Call Admission Control for Voice/Data Integration in Broadband Wireless Networks

Majid Ghaderi and Raouf Boutaba, *Senior Member*, *IEEE*

Abstract—This paper addresses bandwidth allocation for an integrated voice/data broadband mobile wireless network. Specifically, we propose a new admission control scheme called EFGC, which is an extension of the well-known fractional guard channel scheme proposed for cellular networks supporting voice traffic. The main idea is to use two acceptance ratios, one for voice calls and the other for data calls in order to maintain the proportional service quality for voice and data traffic while guaranteeing a target handoff failure probability for voice calls. We describe two variations of the proposed scheme: EFGC-REST, a conservative approach which aims at preserving the proportional service quality by sacrificing the bandwidth utilization, and EFGC-UTIL, a greedy approach which achieves higher bandwidth utilization at the expense of increasing the handoff failure probability for voice calls. Extensive simulation results show that our schemes satisfy the hard constraints on handoff failure probability and service differentiation while maintaining a high bandwidth utilization.

Index Terms—Call admission control, voice/data integration, quality-of-service, broadband wireless networks.

---

1 INTRODUCTION

EMERGING wireless technologies such as 3G and 4G will increase the cell capacity of wireless cellular networks to several Mbps [1]. With this expansion of wireless bandwidth, the next generations of mobile cellular networks are expected to support diverse applications such as voice, data, and multimedia with varying quality of service (QoS) and bandwidth requirements [2]. Wireless links bandwidth is limited and is generally much smaller than that of wireline access links. Therefore, for integrated voice/data mobile networks, it is necessary to develop mechanisms that can provide effective bandwidth management while satisfying the QoS requirements of both types of traffic.

At call-level, two important quality of service parameters are the *call blocking probability* (p_b) and the *call dropping probability* (p_d). An active mobile user in a cellular network may move from one cell to another. The continuity of service to the mobile user in the new cell requires a successful handoff from the previous cell to the new cell. The probability of a handoff failure is called *handoff failure probability* (p_f). During the life of a call, a mobile user may cross several cell boundaries and, hence, may require several successful handoffs. Failure to get a successful handoff at any cell in the path forces the network to drop the call. The probability of such an event is known as the call dropping probability.

Since dropping a call in progress has a more negative impact from the user perspective, handoff calls are given higher priority than new calls in accessing the wireless resources. This preferential treatment of handoffs increases the probability of blocking new calls and, hence, may degrade the bandwidth utilization. The most popular approach to prioritize handoff calls over new calls is by reserving a portion of available bandwidth in each cell to be used exclusively for handoffs. Based on this idea, a number of call admission control (CAC) schemes have been proposed which basically differ from each other in the way they calculate the reservation threshold [3], [4], [5], [6], [7], [8].

Bandwidth allocation has been extensively studied in single-service (voice) wireless cellular networks. Hong and Rappaport [3] 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. [9] 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 (FGC) [9] is optimal for minimizing call blocking probability subject to a hard constraint on call dropping probability. Instead of explicit bandwidth reservation as in GC, the FGC accepts new calls according to a randomization parameter called the *acceptance ratio*. One advantage of FGC over GC is that it distributes the new accepted calls evenly over time, which leads to a more stable control [10].

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 [4], [5], [6], [7], [8], [10], [11]. Information exchange among a cluster of neighboring cells is the approach adopted by all distributed schemes.

In particular, Naghshineh and Schwartz [4] proposed a collaborative admission control known as *distributed call admission control* (DCA). DCA periodically gathers 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

---

- *The authors are with the David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada.*
  E-mail: {mghaderi, rboutaba}@uwaterloo.ca.

load increases [10]. Levine et al. [5] 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. [10] proposed a distributed scheme called 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.

One of the challenges in considering multiservices systems is that the already limited bandwidth has to be shared among multiple traffics. Epstein and Schwartz [12] investigated complete sharing, complete partitioning, and hybrid reservation schemes for two classes of traffic, namely, narrow-band and wide-band traffic. In general, complete sharing strategy achieves the highest bandwidth utilization [12].

Fixed and movable boundary schemes for bandwidth allocation in wireless networks were studied by Wieselthier and Ephremides [13]. They concluded that movable boundary schemes can achieve a better utilization than fixed boundary schemes for voice and data integration. Since then, a number of papers have been published focusing on the performance of fixed and movable boundary schemes given different assumptions and network configurations [14], [15], [16], [17], [18], [19], [20].

In particular, Haung et al. [18] proposed a bandwidth allocation scheme for voice/data integration based on the idea of movable boundaries (MB). In their scheme, bandwidth is divided into two portions that can be dynamically adjusted to achieve the desired performance. However, they completely neglected the prioritization of handoff calls over new calls and treated the two identically. Yin et al. [19] proposed a *dual threshold reservation* (DTR) scheme, which extends the basic guard channel to use two thresholds, one for reserving channels for voice handoff, and the other for limiting the data traffic into the network in order to preserve the voice performance. An extended version of DTR which implements queueing for data calls (DTR-Q) was proposed in [20]. In general, queueing of new/handoff calls can further improve the performance of call admission control [21]. The main limitation of DTR (DTR-Q) is that it is static, i.e., the two reservation thresholds are fixed over time regardless of the state of the network. Interested readers are referred to [22] for a comparison between DTR and MB schemes.

This paper introduces an *extended fractional guard channel call admission mechanism* (EFGC) for integrated voice and data mobile cellular networks. EFGC maximizes the wireless bandwidth utilization while satisfying a target call dropping probability and a relative voice/data service differentiation. The main idea is to use two acceptance ratios for voice and data according to the desired dropping probability of voice calls and the relative priority of voice calls over data calls. Similar to [15], [16], [17], [18], [19], [20], we assume that call dropping is not an important issue for data calls and treat handoff and new data calls in the same way. We define the extended MINBLOCK [9] problem as follows:

For a given cell capacity, maximize the bandwidth utilization subject to a hard constraint on the voice call dropping probability and relative voice/data call blocking probability.

To the best of our knowledge, extending the basic fractional guard channel scheme to address the extended MINBLOCK problem is a novel work. We follow an approach similar to the admission control algorithm proposed by Wu et al. [10] to derive the acceptance ratios for voice and data calls. The main features of EFGC are as follows:

1. EFGC is dynamic; therefore, it adapts to a wide range of system parameters and traffic conditions.
2. EFGC uses separate acceptance ratios for voice and data calls; therefore, it is very straightforward to enforce a relative or even strict service differentiation between voice and data traffic.
3. EFGC is distributed and takes into consideration the information from direct neighboring cells in making admission decisions.
4. The control mechanism is stochastic and periodical to reduce the overhead associated with distributed control schemes. EFGC determines the appropriate control parameters, such as the control interval length, in order to restrict the impact of the network to the direct neighbors only.

The rest of the paper is organized as follows: Our system model, assumptions, and notations are described in Section 2. Section 3 is dedicated to the proposed admission control algorithm and presents the analysis of the proposed algorithm in detail. In Section 4, we discuss the estimation of control parameters such as arrival rates, then we address the multiple handoffs problem and control interval length. Extensive simulation results and their analysis are presented in Section 5. Finally, Section 6 concludes this paper.

2 SYSTEM MODEL

As shown in Fig. 1, we consider a cellular system which carries both voice and data traffic. We assume that wireless bandwidth is channelized, where a channel can be a frequency, a time slot, or a code sequence. We define the basic bandwidth unit (BU) as the smallest amount of bandwidth that can be allocated to a call, e.g., a channel. In this paper, we focus on call-level QoS parameters; therefore, only call-level traffic dynamics are required for resource allocation and admission control. More specifically, we assume that the *effective bandwidth* [23], [24], [25] concept is applied to each call. When employing this concept, an appropriate effective bandwidth is assigned to each call and each call is treated as if it required this effective bandwidth throughout the active period of the call. The feasibility of admitting a given set of connections may then be determined by ensuring that the sum of the effective bandwidths is less than or equal to the total available bandwidth, i.e., the cell capacity.

We assume that each voice call requires b_v BUs and each data call requires b_d BUs for the whole duration of the call. In the system under consideration, voice handoff calls have the highest priority, then come new voice calls, and, last, the new and handoff data calls are considered. As mentioned earlier, there is no prioritization of handoff data calls and, hence, handoff data calls are treated the same as new data calls.

The considered system is not required to be uniform. Each cell can experience a different load, e.g., some cells can
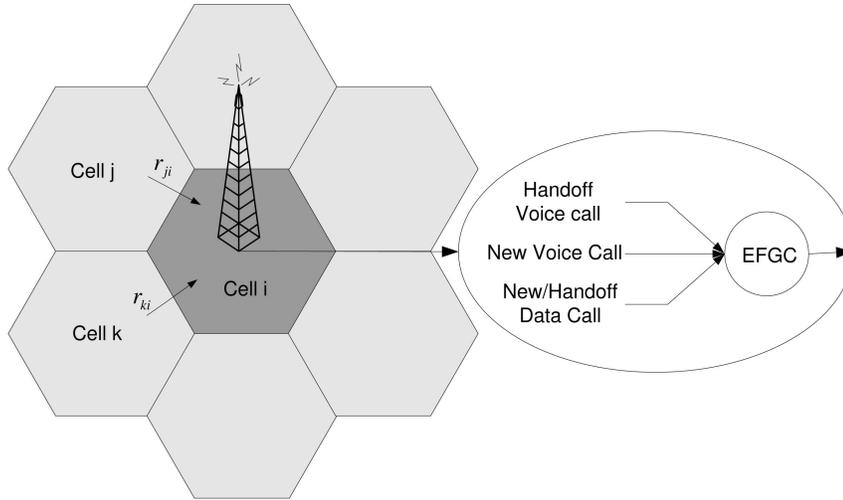
Fig. 1. Integration of voice and data at the base station of a cellular network.

be overutilized while others are underutilized. Let $k = \{v, d\}$ denote the type of traffic, i.e., $k = v$ for voice and $k = d$ for data traffic. Below is the notation which will be used throughout this paper:

- M: Number of cells in the network.
- $\mathcal{A}_i$: The set of the adjacent cells of cell i.
- c_i: The capacity of cell i in terms of BUs.
- $R_i(t)$: Bandwidth requirements (used capacity) in cell i at time t in terms of BUs.
- p_{f_i}: Voice handoff failure probability in cell i.
- p_{QoS}: Target voice handoff failure probability to be guaranteed.
- k: The service index for voice and data with $k = v$ for voice and $k = d$ for data.
- λ_i^k: Type-k new call arrival rate into cell i.
- $1/\mu_k$: Type-k mean call duration.
- $1/h_k$: Type-k mean cell residency time.
- T: Length of the control period.
- b_k: Bandwidth requirement of type-k calls in terms of BUs.
- $N_i^k(t)$: Number of active type-k calls in cell i at time t.
- r_{ji}: Routing probability from cell $j \in \mathcal{A}_i$ to cell i.
- b_i^k: Type-k call blocking probability in cell i.
- a_i^k: Type-k call acceptance ratio in cell i.
- α_i: Relative priority of voice traffic over data traffic in cell i defined as $\alpha_i = a_i^v/a_i^d$.
- α_{QoS}: Target relative priority of voice traffic over data traffic to be guaranteed.
- p_b^k: Network-wide type-k call blocking probability.
- p_d: Network-wide voice call dropping probability.
- $E[z]$: The mean of random variable z.
- $V[z]$: The variance of random variable z.
- $\tilde{z}$: Time-averaged value of random variable z.
- $\hat{z}$: Measured (observed) value of random variable z.

Let random variables t_{d_k} and t_{r_k} denote the call duration (call holding time) and cell residency time of a typical type-k call, respectively. Similar to [3], [9], [10], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], we assume that t_{d_k} and t_{r_k} are exponentially distributed. In the real world, 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.

2.1 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 signaling 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 information exchange is triggered only at the beginning of control periods. 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 signaling 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 reflect 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 signaling overhead will not increase, while the collected information on the network status will be sufficiently accurate for the purpose of a stochastic
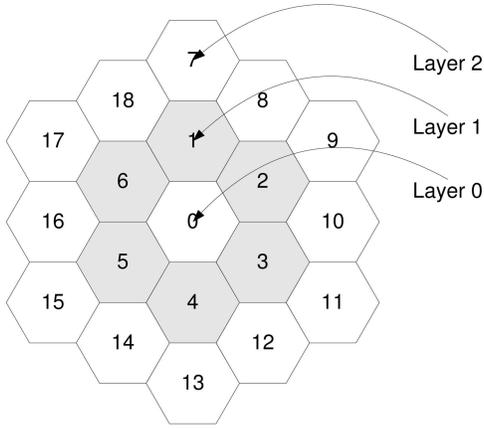
Fig. 2. A cellular system with three layers.

admission control. The reason is that, first, by decreasing the control interval, the probability of multiple handoffs decays to zero exponentially (see Section 4.3), and, 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. 2. 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

$$M(n) = 1 + \mathcal{A} + \cdots + n\mathcal{A}$$
$$= 1 + \frac{1}{2}n(n+1)\mathcal{A}. \tag{1}$$

Therefore, by slightly reducing the control interval, we essentially achieve the same control accuracy but with reduced signaling overhead. The problem of choosing the proper control interval will be further addressed in Section 4.3.

2.2 Handoff Failure and Call Dropping Probabilities

Although call dropping probability is more meaningful for mobile users and service providers, calculating the handoff failure probability is more convenient. Therefore, our calculations in this paper are based on the handoff failure probability, p_f, which can be related to the call dropping probability, p_d, as follows (refer to [3] for more details):

$$p_d = \sum_{H=0}^{\infty} (P_h^v)^H (1-p_f)^{H-1} p_f = \frac{P_h^v p_f}{1 - P_h^v(1-p_f)}, \tag{2}$$

where H is the number of possible handoffs during the life of a call and P_h^v is the handoff probability of a voice call before the call completes, which can be computed by the following equation:

$$P_h^v = \Pr(t_{d_v} > t_{r_v})$$
$$= \int_{t=0}^{\infty} \Pr(t_{d_v} > t_{r_v} | t_{r_v}) \; \Pr(t_{r_v} = t) \; dt \tag{3}$$
$$= \int_{t=0}^{\infty} h_v \exp(-\mu_v t) \exp(-h_v t) \; dt = \frac{h_v}{\mu_v + h_v},$$

therefore,

$$p_f = \frac{p_d}{1 - p_d} \left(\frac{\mu_v}{h_v}\right). \tag{4}$$

It means that, for a given p_d, the equivalent p_f can be easily computed based on (4). Therefore, in this paper, it is assumed that a target handoff failure probability p_{QoS} must be guaranteed for voice calls. Notice that exponential assumption is a necessary condition in deriving (3). Interested readers are referred to [26], [27] for the handoff probability under general call duration and cell residency distributions.

2.3 Time-Dependent Handoff and Stay Probabilities

Here, we compute some useful probabilities required for the rest of our discussion. Let $P_h^k(t)$ denote the probability that a type-k call hands off by time t and remains active until t, given that it has been active at time 0. Also, let $P_s^k(t)$ denote the probability that a type-k call remains active in its home cell until time t, given that it has been active at time 0. Then,

$$P_h^k(t) = \Pr(t_{r_k} \leq t) \; \Pr(t_{d_k} > t)$$
$$= (1 - \exp(-h_k t)) \; \exp(-\mu_k t) \tag{5}$$

and

$$P_s^k(t) = \Pr(t_{r_k} > t) \; \Pr(t_{d_k} > t)$$
$$= \exp(-(\mu_k + h_k)t). \tag{6}$$

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 $\tilde{P}_h^k$ and $\tilde{P}_s^k$ are expressed as

$$\tilde{P}_h^k(t) = \frac{1}{t} \int_0^t P_h^k(t - t') \; dt', \tag{7}$$

$$\tilde{P}_s^k(t) = \frac{1}{t} \int_0^t P_s^k(t - t') \; dt'. \tag{8}$$

These integrals can be easily computed with respect to (5) and (6). Finally, let $P_{ji}^k(t)$ denote the time-dependent handoff probability and $\tilde{P}_{ji}^k(t)$ denote the average time-dependent handoff probability from cell j to cell i where $j \in \mathcal{A}_i$. It is obtained that

$$P_{ji}^v(t) = P_h^v(t) \; r_{ji}, \tag{9}$$
$$\tilde{P}_{ji}^v(t) = \tilde{P}_h^v(t) \; r_{ji}, \tag{10}$$

because voice handoff calls are always accepted if there is enough free bandwidth. Similarly,

$$P_{ji}^d(t) = a_i^d \left[P_h^d(t) \; r_{ji}\right], \tag{11}$$
$$\tilde{P}_{ji}^d(t) = a_i^d \left[\tilde{P}_h^d(t) \; r_{ji}\right], \tag{12}$$

because data calls are always subject to an acceptance ratio a_i^d in cell i.

In the next section, we will use the computed probabilities to find the maximum acceptance ratios for voice and data calls with respect to the prespecified call dropping probability (p_{QoS}) and relative voice/data acceptance probability (α_{QoS}).

3 ADMISSION CONTROL ALGORITHM

The proposed distributed algorithm, EFGC, consists of two components. The first component is responsible for retrieving the required information from the neighboring cells and computing the control parameters. Using the computed control parameters, the second component enforces the admission control locally in each cell. The following sections describe these two components in detail.

3.1 Distributed Control Algorithm

As mentioned earlier, to reduce the signaling overhead, EFGC has a periodic structure. All the information exchange and control parameter computations happen only once at the beginning of each control period of length T. Several steps involved in EFGC 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 voice and data calls presented in the cell at the beginning of the control period denoted by $N_i^v(0)$ and $N_i^d(0)$, respectively.

   b. The number of new voice calls, N_i^v, and new/handoff data calls, N_i^d, which were admitted in the last control period.

2. Each cell i receives $N_j^k(0)$ and N_j^k 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 ratios a_i^v and a_i^d using the technique described in Section 3.3.

4. Finally, the computed acceptance ratios a_i^v and a_i^d are used to admit call requests into cell i using the algorithm presented in Section 3.2.

Assume that all the cells have the same number of adjacent cells. Let $\mathcal{A}$ denote the number of adjacent cells. Also, assume that all the required information can be sent from one cell to another cell in one message. Then, the signaling overhead in terms of the number of exchanged messages in one control period is $\mathcal{A}$ messages per cell.

3.2 Local Admission Control Algorithm

Let (m, n) denote the state of cell i, where there are m voice calls and n data calls active in the cell. Define $\mathcal{S}_i$ as the state space of cell i governed by the EFGC scheme. Then, $\mathcal{S}_i$ can be expressed as

$$\mathcal{S}_i = \{(m, n) \mid mb_v + nb_d \le c_i\}. \tag{13}$$

Let $a_i^k(m, n)$ denote the acceptance ratio for type-k calls where the cell state is (m, n). Fig. 3 shows the state transition diagram of the EFGC scheme in cell i for a typical state $(m, n) \in \mathcal{S}_i$. In this figure, ν_i^k is the type-k handoff arrival rate into cell i. At each state, there are two acceptance ratios for voice and data calls in such a way that

$$\begin{cases} a_i^v(m, n) = 0, & \text{if } (m+1, n) \notin \mathcal{S}_i \\ a_i^d(m, n) = \frac{1}{\alpha_i} \, a_i^v(m, n), & \text{if } (m, n) \in \mathcal{S}_i. \end{cases} \tag{14}$$

There is a service differentiation (α_i) between voice and data calls that govern the relation between these two acceptance ratios. In this paper, we assume that this service differentiation is specified a priori (α_{QoS}) and EFGC should maintain it regardless of traffic conditions.
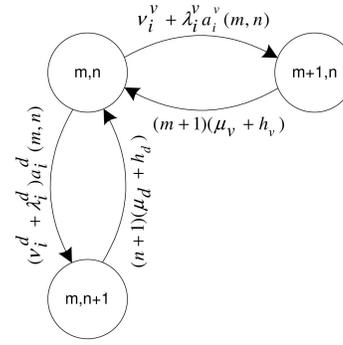
Fig. 3. Extended fractional guard channel transition diagram.

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, the call blocking probability is expressed as $b_i^k = 1 - E[a_i^k]$. Consequently, the average network-wide call blocking probability for the considered network is given by

$$p_b^k = \frac{\sum_{j=i}^{M} \lambda_i^k b_i^k}{\sum_{j=i}^{M} \lambda_i^k}. \tag{15}$$

The pseudocode for the local admission control in cell i is given by the algorithm of Fig. 4. In this algorithm, x_k is a type-k call requesting b_k BUs. The corresponding type-k acceptance ratio is a_i^k. Also, rand$(0, 1)$ is the uniform random generator function. In the next section, we will present a technique to compute the acceptance ratio vector $a_i = (a_i^v, a_i^d)$ in order to complete this algorithm.

3.3 Computing Acceptance Ratios

It is assumed that, by setting the control interval T to an appropriate value, each call experiences at most one handoff during a control period (see Section 4.3 for more detail). Therefore, immediate neighbors of cell i, i.e., $\mathcal{A}_i$, are those which will affect the number of calls and, consequently, the bandwidth usage in cell i during a control period.

The proposed approach for computing the acceptance ratios 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.

2. The computed mean and variance of the number of calls is used to find the mean and variance of the bandwidth requirement process in the cell.

3. Having the mean and variance of the bandwidth requirement process, the actual time-dependent bandwidth requirement process is approximated by a Gaussian distribution.

4. The tail of this Gaussian distribution is used to find the time-dependent handoff failure in each cell i.

5. Time-dependent handoff failure is averaged over the control interval of length T to find an average handoff failure probability for the whole period.

6. Using the computed handoff failure probability and the prespecified QoS constraints, i.e., p_{QoS} and α_{QoS}, acceptance ratios a_i^v and a_i^d are computed.

```
if (x_k is a voice handoff call) then
    if (R_i(t) + b_v ≤ c_i) then
        accept call;
    else
        reject call;
    end if
else /* new voice or new/handoff data call */
    if (R_i(t) + b_k ≤ c_i)&(rand(0,1) < a_i^k) then
        accept call;
    else
        reject call;
    end if
end if
```

Fig. 4. Local call admission control algorithm in cell i.

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 $g_i^k(t)$ and $n_i^k(t)$ denote the number of background and new type-k calls in cell i at time t, respectively. A background type-k call in cell i will remain in cell i with probability $P_s^k(t)$ or will handoff to an adjacent cell j with probability $P_{ij}^k(t)$. A new type-k call which is admitted in cell i at time $t' \in (0, t]$ will stay in cell i with probability $\tilde{P}_s^k(t)$ or will handoff to an adjacent cell j with probability $\tilde{P}_{ij}^k(t)$. 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

$$V_s^k(t) = P_s^k(t)\,(1 - P_s^k(t)), \qquad (16)$$

$$V_{ji}^k(t) = P_{ji}^k(t)\,(1 - P_{ji}^k(t)), \qquad (17)$$

$$\tilde{V}_s^k(t) = \tilde{P}_s^k(t)\,(1 - \tilde{P}_s^k(t)), \qquad (18)$$

$$\tilde{V}_{ji}^k(t) = \tilde{P}_{ji}^k(t)\,(1 - \tilde{P}_{ji}^k(t)), \qquad (19)$$

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

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

$$E[N_i^k(t)] = E[g_i^k(t)] + E[n_i^k(t)], \qquad (20)$$

where

$$E[g_i^k(t)] = N_i^k(0)P_s^k(t) + \sum_{j \in \mathcal{A}_i} N_j^k(0)P_{ji}^k(t), \qquad (21)$$

$$E[n_i^k(t)] = (a_i^k \lambda_i^k t)\tilde{P}_s^k(t) + \sum_{j \in \mathcal{A}_i} (a_j^k \lambda_j^k t)\tilde{P}_{ji}^k(t). \qquad (22)$$

Similarly, the variance is given by

$$V[N_i^k(t)] = V[g_i^k(t)] + V[n_i^k(t)], \qquad (23)$$

where

$$V[g_i^k(t)] = N_i^k(0)V_s^k(t) + \sum_{j \in \mathcal{A}_i} N_j^k(0)V_{ji}^k(t), \qquad (24)$$

$$V[n_i^k(t)] = (a_i^k \lambda_i^k t)\tilde{V}_s^k(t) + \sum_{j \in \mathcal{A}_i} (a_j^k \lambda_j^k t)\tilde{V}_{ji}^k(t). \qquad (25)$$

Note that, given the arrival rate λ_i^k and the acceptance ratio a_i^k, the actual new call arrival rate into cell i is given by $\lambda_i^k a_i^k$ (see Section 4.2). Therefore, the expected number of call arrivals during the interval $(0, t]$ is given by $a_i^k \lambda_i^k t$.

Knowing the bandwidth requirement of each type of call, the mean and variance of bandwidth usage in cell i at time t are given by

$$E[R_i(t)] = b_v E[N_i^v(t)] + b_d E[N_i^d(t)], \qquad (26)$$

$$V[R_i(t)] = b_v^2 V[N_i^v(t)] + b_d^2 V[N_i^d(t)]. \qquad (27)$$

As we mentioned in Section 1, the cellular system considered in this paper is a broadband wireless system with a capacity of several Mbps. In practice, 3G systems and beyond can be considered as broadband wireless systems (for example, a UMTS system can support up to 2 Mbps) [1], [2]. With this range of cell capacity, it is reasonable to apply the central limit theorem (this will be further discussed in Section 5.3). Thus, the bandwidth usage in each cell can be approximated by a Gaussian distribution with mean $E[R_i(t)]$ and variance $V[R_i(t)]$. That is,

$$R_i(t) \sim \mathbf{G}\bigl(E[R_i(t)],\ V[R_i(t)]\bigr). \qquad (28)$$

Therefore, the original problem of maintaining a target handoff failure probability p_{QoS} is reduced to maintaining the bandwidth usage below the available capacity c_i at any point in time $t \in (0, T]$. Approximating the handoff failure probability by the overload probability, the time-dependent handoff failure probability $P_{f_i}(t)$ can be computed as follows:

$$P_{f_i}(t) = \Pr\bigl(R_i(t) > c_i\bigr), \qquad (29)$$

therefore,

$$P_{f_i}(t) = \frac{1}{2}\ \text{erfc}\left(\frac{c_i - E[R_i(t)]}{\sqrt{2\,V[R_i(t)]}}\right), \qquad (30)$$

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

$$\text{erfc}(c) = \frac{2}{\sqrt{\pi}}\int_c^\infty e^{-t^2}\,dt. \qquad (31)$$

Then, the average handoff failure probability over a control period is given by

$$\tilde{P}_{f_i} = \frac{1}{T}\int_0^T P_{f_i}(t)\,dt. \qquad (32)$$

Finally, to guarantee the target handoff failure p_{QoS}, we should have

$$\tilde{P}_{f_i} = p_{\text{QoS}}. \qquad (33)$$

To solve (33) for $a_i = (a_i^v, a_i^d)$, we need one more equation. This equation can be derived with respect to the required service differentiation. Given the service condition $a_d = f(a_v)$,
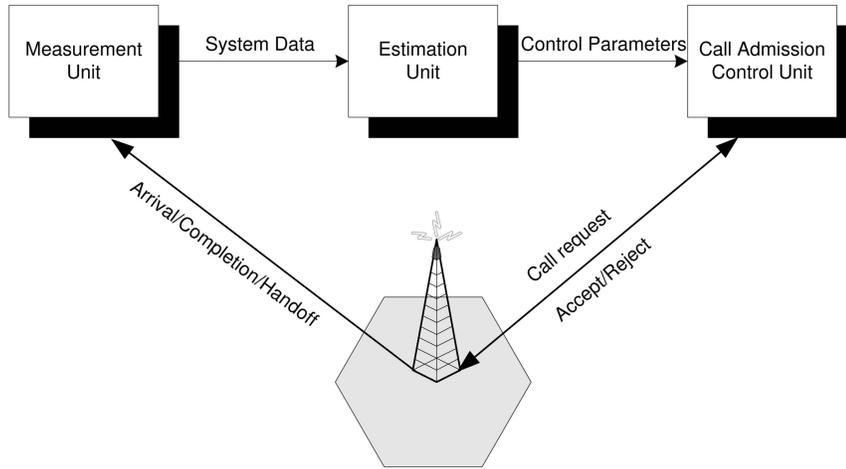
Fig. 5. Control unit diagram.

the acceptance ratio vector $a_i = (a_i^v, a_i^d)$ can be found by numerically solving (33). Function f is such that $0 \leq f(a_i^v) \leq 1$ and $f(0) = 0$. In addition, f is uniformly increasing over $[0, 1]$. The boundary condition is that $a_i \in [0, 1] \times [0, 1]$; hence, if $\tilde{P}_{f_i}$ is less than p_{QoS} even for $a_i^v = 1$, then a_i is set to $(1, f(1))$. Similarly, if $\tilde{P}_{f_i}$ is greater than p_{QoS} even for $a_i^v = 0$, then a_i is set to $(0, 0)$. In this paper, we only consider a constant service differentiation function denoted by α_i, where $a_i^d = a_i^v / \alpha_i$.

Finally, (33) can be solved using the bisection method [28]. Let ξ denote the required numerical precision. Then, the computational complexity of this technique is $O(\log 1/\xi)$, given that all mathematical operations (including exponentiation and integration) can be performed in $O(1)$.

4 CONTROL PARAMETERS

In previous sections, we assumed that several parameters are known to the admission control algorithm a priori. Among these parameters are the call arrival rates, mean call durations, mean cell residency times, and routing probabilities. In practice, all these parameters can be extracted from measured field data using an estimation technique. Measurement and estimation units are used for providing the required parameters to the admission control unit as shown in Fig. 5. One useful estimation technique is presented in the following section.

4.1 Parameter Estimation

A common technique for estimating the mean values from measurement data is the *exponentially weighted moving average* (EWMA) technique. Let z denote a control parameter to be estimated, e.g., arrival rate, and $\hat{z}$ its measured (observed) value. A moving average estimator for z at nth step is given by $z(n) = (1 - \epsilon)\,\hat{z}(n - 1) + \epsilon z(n - 1)$, where ϵ is a weighting factor that should be specified with respect to the sampled observations of z. In general, a small value of ϵ can keep track of the changes more accurately, but is too sensitive to temporary fluctuations. On the other hand, a large value of ϵ is more stable but could be too slow in adapting to real traffic changes. By using this estimator, it can be verified that $E[z] = E[\hat{z}]$. However, EFGC is independent of the estimation technique and, hence, it is possible to use more sophisticated estimation techniques to achieve more accurate estimations (refer to [29], [30]).

We use the EWMA technique to compute the new call arrival rate λ into a cell of the network. The only unknown parameter is the estimation coefficient ϵ. As mentioned before, the accuracy of the EWMA estimation depends on ϵ. The goal is to choose ϵ in such a way to minimize the estimation error. To measure the estimation error, we use the mean squared error (MSE) of the estimations as expressed by

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda(i) - \hat{\lambda}(i) \right)^2, \quad (34)$$

where N is the number of measurements. In our experiments, we found that $\epsilon = 0.8$ minimizes the prediction error and achieves sufficiently accurate predictions.

4.2 Actual New Call Arrival Rate

In Section 3.3, we used products $a_j^k \lambda_j^k$ 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 $\bar{\lambda}_j^k$, as follows:

$$\bar{\lambda}_j^k = a_j^k \lambda_j^k. \quad (35)$$

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

Let $\bar{\lambda}_j^k(n)$ denote the actual new call arrival rate into cell j during the nth control period. Also, let $N_j^k(n)$ denote the number of new calls that were accepted in cell j during the nth control period. Similar to [4], [10], an estimator for $\bar{\lambda}_j^k$ is expressed as

$$\bar{\lambda}_j^k(n + 1) = (1 - \epsilon)\frac{N_j^k(n)}{T} + \epsilon \bar{\lambda}_j^k(n), \quad (36)$$

where $\bar{\lambda}_j^k(n + 1)$ is the actual new call arrival rate into cell j at the beginning of the $(n + 1)$th control period. Note that $\bar{\lambda}_j^k(n)$ 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.

TABLE 1
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

4.3 Control Interval

The idea behind the 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 3, this minimizes the signaling overhead and operational complexity of EFGC. 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

$$r_{ij} = \begin{cases} 1/\mathcal{A}, & j \in \mathcal{A}_i, \\ 0, & j \notin \mathcal{A}_i. \end{cases} \tag{37}$$

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 can then write

$$\delta = \sum_{n=2}^{\infty} q_{ij}(n). \tag{38}$$

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

$$q(n) = \frac{(hT)^n}{n!} e^{-(h+\mu)T}. \tag{39}$$

In order to compute $q_{ij}(n)$ based on (39), 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

$$q_{ij}(n) = \frac{L_{ij}(n)}{\mathcal{A}^n}\ q(n). \tag{40}$$

Consider the network depicted in Fig. 2. Let $T = 20\ s$, $1/\mu = 180\ s$, $1/h = 100\ s$, and $\mathcal{A} = 6$. Table 1 shows the maximum probability of multiple handoffs from any cell j to cell 0, $P_{j0}(n)$, based on the number of handoffs, n. For each n, we have also determined which layer has the
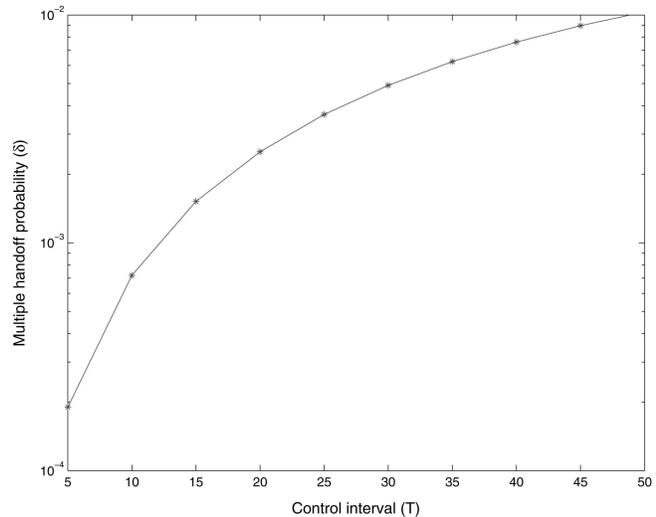
Fig. 6. Effect of T on multiple handoffs probability.

maximum paths to cell 0. Interestingly, cell 0 has the most paths to itself through other cells. We have also illustrated in Fig. 6 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

$$L_{ij}(n) \leq \mathcal{A}^{n-1}, \quad n \geq 1 \tag{41}$$

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

$$q_{ij}(n) \leq \frac{1}{\mathcal{A}} \frac{(hT)^n}{n!} e^{-(h+\mu)T}, \quad n \geq 1. \tag{42}$$

Using (38) and (42), it is obtained that

$$\begin{aligned} \delta &\leq \sum_{n=2}^{\infty} \frac{1}{\mathcal{A}} \frac{(hT)^n}{n!} e^{-(h+\mu)T} \\ &= \frac{e^{hT} - hT - 1}{\mathcal{A} e^{(h+\mu)T}}. \end{aligned} \tag{43}$$

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

$$T \leq \frac{\mathcal{A}\delta(\mu + h) + h\sqrt{2\mathcal{A}\delta}}{\mathcal{A}\delta(\mu + h)^2 - h^2}, \tag{44}$$

which finally leads to the following simple relation

$$T \approx \frac{\sqrt{2\mathcal{A}\delta}}{h}. \tag{45}$$

5 SIMULATION RESULTS

5.1 Greedy EFGC

The basic EFGC introduced in Section 3 may seem to be too conservative about accepting data calls. We refer to this restrictive version of EFGC by EFGC-REST (or simply REST). REST is a conservative approach which aims at

satisfying the specified priority function f over time. In other words, REST always uses the acceptance ratio $a_i = (a_i^v, f(a_i^v))$ regardless of the congestion situation to impose an exact priority function.

It is observed that, in some states of the system, it is possible to increase the acceptance ratio of data calls beyond the limit returned by the service differentiation function. For example, when the network is not congested (at light traffic loads), we found that by increasing the priority of data traffic, the overall utilization of the wireless bandwidth is increased while the handoff failure remains almost untouched. This relaxed version is called EFGC-UTIL (or simply UTIL) due to its greedy behavior in maximizing the bandwidth utilization. To find the data acceptance ratio in cell i, UTIL follows the following steps:

1. Find a_i^v using (33).
2. If ($a_i^v == 1$), then find the maximum value of $a_i^d \in [f(1), 1]$, which satisfies (33).

It is worth noting that the computational complexity of EFGC-UTIL is the same as EFGC-REST, i.e., $O(\log 1/\xi)$.

5.2 Simulation Parameters

Simulations were performed on a two-dimensional cellular system consisting of 19 hexagonal cells (see Fig. 2). Opposite sides wrap-around to eliminate the finite size effect. It is assumed that mobile users move along the cell areas according to a uniform routing pattern. In other words, all neighboring cells have the same chance to be chosen by a call for handoff, i.e., $r_{ji} = 1/6$. For ease of illustrating the results, the simulated system is uniform, i.e., input load is the same for every cell, although EFGC, as well as the simulation program, are designed to handle the nonuniform case as well. Therefore, unless explicitly specified, the subscript i is omitted hereafter.

The common parameters used in the simulation are as follows: All the cells have the same capacity $c = 5$ Mbps, which is equal to 160 BU assuming each BU is equal to 32 Kbps (encoded voice using ADPCM requires 32 Kbps). Target handoff failure probability for voice calls is $p_{QoS} = 0.01$ and $T = 20$ s. We use normalized load in simulations which is simply the total arrival load per BU. Let ρ denote the total normalized arrival load into a cell, then

$$\rho = \frac{1}{c}(\rho_v + \rho_d), \qquad (46)$$

where ρ_v and ρ_d are, respectively, voice and data load given by

$$\rho_v = b_v \lambda_v / \mu_v, \qquad (47)$$
$$\rho_d = b_d \lambda_d / \mu_d. \qquad (48)$$

For each load, simulations were done by averaging over eight samples, each for 10 hours of simulation time. Load distribution between voice and data traffic is fixed over time. At any load, 60 percent of the load is due to voice calls and the remaining 40 percent is composed of data calls. Table 2 summarizes service and traffic parameters for both traffic types. In this table, *priority* refers to the relative priority (service differentiation) of voice and data calls. It means that new voice calls have higher priority than data calls for the admission control algorithm. In particular, the probability of accepting a new voice call is at least twice the probability of accepting a data call (new/handoff) at any

TABLE 2
Voice/Data Service Parameters

Type	Priority	$1/\mu$ (s)	$1/h$ (s)	BU	Load
voice	1	180	100	1	60%
data	0.5	1000	800	2	40%

time and any load. Equivalently, this is achieved by setting $\alpha_{QoS} = 2$. As mentioned earlier, this relative priority can be any service differentiation function. In our simulations, for the sake of simplicity, we have chosen a constant service differentiation function.

We have also implemented the DTR scheme introduced in Section 1 for comparison purposes. Since DTR is designed for a static traffic pattern, the handoff failure probability increases rapidly with the network load when the guard channels for handoff are few, but remains too low when the guard channels are many. Here, we choose the two thresholds in such a way that DTR achieves its objectives when the network starts to get overloaded. Hence, the voice threshold is set to 155 BUs and the data threshold is set to 151 BUs. Using these thresholds at load 2, p_f and $\alpha = a_v/a_d$ were found to be 0.01 and 2, respectively.

5.3 Gaussian Approximation

When the network is not congested and each cell has only a few active calls, it is clear that Gaussian approximation is not good. However, at light loads, the admission algorithm does not require a high precision estimation of the load since there is no congestion in the network. As the load increases, the number of active calls in each cell increases rapidly until no more calls can be accepted. Due to the high capacity of broadband systems, it is expected to have enough active calls in each cell so that central limit theorem can be applied.

Other researchers have also successfully applied Gaussian approximation for similar purposes. Naghshineh and Schwartz [4] and Epstein and Schwartz [7] used the same kind of approximation. The main difference is that we extend their single point approximation at the end of the control period to a time dependent approximation over the whole control period. Wu et al. [10] also realized that, for large system sizes, as is the case in this paper, the cell occupancy distribution evolves into a Gaussian distribution. Investigating the tail behavior of the bandwidth usage distribution is beyond the scope of this paper; instead, we rely on the results from other researchers [4], [7], [10], [23].

5.4 Results and Analysis

5.4.1 Effect of Arrival Load

The first set of simulation results show the main performance parameters of EFGC. Fig. 7 shows the handoff failure probability for the three schemes for a wide range of loads. Both UTIL and REST maintain a constant failure probability independent of the load. For DTR, it grows very rapidly with the load (which was expected). With light loads (load $<$ 2), DTR and REST have almost the same handoff failure probability while UTIL has slightly higher handoff failure probability. But, with high loads (load $>$ 2), UTIL and

Fig. 7. Voice handoff failure probability.

Fig. 8. Voice/Data relative acceptance probability (α).

REST converge to exactly the same handoff failure probability while DTR has much higher handoff failure probability. Fig. 9b shows that, although REST has better failure probability in light loads, this is accomplished at the expense of the data call blocking probability. However, even in this region (load < 2), UTIL satisfies the target handoff failure probability p_{QoS}.

One of the objectives of EFGC is to maintain the relative service priority between voice and data calls. In our simulations, this relative priority is fixed and indicates that the acceptance probability of new voice calls should be twice the acceptance probability of new data calls. Fig. 8 gives the service differentiation $\alpha = a_v/a_d$ for different loads. It shows that EFGC maintains an almost constant service priority between the two types of traffic. More precisely, REST preserves $\alpha = 2$ for the whole range of loads while UTIL has $\alpha = 1$ in light loads and $\alpha = 2$ in high loads as expected. This can be explained by the fact that, in light loads, UTIL accepts data calls as long as there is free

bandwidth (without violating the target voice handoff failure probability). As the load increases, service priority of DTR increases rapidly. Fig. 9b shows that, at high loads, almost no data calls are accepted. In other words, DTR is not fair and leads to starvation of data traffic. It is worth mentioning that, although, in this simulation, the service differentiation is fixed, the EFGC can satisfy more complex priority disciplines such as state dependent priorities.

Fig. 9 shows the new voice and new/handoff data call acceptance probabilities, respectively. Again for high loads, UTIL and REST converge to the same result but the difference in their performance at light loads is significant. For data traffic at light loads, the acceptance probability of UTIL is almost twice that of REST. This explains why the utilization of UTIL is superior to REST. It can be seen that DTR has slightly higher acceptance probability for voice but much lower acceptance probability for data in comparison to UTIL and REST.
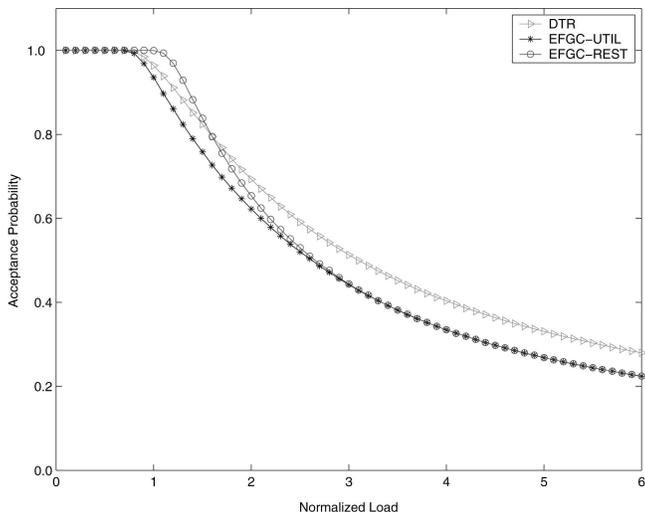
(a)
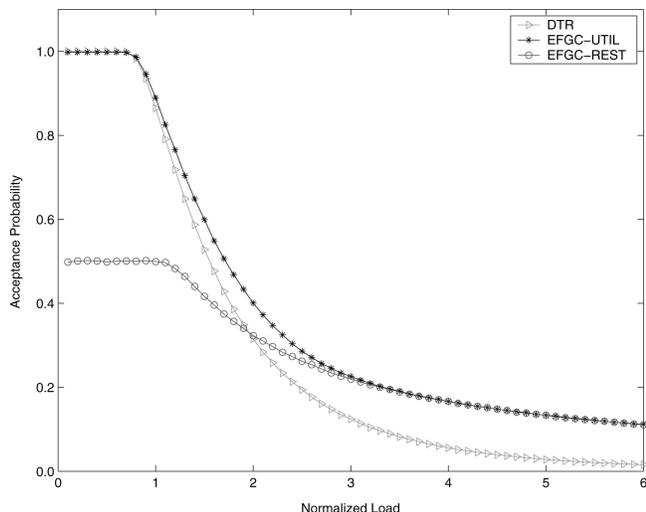
(b)

Fig. 9. Acceptance probability of voice and data. (a) New voice calls acceptance probability. (b) New/handoff data calls acceptance probability.
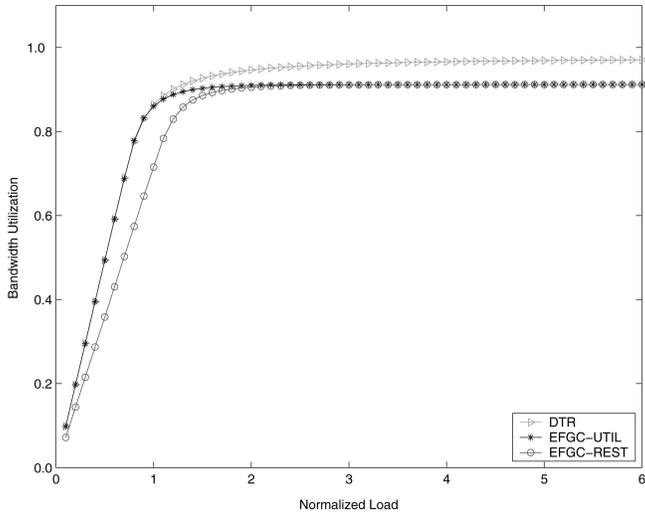
Fig. 10. Wireless bandwidth utilization.

Finally, Fig. 10 shows the wireless bandwidth utilization under the three bandwidth allocation mechanisms. Although DTR performs poorly in terms of handoff failure probability and service priority, its utilization is slightly better than EFGC. Interestingly, UTIL has exactly the same utilization level as DTR at light loads but higher than that of REST. In this simulation, voice traffic constitutes the larger portion of the total load. As the percentage of data traffic increases, the utilization of DTR is expected to drop. This will be investigated next.

5.4.2 Effect of Load Sharing

In previous simulations, the load sharing factor $\beta(\beta > 0)$ is set to 1.5, where

$$\beta = \frac{\text{arriving data traffic load}(\rho_v)}{\text{arriving voice traffic load}(\rho_d)}. \tag{49}$$
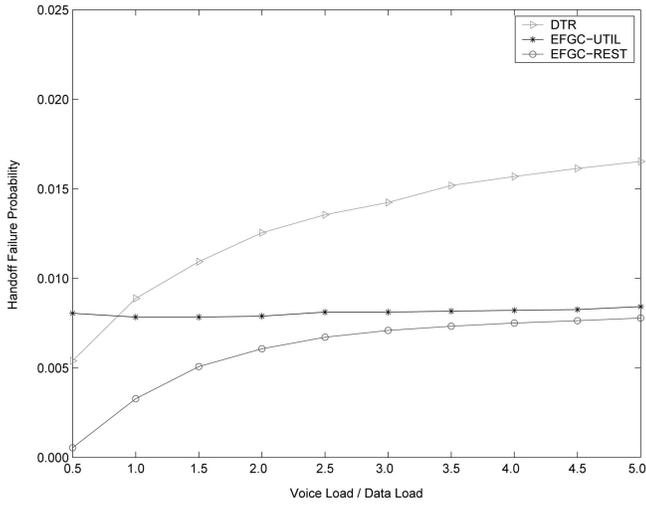
(a)
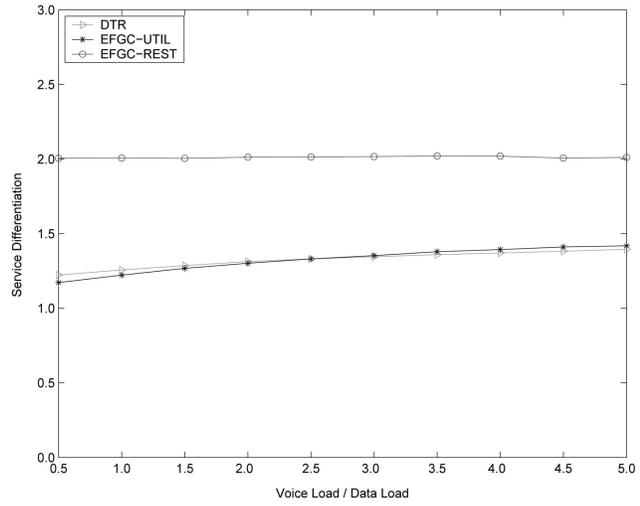
(b)
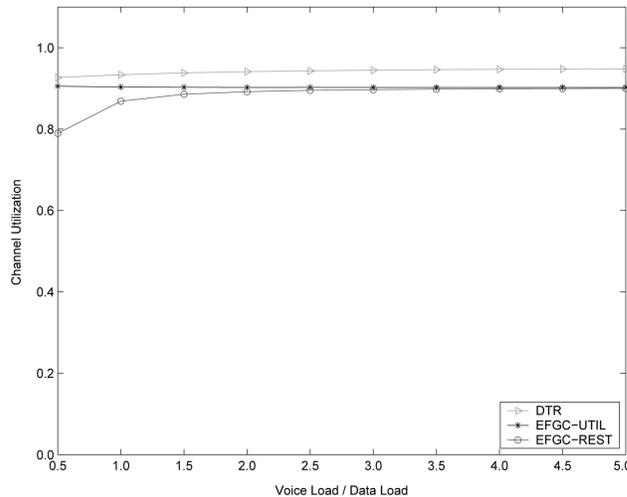
(c)

Fig. 11. Effect of load sharing (β). (a) Voice handoff failure probability. (b) Relative acceptance probability (α). (c) Wireless bandwidth utilization.

(a)

(b)

(c)
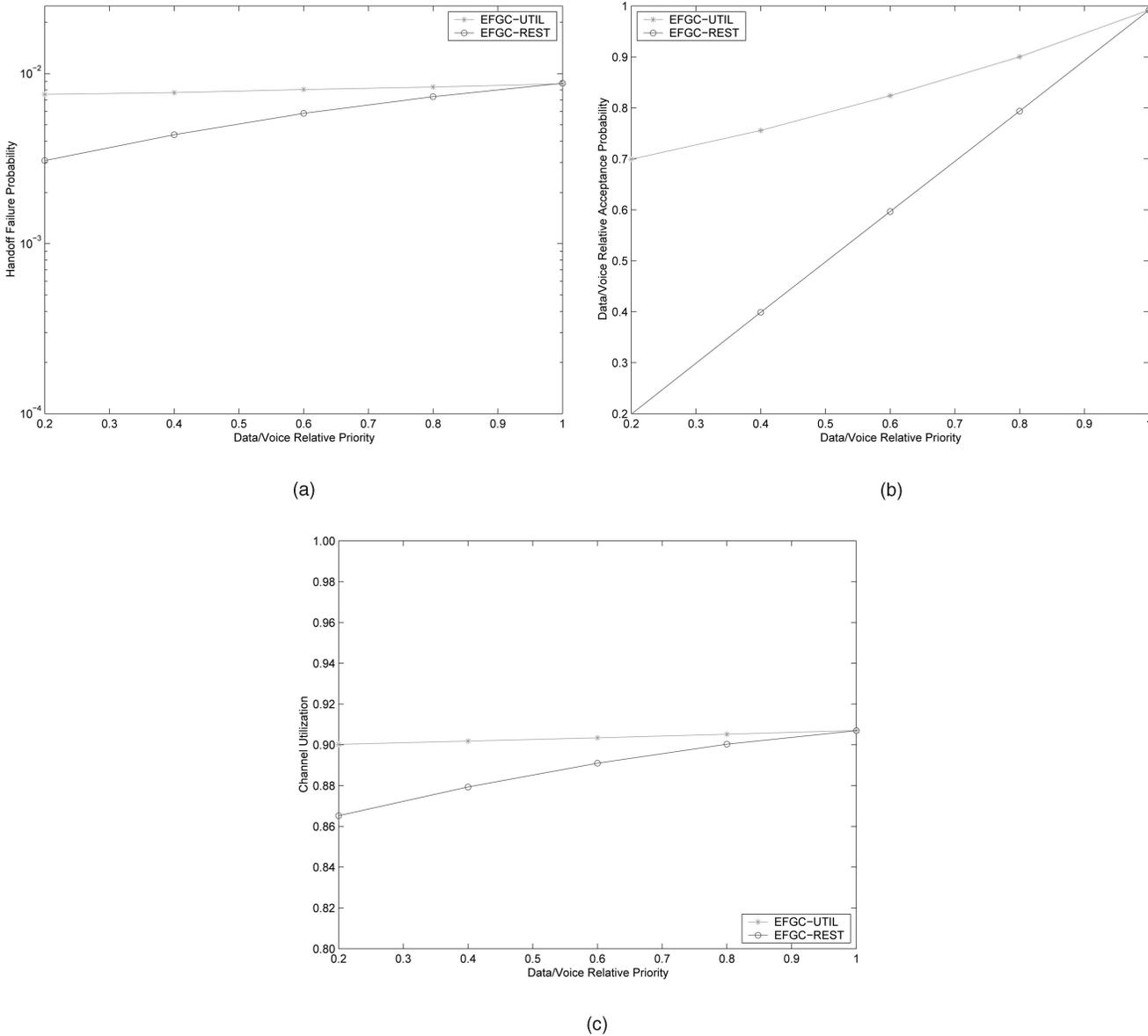
Fig. 12. Effect of voice priority. (a) Voice handoff failure probability. (b) Relative acceptance probability ($1/\alpha$). (c) Wireless bandwidth utilization.

Due to the priority of voice calls over data calls, varying β will affect the behavior of EFGC. As shown in Fig. 11, EFGC is insensitive to the load sharing factor. In these plots, the X axis indicates the load sharing factor β. It is assumed that most of the traffic is composed of voice calls, hence, β varies between 0.5 and 5.

For this set of simulations, normalized arrival load is set to 1.5 Erlang and voice priority is set to 2 ($\alpha = 2$). As expected, DTR is not able to adjust to changes in load shares, although the total load is fixed. Interestingly, as β increases, EFGC-UTIL and EFGC-REST converge to the same value for handoff failure probability. The reason is that, by increasing β, voice traffic will dominate data traffic. Therefore, a larger portion of the available bandwidth is allocated to voice traffic in such a way that there is no extra free bandwidth to be assigned to data traffic (more than their guaranteed share).

The primary goal of the following set of simulations is to show the stability of EFGC under various QoS requirements

(p_{QoS} and α_{QoS}) and the insensitivity of EFGC to the exponential assumption we made about the cell residency time.

5.4.3 Effect of Voice Priority

Fig. 12 shows the effect of changing the relative priority of data calls and voice calls. In this set of plots, the X axis indicates the quantity $1/\alpha$, where

$$1/\alpha = \frac{\text{data calls acceptance probability } (a_d)}{\text{voice calls acceptance probability } (a_v)}. \qquad (50)$$

In the simulations, the total arrival load is set to 1.5 Erlang, which consists of 60 percent voice traffic and 40 percent data traffic (i.e., a load sharing factor of 1.5). It is found that, regardless of α, EFGC is able to satisfy the target α_{QoS} while providing the desired service differentiation. The straight lines in Fig. 12b indicate that any value of service differentiation can be strictly guaranteed with EFGC.

(a)

(b)

(c)
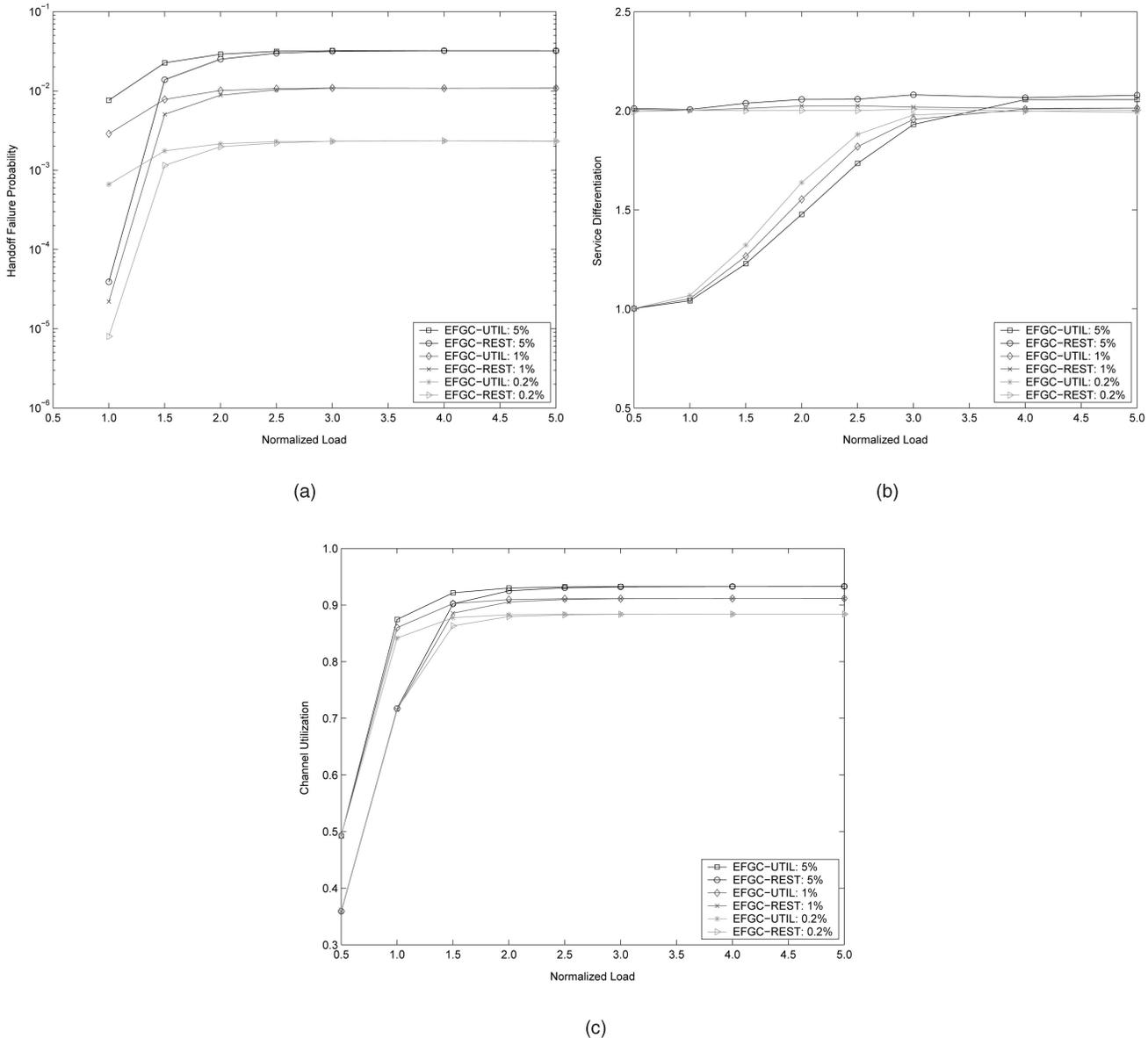
Fig. 13. Effect of handoff failure probability (QoS). (a) Voice handoff failure probability. (b) Relative acceptance probability (α). (c) Wireless bandwidth utilization.

As indicated in these figures, UTIL and REST converge to the same control policy as α tends toward 1. This was expected because the two schemes differ from each other with respect to α. In this case, available resources are completely shared among voice and data traffic and channel utilization is maximized. However, for large values of α (small values of $1/\alpha$), UTIL has a superior performance over REST. For example, at $\alpha = 1/0.2$, UTIL has 4 percent better utilization.

5.4.4 Effect of Handoff Failure Probability (QoS)

In cellular systems, the target p_{QoS} is typically set to 1 percent. To show the adaptiveness of EFGC, simulations were performed for $p_{\mathrm{QoS}} = [0.2 \text{ percent}, 1 \text{ percent}, 5 \text{ percent}]$. Notice that $p_{\mathrm{QoS}} = 0.2$ percent is an extremely low handoff failure probability. As shown in Fig. 13, handoff failure and service differentiation are fully satisfied regardless of the target

QoS requirements. In particular, Fig. 13a shows the stability of EFGC under different target dropping requirements.

5.4.5 Effect of Nonexponential Cell Residency

The first part of our analysis, which gives the equations describing the mean and variance of channel occupancy (i.e., number of busy channels in a cell), is based on the exponential cell residency time assumption. This assumption may not be correct in practice and needs more careful investigation as pointed out in [31], [32], [33] and references there in. Although exponential distributions are not accurate in practice, the models based on the exponential assumption are tractable and do provide mean value analysis which indicates the system performance trend.

Using real measurements, Jedrzycki and Leung [31] 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

(a)

(b)

(c)
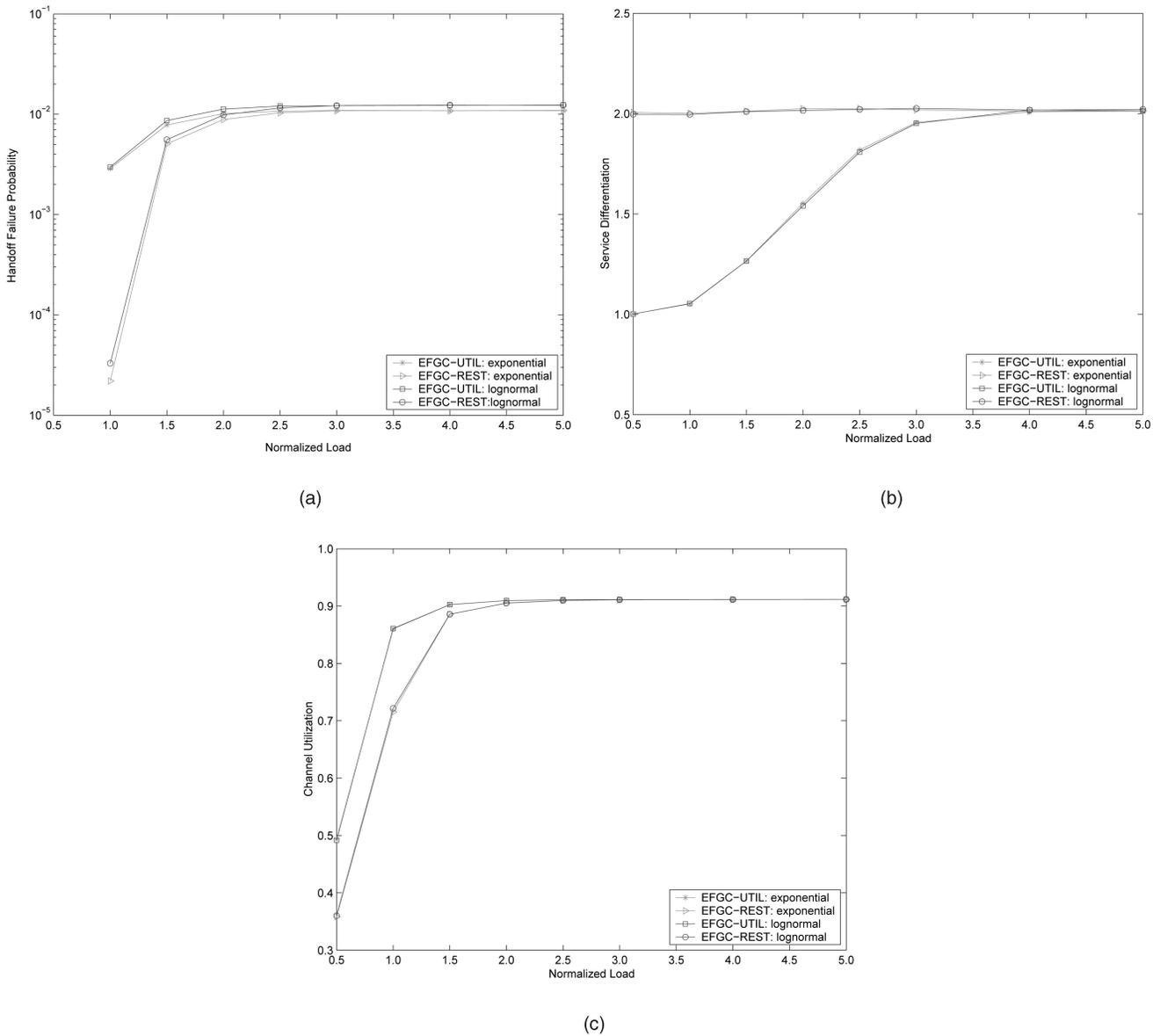
Fig. 14. Effect of nonexponential cell residency. (a) Voice handoff failure probability. (b) Relative acceptance probability (α). (c) Wireless bandwidth utilization.

under more realistic lognormal distribution. The mean and variance of both distributions are the same (refer to Table 2). Fig. 14 shows 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 less than the average cell residency time.

6 CONCLUSION

In this paper, we proposed a new admission control algorithm for voice/data integration in broadband wireless networks. Our algorithm is a natural extension of the well-known fractional guard channel proposed for voice cellular systems. EFGC always achieves the predetermined call dropping probability for voice calls while keeping the relative blocking probability of voice and data calls within a target threshold. We then described two versions of the

EFGC, namely, EFGC-UTIL and EFGC-REST. EFGC-UTIL follows a greedy approach to maximize the bandwidth utilization while EFGC-REST maintains the relative service priority. Both versions converged to the same result for high traffic loads. The major advantage of EFGC is its insensitivity to network traffic load. The dropping probability of voice calls and relative blocking probability of voice and data calls is maintained at a stable level over a wide range of traffic loads. From the simulation results, we conclude that EFGC-UTIL is a better candidate for integrated voice/data cellular networks.

We are currently investigating the case of multiple classes of traffic where each class has its own QoS requirements in terms of call blocking and dropping probabilities. EFGC can readily support multiple classes of traffic by assigning a separate acceptance ratio to each class. However, computing these acceptance ratios in order to satisfy the desired QoS is not trivial.

ACKNOWLEDGMENTS

REFERENCES

[1] U. Varshney and R. Jain, "Issues in Emerging 4G Wireless Networks," *Computer,* vol. 34, no. 6, pp. 94-96, June 2001.

[2] S.Y. Hui and K.H. Yeung, "Challenges in the Migration to 4G Mobile Systems," *Computer,* vol. 41, no. 12, pp. 54-59, Dec. 2003.

[3] D. Hong and S.S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures," *IEEE Trans. Vehicular Technology,* vol. 35, no. 3, pp. 77-92, Aug. 1986, see also: CEAS Technical Report No. 773, College of Eng. and Applied Sciences, State Univ. of New York, June 1999.

[4] M. Naghshineh and M. Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks," *IEEE J. Selected Areas in Comm.,* vol. 14, no. 4, pp. 711-717, May 1996.

[5] 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.

[6] S. Choi and K.G. Shin, "Predictive and Adaptive Bandwidth Reservation for Handoffs in QoS-Sensitive Cellular Networks," *Proc. ACM SIGCOMM '98,* vol. 27, pp. 155-166, Oct. 1998.

[7] B.M. Epstein and M. Schwartz, "Predictive QoS-Based Admission Control for Multiclass Traffic in Cellular Wireless Networks," *IEEE J. Selected Areas in Comm.,* vol. 18, no. 3, pp. 523-534, Mar. 2000.

[8] A. Aljadhai and T.F. Znati, "Predictive Mobility Support for QoS Provisioning in Mobile Wireless Networks," *IEEE J. Selected Areas in Comm.,* vol. 19, no. 10, pp. 1915-1930, Oct. 2001.

[9] 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.

[10] 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.

[11] 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.

[12] B. Epstein and M. Schwartz, "Reservation Strategies for Multi-Media Traffic in a Wireless Environment," *Proc. IEEE Vehicular Technology Conf.,* vol. 1, pp. 165-169, July 1995.

[13] J.E. Wieselthier and A. Ephremides, "Fixed- and Movable-Boundary Channel-Access Schemes for Integrated Voice/Data Wireless Networks," *IEEE Trans. Comm.,* vol. 43, no. 1, pp. 64-74, Jan. 1995.

[14] M.C. Young and Y.-R. Haung, "Bandwidth Assignment Paradigms for Broadband Integrated Voice/Data Networks," *Elsevier Computer Comm. J.,* vol. 21, no. 3, pp. 243-253, 1998.

[15] H.-H. Liu, J.-L.C. Wu, and W.-C. Hsieh, "Delay Analysis of Integrated Voice and Data Service for GPRS," *IEEE Comm. Letters,* vol. 6, no. 8, pp. 319-321, Aug. 2002.

[16] D.-S. Lee and C.-C. Chen, "QoS of Data Traffic with Voice Handoffs in a PCS Network," *Proc. IEEE GLOBECOM '02,* vol. 2, pp. 1534-1538, Nov. 2002.

[17] M.A. Marsan, P. Laface, and M. Meo, "Packet Delay Analysis in GPRS Systems," *Proc. IEEE INFOCOM '03,* vol. 2, pp. 970-978, Mar. 2003.

[18] Y.-R. Haung, Y.-B. Lin, and J.-M. Ho, "Performance Analysis for Voice/Data Integration on a Finite-Buffer Mobile System," *IEEE Trans. Vehicular Technology,* vol. 49, no. 2, pp. 367-378, Mar. 2000.

[19] L. Yin, B. Li, Z. Zhang, and Y.-B. Lin, "Performance Analysis of a Dual-Threshold Reservation (DTR) Scheme for Voice/Data Integrated Mobile Wireless Networks," *Proc. IEEE Wireless Comm. & Networking Conf.,* vol. 1, pp. 258-262, Sept. 2000.

[20] H. Wu, L. Li, B. Li, L. Yin, I. Chlamtac, and B. Li, "On Handoff Performance for an Integrated Voice/Data Cellular System," *Proc. IEEE Int'l Symp. Personal Indoor and Mobile Radio Comm.,* vol. 5, pp. 2180-2184, Sept. 2002.

[21] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a Cutoff Priority Cellular Radio System with Finite Queueing and Reneging/Dropping," *IEEE/ACM Trans. Networking,* vol. 2, no. 2, pp. 166-175, Apr. 1994.

[22] B. Li, L. Li, B. Li, and X.-R. Cao, "On Handoff Performance for an Integrated Voice/Data Cellular System," *ACM/Kluwer Wireless Networks,* vol. 9, no. 4, pp. 393-402, July 2003.

[23] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks," *IEEE J. Selected Areas in Comm.,* vol. 9, no. 7, pp. 968-981, 1991.

[24] F.P. Kelly, "Notes on Effective Bandwidths," *Stochastic Networks: Theory and Applications,* F.P. Kelly, S. Zachary, and I. Ziedins, eds., pp. 141-168, Oxford, UK: Oxford Univ. Press, 1996.

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

[26] Y. Fang, I. Chlamtac, and Y.-B. Lin, "Channel Occupancy Times and Handoff Rate for Mobile Computing and PCS Networks," *IEEE Trans. Computers,* vol. 47, no. 6, pp. 679-692, June 1998.

[27] Y. Fang and I. Chlamtac, "Analytical Generalized Results for Handoff Probability in Wireless Networks," *IEEE Trans. Comm.,* vol. 50, no. 3, pp. 396-399, Mar. 2002.

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

[29] P. Whittle, *Prediction and Regulation by Linear Least-Square Methods,* second ed. Minnesota: Univ. of Minnesota Press, 1983.

[30] P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods,* second ed. New York: Springer-Verlag, 1991.

[31] C. Jedrzycki and V.C.M. Leung, "Probability Distribution of Channel Holding Time in Cellular Telephone Systems," *Proc. IEEE Vehicular Technology Conf.,* vol. 1, pp. 247-251, May 1996.

[32] M.M. Zonoozi and P. Dassanayake, "User Mobility Modeling and Characterization of Mobility Patterns," *IEEE J. Selected Areas in Comm.,* vol. 15, no. 7, pp. 1239-1252, Sept. 1997.

[33] R. Guerin, "Channel Occupancy Time Distribution in a Cellular Radio System," *IEEE Trans. Vehicular Technology,* vol. 35, no. 3, pp. 89-99, 1987.

Majid Ghaderi received the BS and MS degrees in computer engineering from the Sharif University of Technology, Iran, in 1999 and 2001, respectively. He is currently a PhD student in the David R. Cheriton School of Computer Science at the University of Waterloo, Canada. His research interests are in the area of mobile communication systems, including radio resource management, admission control, and quality of service.

Raouf Boutaba is currently an associate professor in the David R. Cheriton School of Computer Science at the University of Waterloo, Canada. Before that, he was with the Department of Electrical and Computer Engineering at the University of Toronto, Canada. Before joining academia, he founded and was the director of the telecommunications and distributed systems division of the Computer Science Research Institute of Montreal (CRIM). Dr. Boutaba conducts research in the areas of network and distributed systems management and resource management in multimedia wired and wireless networks. He has published more than 120 papers in refereed journals and conference proceedings. He was the recipient of the Premier's Research Excellence Award, a fellow of the faculty of mathematics of the University of Waterloo, a senior member of the IEEE, and a distinguished lecturer of the IEEE Computer Society. He is the chairman of the IFIP Working Group on Networks and Distributed Systems, the vice chair of the IEEE Communications Society Technical Committee on Information Infrastructure, and the chair of the IEEE Communications Society Committee on Standards. He is the founder and editor-in-chief of the *IEEE ComSoc eTransactions on Network and Service Management*, on the advisory editorial board of the *Journal of Network and Systems Management*, on the editorial board of the *KIKS/IEEE Journal of Communications and Networks*, and on the editorial board of the *Journal of Computer Networks*.