number of control periods,

```
if (x is a handoff call) then
   accept call;
else /* x is a new call */
   if (rand(0,1) < a_i) & (N_i(t) ≤ M_i) then
      accept call;
   else
      reject call;
   end if
end if
```

Fig. 3. Local call admission control algorithm in cell i.

the call blocking probability in cell i denoted by p_{b_i} is expressed as

$$p_{b_i} = 1 - \tilde{a}_i \, . \tag{9}$$

Consequently, the average network-wide call blocking probability for the considered network is given by

$$p_b = \frac{\sum_{j=i}^{\mathcal{B}} \lambda_i p_{b_i}}{\sum_{j=i}^{\mathcal{B}} \lambda_i} \, . \tag{10}$$

The pseudo-code for the local admission control in cell i is given by the algorithm in Fig. 3. In this algorithm, x is a call requesting a connection into cell i. The acceptance ratio for the respective control period is a_i. Also, $\texttt{rand}(0,1)$ is a uniform random generator function. In the next section, we will present a technique to compute the acceptance ratio a_i in order to complete this algorithm.

V. Computing the Acceptance Ratio

Assuming the target loss probability is sufficiently small, we approximate the packet loss probability by the overflow probability in each cell. Similar approximation is used for computing the effective bandwidth of a call in [4] and [19]. In particular, it is shown in [18] that, for a given network configuration, the overflow probability is superior to the packet loss probability. However, as the overflow probability decays to zero, both measures converge to the same value and the difference becomes negligible.

The approximated packet loss probability here can be interpreted as a tight upper bound on the actual packet loss probability. Therefore, the time-dependent packet loss probability at time t in cell i is given by

$$L_i(t) = \Pr\big(R_i(t) > c_i\big), \tag{11}$$

where $R_i(t)$ denotes the total (new and handoff) packet arrival rate into cell i at time t.

The proposed approach for computing the acceptance ratio includes the following steps:

1) Each cell i uses the information received from its adjacents and the information available locally to find the time-dependent mean and variance of the number of calls in the cell using (22) and (25), respectively.
2) The computed mean and variance of the number of calls is used to find the mean and variance of the packet arrival process in the cell using (16) and (17), respectively.
3) Having the mean and variance of the packet arrival process, the time-dependent packet arrival process is approximated by a Gaussian distribution.
4) The tail of this Gaussian distribution is used to find the time-dependent packet loss probability in each cell i.
5) Time-dependent packet loss probability is averaged over a control interval of length T to find an average packet loss probability as expressed by (31).
6) Using the computed packet loss probability and the prespecified QoS constraint, i.e. $\tilde{L}_i \leq P_L$, acceptance ratio a_i is computed from (32).

The motivation behind Gaussian traffic characterization is that it is very natural when a large number of sources are multiplexed (motivated by the central limit theorem [20]), as is expected to be the case in future wireless networks. It is expected that future wireless technologies such as 3G and 4G increase the available cell capacities to several Mbps [1], [2]. In such networks, the number of active calls (and consequently, the number of packets being transmitted) is so high that the central limit theorem can be successfully applied to model the packet arrival process in each cell.

In fact, it has been observed that the aggregation of even a fairly small number of traffic streams is usually sufficient for the Gaussian characterization of the input process [19], [21]. Further, Gaussian processes are completely specified by their first two moments. This makes Gaussian traffic characterization ideal from a measurement point of view, since measuring statistics beyond the second moment is usually impractical. Also, Gaussian processes can have an arbitrary correlation structure and this includes self-similar processes [22], [23] as well.

In following subsections we show the derivation of packet loss probability which is used to find the acceptance ratio.

A. Traffic Characterization

Let r_n denote the packet generating process of an individual call n. It is assumed that individual packet generating processes are independent and identically distributed (iid) random variables with the mean and variance $E[r]$ and $V[r]$, respectively. Then, $R_i(t)$, the total packet arrival rate in cell i at time t, is expressed as the summation of packet generating process of individual calls. That is

$$R_i(t) = \sum_{n=1}^{N_i(t)} r_n \,. \tag{12}$$

where $N_i(t)$ denotes the number of calls at time t. Our objective is to apply the central limit theorem to approximate $R_i(t)$ by a Gaussian distribution. At first we have to specify the parameters of $R_i(t)$, namely, mean and variance.

Let Φ_r denote the moment generating function of r_n, i.e. $\Phi_r(\theta) = E[e^{\theta r_n}]$. Also, let Φ_R denote the moment generating function of $R_i(t)$, then $\Phi_R(\theta)$ can be found as follows:

$$\begin{aligned} \Phi_R(\theta)_{|N_i(t)=N} &= E[e^{\theta R_i(t)} | N_i(t) = N] \\ &= E[e^{\theta \sum_{n=1}^{N_i(t)} r_n} | N_i(t) = N] \\ &= E[e^{\theta \sum_{n=1}^{N} r_n}] \\ &= \{\Phi_r(\theta)\}^N, \end{aligned} \tag{13}$$

therefore,

$$E[e^{\theta R_i(t)} | N_i(t)] = \{\Phi_r(\theta)\}^{N_i(t)}, \tag{14}$$

and,

$$E[e^{\theta R_i(t)}] = E[\{\Phi_r(\theta)\}^{N_i(t)}] \,. \tag{15}$$

Using these equations, it is obtained that

$$E[R_i(t)] = E[N_i(t)]E[r], \tag{16}$$

$$V[R_i(t)] = E[N_i(t)]V[r] + V[N_i(t)]E^2[r] \,. \tag{17}$$

As expected, the variance of the total packet arrival rate is a function of both variance of individual call packet generating process and the variance of the number of calls at time t. This indicates that static treatment of the number of calls in a cell, i.e. assuming that there is $E[N_i(t)]$ calls in a cell, is not accurate and must be avoided.

Therefore, in order to compute $E[R_i(t)]$ and $V[R_i(t)]$ we have to first have $E[r]$ and $V[r]$. In this paper, we assume that $E[r]$ and $V[r]$ are known to the admission controller a priori. This is a minimal set of requirements since it does not assume anything specific about the actual packet generating process of the individual calls. Two cases can happen in practice:

1) The traffic generation process of individual calls can be described by means of an analytical model. In this case, $E[r]$ and $V[r]$ are simply computed using probabilistic techniques [20].
2) The traffic generation process of individual calls can not be described by means of an analytical model. In this case, $E[r]$ and $V[r]$ are simply measured from real traffic data [21].

In next section, we compute $E[N_i(t)]$ and $V[N_i(t)]$ based on the mobility information available locally and the information obtained from neighboring cells.

B. Mobility Characterization

The number of calls in cell i at time t is affected by two factors: (1) the number of background (existing) calls which are already in cell i or its adjacent cells, and, (2) the number of new calls which will arrive in cell i and its adjacent cells during the period $(0, t]$ ($0 < t \leq T$). Let $g_i(t)$ and $n_i(t)$ denote the number of background and new calls in cell i at time t, respectively.

A background call in cell i will remain in cell i with probability $P_s(t)$ or will handoff to an adjacent cell j with probability $P_{ij}(t)$. A new call which is admitted in cell i at time $t' \in (0, t]$ will stay in cell i with probability $\tilde{P}_s(t)$ or will handoff to an adjacent cell j with probability $\tilde{P}_{ij}(t)$. Therefore, the number of background calls which remain in cell i and the number of handoff calls which come into cell i during the interval $(0, t]$ are binomially distributed. For a binomial distribution with parameter q, the variance is given by $q(1-q)$. Using this property it is obtained that

$$V_s(t) = P_s(t)\,(1 - P_s(t)), \tag{18}$$

$$V_{ji}(t) = P_{ji}(t)\,(1 - P_{ji}(t)), \tag{19}$$

$$\tilde{V}_s(t) = \tilde{P}_s(t)\,(1 - \tilde{P}_s(t)), \tag{20}$$

$$\tilde{V}_{ji}(t) = \tilde{P}_{ji}(t)\,(1 - \tilde{P}_{ji}(t)), \tag{21}$$

where, $V_s(t)$ and $V_{ji}(t)$ denote the time-dependent variance of stay and handoff processes, and, $\tilde{V}_s(t)$ and $\tilde{V}_{ji}(t)$ are their average counterparts, respectively.

The number of calls in cell i is the summation of the number of background calls, $g_i(t)$, and new calls, $n_i(t)$. Therefore, the mean number of active calls in cell i at time t is given by

$$E[N_i(t)] = E[g_i(t)] + E[n_i(t)], \tag{22}$$

where,

$$E[g_i(t)] = N_i(0)P_s(t) + \sum_{j\in\mathcal{A}_i} N_j(0)P_{ji}(t), \tag{23}$$

$$E[n_i(t)] = (a_i\lambda_i t)\tilde{P}_s(t) + \sum_{j\in\mathcal{A}_i} (a_j\lambda_j t)\tilde{P}_{ji}(t)\,. \tag{24}$$

Similarly the variance is given by

$$V[N_i(t)] = V[g_i(t)] + V[n_i(t)], \tag{25}$$

where,

$$V[g_i(t)] = N_i(0)V_s(t) + \sum_{j\in\mathcal{A}_i} N_j(0)V_{ji}(t), \tag{26}$$

$$V[n_i(t)] = (a_i\lambda_i t)\tilde{V}_s(t) + \sum_{j\in\mathcal{A}_i} (a_j\lambda_j t)\tilde{V}_{ji}(t)\,. \tag{27}$$

Note that given the arrival rate λ_i and the acceptance ratio a_i, the *actual new call arrival rate* into cell i is given by $\lambda_i a_i$. Therefore, the expected number of call arrivals during the interval $(0, t]$ is given by $a_i\lambda_i t$.

C. Packet Loss Probability

As mentioned earlier, the packet arrival distribution in each cell can be approximated by a Gaussian distribution:

$$R_i(t) \sim \mathbf{G}\Big(E[R_i(t)],\ V[R_i(t)]\Big), \tag{28}$$

where, $E[R_i(t)]$ and $V[R_i(t)]$ are given by (16) and (17), respectively.

Hence, the original admission control problem is reduced to maintaining the packet arrival rate below the available capacity c_i with probability $1-P_L$ at any point in time $t \in (0, T]$. Using (11) and (28) it is obtained that

$$L_i(t) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{c_i - E[R_i(t)]}{\sqrt{2\,V[R_i(t)]}}\right), \tag{29}$$

where $\mathrm{erfc}(c)$ is the complementary error function defined as

$$\mathrm{erfc}(c) = \frac{2}{\sqrt{\pi}} \int_c^{\infty} e^{-t^2}\, dt\,. \tag{30}$$

Using (29), the average packet loss probability over a control period of length T is given by

$$\tilde{L}_i = \frac{1}{T} \int_0^{T} L_i(t)\, dt\,. \tag{31}$$

Then, the acceptance ratio, a_i, can be found by numerically solving (refer to [24]) equation

$$\tilde{L}_i = P_L\,. \tag{32}$$

The boundary condition is that $a_i \in [0, 1]$, hence if $\tilde{L}_i$ is less than P_L even for $a_i = 1$ then a_i is set to 1. Similarly, if $\tilde{L}_i$ is greater than P_L even for $a_i = 0$, then a_i is set to 0.

D. Actual New Call Arrival Rate

In section V-C, we used products $a_j\lambda_j$ to compute the mean and variance of the number of calls in cell i ($j \in \mathcal{A}_i$). Let us define the *actual new call arrival rate* into cell j, denoted by $\bar{\lambda}_j$, as follows

$$\bar{\lambda}_j = a_j\lambda_j\,. \tag{33}$$

In order to compute a_i for the new control period we need to know $\bar{\lambda}_j$ for every adjacent cell j ($j \in \mathcal{A}_i$). Similarly, cell j needs to know $\bar{\lambda}_i$ in order to be able to compute a_j. Therefore, every cell depends on its adjacents and vice versa. To break this dependency, instead of using the actual value of $\bar{\lambda}_j$, each cell i estimates the actual new call arrival rates of its adjacents for the new control period.

Let $\bar{\lambda}_j(n)$ denote the actual new call arrival rate into cell j during the n-th control period. Also, let $N_j(n)$ denote the number of new calls that were accepted in cell j during the n-th control period. An estimator for $\bar{\lambda}_j$ is expressed as

$$\bar{\lambda}_j(n+1) = (1-\epsilon)\frac{N_j(n)}{T} + \epsilon\bar{\lambda}_j(n), \tag{34}$$

where, $\bar{\lambda}_j(n+1)$ is the actual new call arrival rate into cell j at the beginning of the $(n+1)$-th control period. Note that $\bar{\lambda}_j(n)$ is known at the beginning of the $(n+1)$-th control period. In our simulations we found that $\epsilon = 0.3$ leads to a good estimation of the actual new call arrival rate.

E. Control Interval

The idea behind at-most-one handoff assumption is that by setting control interval appropriately, the undesired multiple handoffs during a control period can be avoided. As discussed in section III-B, this minimizes the signalling overhead and operational complexity of PFG. In this section, we address the control interval selection problem.

Consider a symmetric network where each cell has exactly $\mathcal{A}$ neighbors, and the probability of handoff to every neighbor is the same. Then, the routing probability r_{ij} from cell i to cell j is given by

$$r_{ij} = \begin{cases} 1/\mathcal{A}, & j \in \mathcal{A}_i, \\ 0, & j \notin \mathcal{A}_i \,. \end{cases} \tag{35}$$

Let $q(n)$ denote the probability that an active call experiences n handoffs during time interval T. Also, let $q_{ij}(n)$ denote the probability that a call originally in cell i moves to cell j over a path consisting of n handoffs during time interval T. Define δ as the multiple handoffs probability from cell i to cell j. We then can write

$$\delta = \sum_{n=2}^{\infty} q_{ij}(n) \,. \tag{36}$$

Our goal is to find a relation between T and δ in order to be able to control δ by controlling T.

For an effective control (p_f in the range of 10^{-4} to 10^{-2}) we can assume that p_f is effectively zero. Similarly, if $\delta \approx p_f$ for a given T, we can assume that the multiple handoffs probability is zero. Since cell residency is exponential, the number of handoffs a call experiences during an interval is Poisson distributed with mean hT, given that the call is active during the whole interval. Therefore, it is obtained that

$$q(n) = \frac{(hT)^n}{n!} e^{-(h+\mu)T} \,. \tag{37}$$

In order to compute $q_{ij}(n)$ based on (37), we need to find the probability of moving from cell i to cell j by n handoffs. Let $L_{ij}(n)$ denote the number of paths consisting of n handoffs from i to j, then

$$q_{ij}(n) = \frac{L_{ij}(n)}{\mathcal{A}^n} \, q(n) \,. \tag{38}$$

Consider the network depicted in Fig. 1. Let $T = 20\,s$, $1/\mu = 180\,s$, $1/h = 100\,s$ and $\mathcal{A} = 6$. Table (I) shows the maximum probability of multiple handoffs from any cell j to cell 0, $P_{j0}(n)$,

TABLE I

MULTIPLE HANDOFFS PROBABILITY FOR $T = 20\,s$.

n	Layer	$\max\{L_{j0}(n)\}$	$\max\{P_{j0}(n)\}$
0	0	1	0.73263
1	0	1	0.02442
2	0	6	0.00244
3	1	15	0.00007
4	0	90	0.00000
5	0	360	0.00000

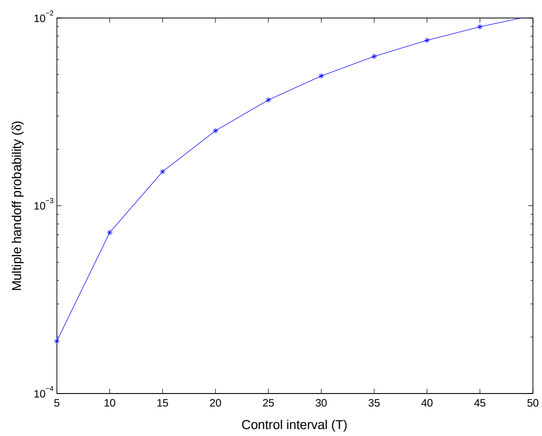


Fig. 4. Effect of T on multiple handoffs probability.

based on the number of handoffs, n. For each n, we have also determined which layer has the maximum paths to cell 0. Interestingly, cell 0 has the most paths to itself through other cells. We have also illustrated in Fig. 4 the impact of the control interval T on the multiple handoffs probability δ for the same set of parameters.

Consider cell i and all the cells around it forming circular layers. From cell i, all the cells up to layer n are accessible with n handoffs assuming that cell i forms layer 0. It can be shown that

$$L_{ij}(n) \leq \mathcal{A}^{n-1}, \quad n \geq 1 \tag{39}$$

because for $n \geq 1$, at each level there are at least $\mathcal{A}$ cells which have the same number of paths

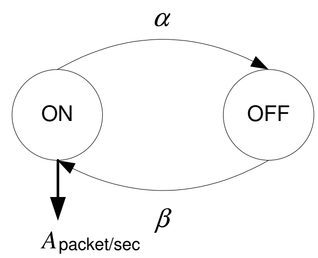


Fig. 5. ON/OFF model.

to the destination cell i. Therefore

$$q_{ij}(n) \leq \frac{1}{\mathcal{A}} \frac{(hT)^n}{n!} e^{-(h+\mu)T}, \quad n \geq 1 \,. \tag{40}$$

Using (36) and (40), it is obtained that

$$\begin{aligned} \delta &\leq \sum_{n=2}^{\infty} \frac{1}{\mathcal{A}} \frac{(hT)^n}{n!} e^{-(h+\mu)T} \\ &= \frac{e^{hT} - hT - 1}{\mathcal{A} e^{(h+\mu)T}} \,. \end{aligned} \tag{41}$$

Using the Taylor expansion of exponential terms for $\delta \ll \frac{1}{\mathcal{A}}(\frac{h}{\mu+h})$, it is obtained that

$$T \leq \frac{\mathcal{A}\delta(\mu+h) + h\sqrt{2\mathcal{A}\delta}}{\mathcal{A}\delta(\mu+h)^2 - h^2}, \tag{42}$$

which finally leads to the following simple relation

$$T \approx \frac{\sqrt{2\mathcal{A}\delta}}{h} \,. \tag{43}$$

VI. Simulation Results

A. Simulation Parameters

Simulations were performed on a two-dimensional cellular system consisting of 19 hexagonal cells (see Fig. 1). Opposite sides wrap-around to eliminate the finite size effect. As the basic traffic type, packetized voice calls are generated for simulation purposes. For packetized voice, a packet loss probability of $P_L = 0.01$ is acceptable.

The common parameters used in the simulation are as follows. All the cells have the same capacity c. Target packet loss probability is $P_L = 0.01$, control interval is set to $T = 20\,s$ and all the neighboring cells have the same chance to be chosen by a call for handoff, i.e. $r_{ji} = 1/6$.

Besides, for ease of illustrating the results, we assumed that the system is uniform and the input load is the same for every cell, although PFG is designed to handle the nonuniform case and also, the simulation can accommodate arbitrary load distributions. In all of the cases simulated, *normalized load* is used to show a fair comparison of performance measures irrespective of the absolute values of cell capacity and arrival rates, where the normalized load is defined as

$$\rho = \frac{1}{M_i}\Big(\frac{\lambda}{\mu}\Big), \tag{44}$$

where M_i is given by (1).

For each load, simulations were done by averaging over 8 samples, each for $10^4\, s$ of simulation time. Call duration and cell residency times are exponentially distributed with means $\mu^{-1} = 180\, s$ and $h^{-1} = 100\, s$, respectively (except the last simulated case). We found this set of parameters more or less common and reasonable for a simulation setup (see for example [16]).

B. Traffic Model

The two-state Markov model shown in Fig. 5 is used to describe the traffic generation process of voice calls. It has been shown that this simple process can model voice and video traffic sources [18] as well as other complicated traffics [19]. In this model, α and β are transition rates to OFF and ON states, respectively, from ON and OFF states. While in the ON state, traffic is generated at a constant rate of A packet/sec. The *activity factor* of such a traffic source is defined to be the probability of being in the ON state and is given by

$$\eta = \frac{\beta}{\alpha + \beta}\,. \tag{45}$$

For this traffic model, the mean and variance of the traffic generated is given by $E[r] = \eta A$ and $V[r] = \eta(1-\eta)A^2$ [18]. Commonly used parameters for human speech representation are $\alpha^{-1} = 1.2\, s$ and $\beta^{-1} = 1.8\, s$ [25], [26]. Using an 8 Kbps encoded voice source, it is obtained that $A = 100$ packet/sec and hence, $E[r] = 40$ packet/sec and $V[r] = 50$ packet/sec assuming that each packet is 80 bits.

C. Conservative PFG

As mentioned earlier, PFG does not drop any handoff call, instead some packets may be dropped to accommodate the incoming handoff packets. To have an intuition about the impact

```
if (x is a handoff call) then
  if (N_i(t) ≤ M_i) then
    accept call;
  else
    reject call;
  end if
else /* x is a new call */
  if (rand(0,1) < a_i) & (N_i(t) ≤ M_i) then
    accept call;
  else
    reject call;
  end if
end if
```

Fig. 6. Pseudo-code of PFG-DP algorithm in cell i.

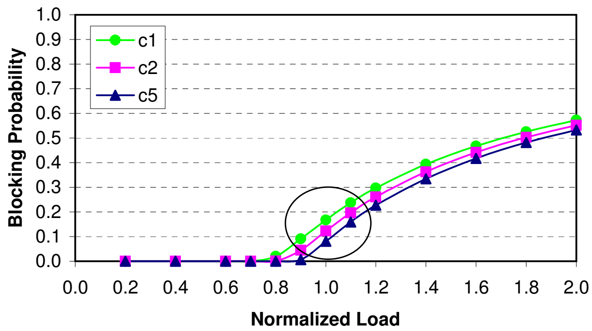
of accepting handoffs even during the overloaded state, we have also implemented a slightly different version of PFG in addition to the original PFG represented in Fig. 2.

This modified version drops handoffs during the overloaded state. We refer to the original algorithm by PFG-D0 and the modified one by PFG-DP where D0 and DP stand for zero dropping probability and P dropping probability, i.e. if we use PFG-DP instead of PFG-D0 then there will be P percent call dropping. Our purpose is to find the value of P for some simulated scenarios to see how far it is from zero. Notice that, having $N_i(t) > M_i$ ($t \in (0, T]$) indicates that cell i is in the overloaded state at time t. The pseudo-code for PFG-DP in cell i is given by the algorithm in Fig. 6.

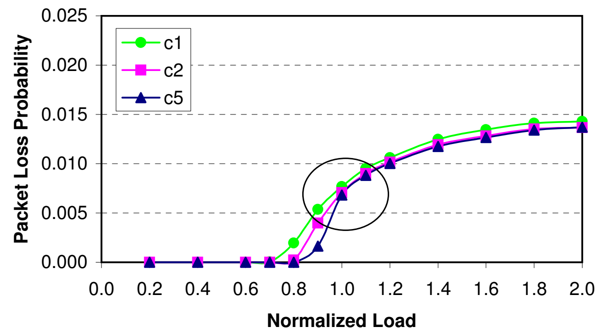
D. Results and Analysis

As mentioned earlier, PFG is the first to achieve zero call dropping while guaranteeing a hard constraint on packet loss probability. To the best of our knowledge there is no existing scheme which takes into consideration a combination of call-level and packet-level QoS parameters as defined by packetized MINB in section II. Therefore, we are not able to compare the performance of PFG with any other scheme. Instead, by doing extensive simulations, we have shown that PFG can achieve its defined goals.

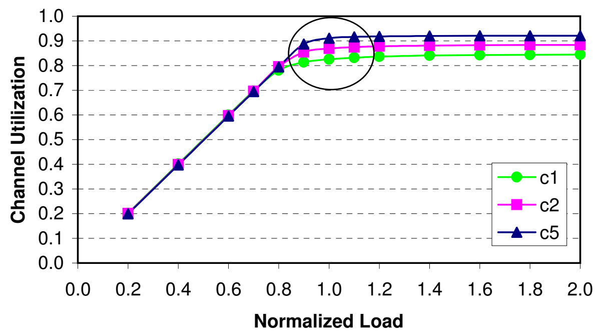
In the rest of this section, we present our simulation results. Several scenarios have been considered to investigate the impact of major factors such as cell capacity, control interval, cell residency and mobility on the performance of PFG. In all the simulated cases, PFG is stable and achieves accurate results.



(a) Blocking probability.



(b) Packet loss probability.



(c) Channel utilization.

Fig. 7. PFG-D0 performance results.

TABLE II

CELL CAPACITY PROFILES.

Profile	Capacity
$c1$	1 Mbps
$c2$	2 Mbps
$c5$	5 Mbps

1) Effect of cell capacity: Intuitively, increasing the cell capacity leads to a better Gaussian approximation, and hence, a more accurate admission decision. To investigate the effect of cell capacity, we considered three different capacity configurations as shown in Table II. Keep in mind that the system under consideration is a broadband wireless network such as 3G and 4G systems. Normalized loads in range $[0 \dots 2]$ are simulated, where the normalized load is defined by (44).

In Figs. 7(a), 7(b) and 7(c) we have circled a region around load $\rho = 1.0$. This is the most

TABLE III

PFG-DP CALL DROPPING PROBABILITY.

Load	$c1$	$c2$	$c5$
0.2	0.000000	0.000000	0.000000
0.6	0.000000	0.000000	0.000000
1.0	0.000000	0.000000	0.000000
1.4	0.000007	0.000002	0.000001
1.8	0.000012	0.000006	0.000005

interesting part of the system which is likely to happen in practice. In the following discussion we refer to this region as the *operating region* of the system.

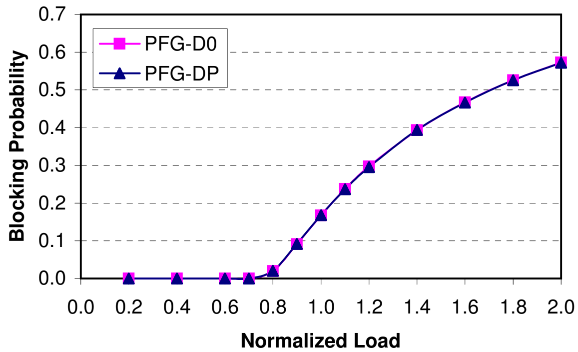
Fig. 7(a) shows the new call blocking probability. It is clear from the figure that as the cell capacity increases the blocking probability decreases which can be explained from the central limit theorem and Gaussian approximation used in section V-C. As the system capacity increases, the Gaussian modeling leads to more accurate approximation and hence, decreased call blocking probability.

The packet loss probability, $\tilde{L}_i$, is depicted in Fig. 7(b). Although $\tilde{L}_i$ goes beyond the target limit for high system loads, it is completely satisfactory for the operating region. Nevertheless, it is quite possible to modify PFG-D0 in order to make it more conservative for high loads. Similar to call blocking, as the capacity increases the PFG-D0 efficiency improves.

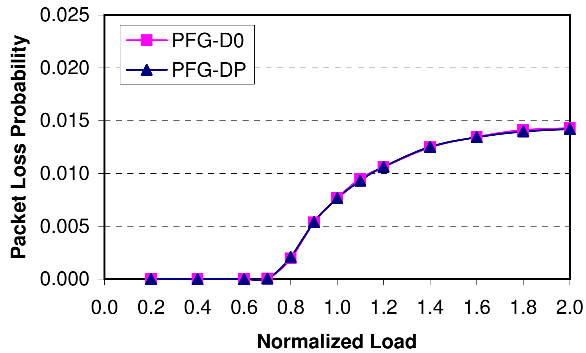
Fig. 7(c) depicts the wireless bandwidth utilization under the three different system capacities. As explained before, increased accuracy of the Gaussian approximation for high system capacity leads to a better channel utilization. After all, $c1$ produces rather accurate results and increasing the capacity beyond it produces only marginal improvements.

2) Effect of accepting handoffs in overloaded state: To investigate the impact of accepting handoffs during the overloaded state (in which $N_i(t) > M_i$), we ran PFG-DP for the same simulation configuration we ran PFG-D0. Table III shows the call dropping probabilities for different loads and capacities. It is observed that the call dropping probability is almost zero in all the simulated configurations. It means that basically there is no difference between two schemes in terms of the call dropping probability.

Fig. 8 shows the call blocking and packet loss probabilities of PFG-D0 versus PFG-DP when



(a) Blocking probability.



(b) Packet loss probability.

Fig. 8. PFG-D0 vs. PFG-DP.

TABLE IV

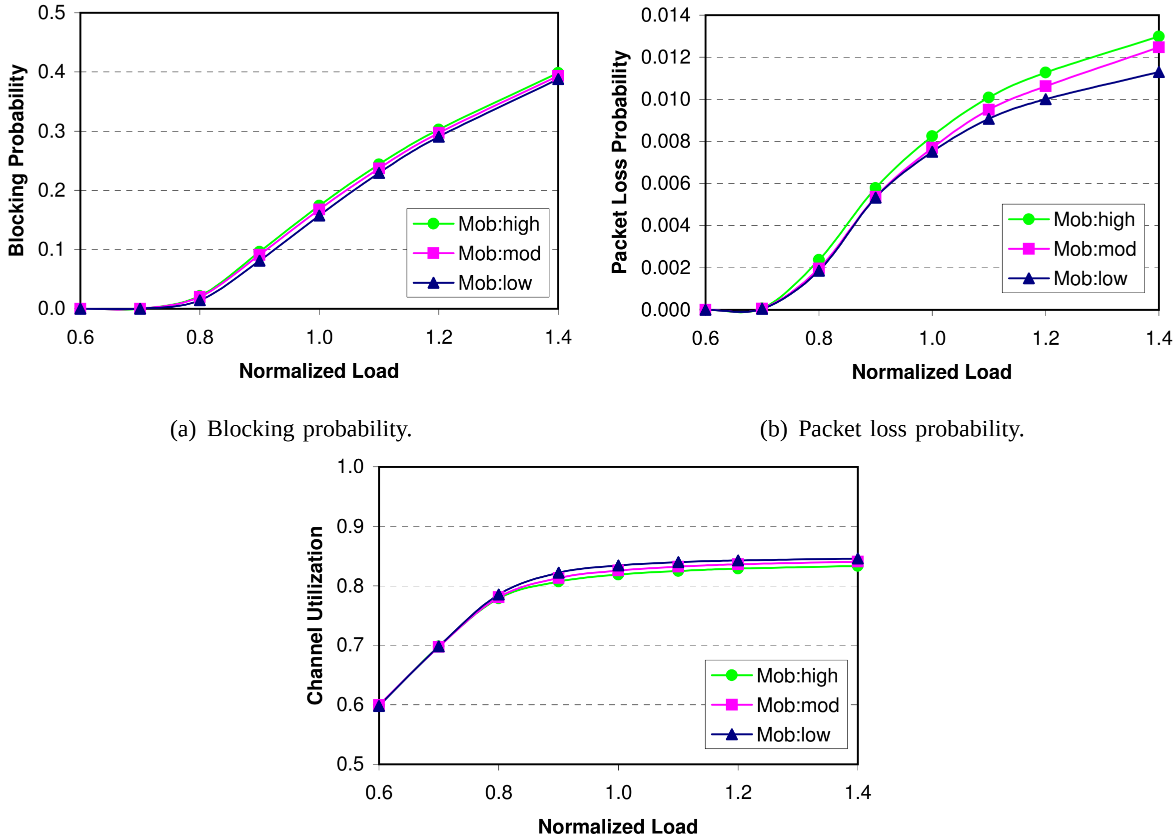
MOBILITY PROFILES.

Profile	$1/\mu\,(s)$	$1/h\,(s)$	α
Mob: high	180	20	9.00
Mob: mod	180	100	1.80
Mob: low	180	500	0.36

the system capacity is set to $c1$ (1 Mbps). Overall, there is no difference between the two schemes. It can be seen from Fig. 8(b) that the packet loss probability is almost the same for both schemes indicating that accepting handoffs during the overloaded state has a negligible effect on the admission control performance. Fig. 8(a) further confirms the same result.

3) Effect of mobility: To increase the capacity of cellular networks, micro/pico cellular architectures will be deployed in the future. The smaller cell size of these architectures leads to a higher handoff rate. Define the *mobility factor* to be $\alpha = h/\mu$. Intuitively, α shows the average number of handoff attempts a call makes during its life time. As the mobility factor increases the handoff arrival rate increases as well. To investigate the impact of mobility on PFG, we have simulated three mobility cases for the base capacity $c1$ as shown in Table IV. In this table, $\alpha = 9.00$ represents a highly mobile scenario such as vehicular users in a high way; $\alpha = 1.80$ is a common scenario typically used in similar research papers [12], [16] and shows an urban area mobility, and finally, $\alpha = 0.36$ represents a low mobility case.

Observed from Fig. 9, PFG is almost insensitive to the mobility rate of users. As shown



(a) Blocking probability.

(b) Packet loss probability.

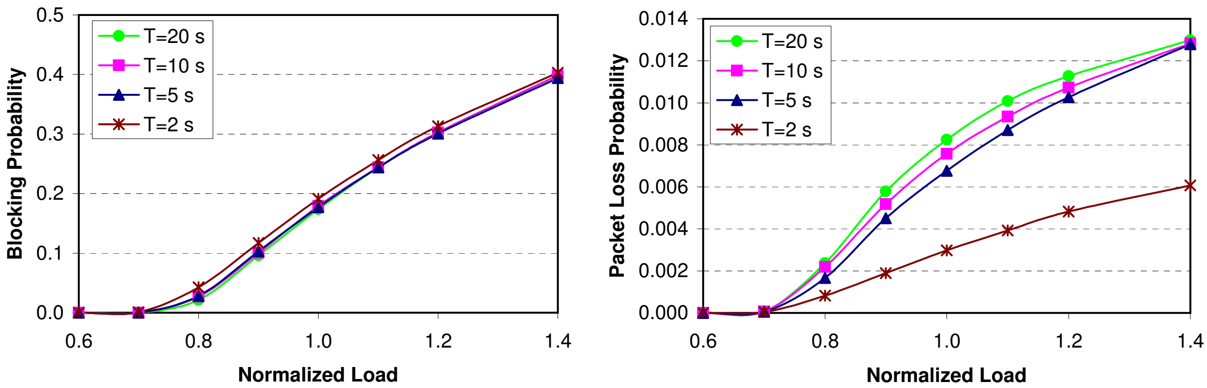
(c) Channel utilization.

Fig. 9. Mobility impact on PFG-D0 performance.

in Figs. 9(a) and 9(c), the call blocking probability and channel utilization are almost match. Furthermore, Fig. 9(b) shows that the effect of mobility on packet loss probability is not very significant. In all three cases, PFG is able to satisfy the target packet loss probability in the operating region of the system. In general, handoff degrades the performance of cellular systems.

4) Effect of Control Interval: Both signalling overhead and accuracy of PFG are affected by the control interval. Although increasing the control interval reduce the signalling overhead, the admission control accuracy will deteriorate. Therefore, there must be a compromise between the incurred overhead and the achieved accuracy. As we showed in subsection V-E, this compromise depends on the mobility of users.

Fig. 10 shows the effect of control interval on the performance of PFG. The simulated scenarios consider the high mobility profile in Table IV, where the mobility factor is set to $\alpha = 9$. It is observed that by reducing the control interval T, the accuracy of PFG in terms of the achieved



(a) Blocking probability.

(b) Packet loss probability.

Fig. 10. Effect of control interval.

packet loss probability increases.

An interesting question is that what is the appropriate control interval for high mobility scenario to achieve the same performance as moderate mobility scenario? Fig. 9 shows that there is a small discrepancy between two scenarios when the control interval is the same and equal to $T = 20\, s$.

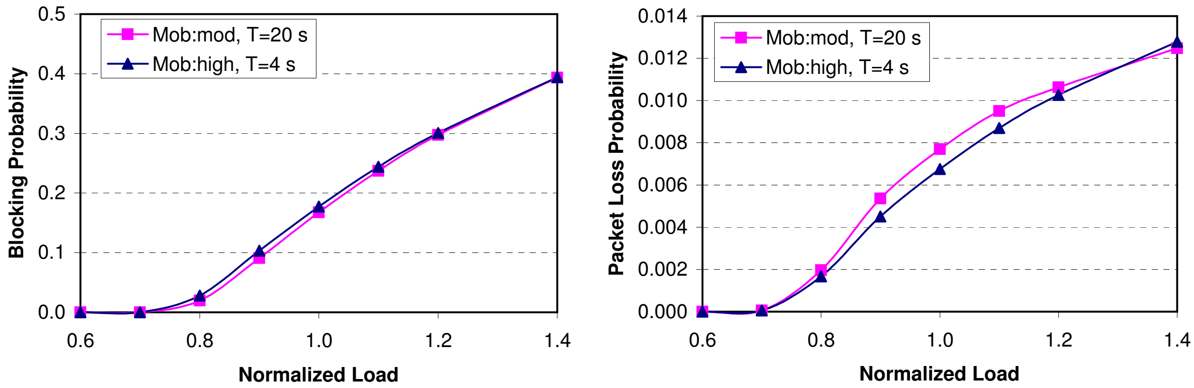
Using (43), it is obtained that

$$\frac{T_{\text{Mob: high}}}{T_{\text{Mob: mod}}} = \frac{\alpha_{\text{Mob: mod}}}{\alpha_{\text{Mob: high}}}. \tag{46}$$

Therefore, $T_{\text{Mob: high}}$ must be set to $\frac{1}{5}T_{\text{Mob: mod}}$ in order to see the same performance results. Fig. 11 shows the simulation results for high mobility and moderate mobility scenarios where $T_{\text{Mob: high}} = 4\, s$ and $T_{\text{Mob: mod}} = 20\, s$.

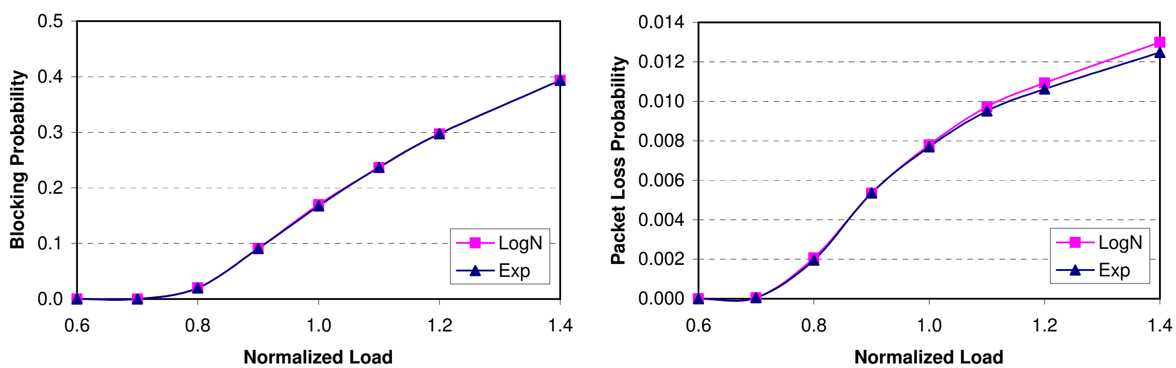
5) Effect of non-exponential cell residence times: The first part of our analysis, which gives the equations describing the mean and variance of the traffic generation process, is based on the assumption of the exponential cell residency time. As mentioned earlier, exponential distributions provide the mean value analysis, which indicates the performance trend of the system. However, in practice, cell residence times are usually non-exponentially distributed. In this section, we investigate the sensitivity of PFG to exponential cell residency assumption.

Using real measurements, Jedrzycki and Leung [27] showed that a lognormal distribution is a more accurate model for cell residency time. We now compare the results obtained under exponential distribution with those obtained under more realistic lognormal distribution. The



(a) Blocking probability. (b) Packet loss probability.

Fig. 11. Robustness to mobility patterns.



(a) Blocking probability. (b) Packet loss probability.

Fig. 12. Lognormal vs. Exponential residence time.

mean and variance of both distributions are the same. Fig. 12 shows the call blocking and packet loss probability of exponential cell residency versus lognormal cell residency. It is observed that the exponential cell residency achieves sufficiently accurate control. In other words, the control algorithm is rather insensitive to this assumption due to its periodic control in which the length of the control interval is much smaller than the mean residency time.

VII. Conclusion

In this paper we presented a novel scheme for admission control and hence QoS provisioning for packet-switched cellular systems. In essence, our approach is the natural generalization of the well-known effective bandwidth [19] proposed for wireline networks. Through analysis and simulation, we showed that the proposed scheme, PFG, is able not only to improve utilization of scarce wireless bandwidth thanks to the statistical multiplexing of VBR traffic sources but also to eliminate the undesirable call dropping event inherent to circuit-switched cellular systems.

In wireless multimedia networks, there are different service classes, each of which has its own packet and call level QoS constraints. We are currently investigating the extension of PFG to multiple service classes where each service class has its own QoS requirements. Also, a preliminary work shows that by embedding the loss rate into equation (31), PFG is able to have a more precise control on actual packet loss probability.

REFERENCES

[1] U. Varshney and R. Jain, "Issues in emerging 4G wireless networks," *IEEE Computer*, vol. 34, no. 6, pp. 94–96, June 2001.

[2] T. Zahariadis and D. Kazakos, "(R)evolution toward 4G mobile communication systems," *IEEE Wireless Commun. Mag.*, vol. 10, no. 4, pp. 6–7, Aug. 2003.

[3] S. Y. Hui and K. H. Yeung, "Challenges in the migration to 4G mobile systems," *IEEE Computer*, vol. 41, no. 12, pp. 54–59, Dec. 2003.

[4] A. W. Berger and W. Whitt, "Extending the effective bandwidth concept to networks with priority classes," *IEEE Commun. Mag.*, vol. 36, no. 8, pp. 78–83, 1998.

[5] Q. Gao and A. Acampora, "Performance comparisons of admission control strategies for future wireless networks," in *Proc. IEEE WCNC'02*, vol. 1, Orlando, USA, Mar. 2002, pp. 317–321.

[6] 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. 35, no. 3, pp. 77–92, Aug. 1986, see also: CEAS Tech. Rep. No. 773, College of Engineering and Applied Sciences, State University of New York, June 1999.

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

[8] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE J. Select. Areas Commun.*, vol. 14, no. 4, pp. 711–717, May 1996.

[9] D. Levine, I. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Trans. Networking*, vol. 5, no. 1, pp. 1–12, Feb. 1997.

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

[11] J. M. Peha and A. Sutivong, "Admission control algorithms for cellular systems," *ACM/Kluwer Wireless Networks*, vol. 7, no. 2, pp. 117–125, Mar. 2001.

[12] B. Li, L. Yin, K. Y. M. Wong, and S. Wu, "An efficient and adaptive bandwidth allocation scheme for mobile wireless networks using an on-line local estimation technique," *ACM/Kluwer Wireless Networks*, vol. 7, no. 2, pp. 107–116, 2001.

[13] K. Mitchell and K. Sohraby, "An analysis of the effects of mobility on bandwidth allocation strategies in multi-class cellular wireless networks," in *Proc. IEEE INFOCOM'01*, vol. 2, Anchorage, USA, Apr. 2001, pp. 1005–1011.

[14] T. Zhang, E. Berg, J. Chennikara, P. Agrawal, J. C. Chen, and T. Kodama, "Local predictive resource reservation for handoff in multimedia wireless IP networks," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. 1931–1941, Oct. 2001.

[15] 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.

[16] S. Wu, K. Y. M. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, no. 2, pp. 257–271, Apr. 2002.

[17] W.-S. Soh and H. S. Kim, "QoS provisioning in cellular networks based on mobility prediction techniques," *IEEE Commun. Mag.*, vol. 41, no. 1, pp. 86–92, Jan. 2003.

[18] M. Schwartz, *Broadband Integrated Networks*. New Jersey, USA: Prentice Hall, 1996.

[19] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, no. 7, pp. 968–981, 1991.

[20] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York, USA: McGraw-Hill, 1965.

[21] D. Y. Eun and N. B. Shroff, "A measurement-analytic approach for QoS estimation in a network based on the dominant time scale," *IEEE/ACM Trans. Networking*, vol. 11, no. 2, pp. 222–235, Apr. 2003.

[22] W. E. Leland, M. Taque, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[23] J. Beran, R. Sherman, M. S. Taque, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, no. 2, pp. 1566–1579, Feb. 1995.

[24] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, UK: Cambridge University Press, 1992.

[25] P. T. Brady, "A model for generating on-off speech patterns in two way conversations," *Bell System Tech. J.*, vol. 48, pp. 2445–2472, Sept. 1969.

[26] J. N. Diagle and J. D. Langford, "Models for analysis of packet voice communications systems," *IEEE J. Select. Areas Commun.*, vol. 4, no. 6, pp. 847–855, Sept. 1986.

[27] C. Jedrzycki and V. C. M. Leung, "Probability distribution of channel holding time in cellular telephone systems," in *Proc. IEEE VTC'96*, vol. 1, Atlanta, GA, May 1996, pp. 247–251.