Wireless Pers Commun (2007) 43:857–878
DOI 10.1007/s11277-007-9307-y

Call admission and handoff control in multi-tier cellular networks: algorithms and analysis

Vijoy Pandey · Dipak Ghosal · Biswanath Mukherjee · Xiaoxin Wu

Received: 24 November 2006 / Accepted: 3 April 2007 / Published online: 6 June 2007

Abstract In this paper, we investigate a call-admission and handoff-control framework for multi-tier cellular networks. We first propose and compare Call-Admission Control (CAC) algorithms based on the cell-dwelling time, by studying their impact on the handoff-call dropping and new-call blocking probabilities and the channel partitioning between the two tiers. Our results show that a simple, cell-dwelling-time-insensitive algorithm performs better under various mixes of user mobilities and call types. Moreover, there is an optimal channel partition of the overall spectrum between the tiers which minimizes the dropping and blocking probabilities for the two different CAC algorithms studied in this paper. Once the call is admitted into the network, we propose and compare various handoff-queuing strategies to reduce the call dropping probability. We show that implementing a queuing framework in one of the tiers (especially the upper, i.e., macrocellular, tier), results in a significant reduction in the dropping probability.

Keywords Call admission and handoff control · Multi-tier cellular networks · Dropping probability · Handoff queuing · Analytical models

V. Pandey · D. Ghosal (✉)· B. Mukherjee
Department of Computer Science, University of California, Davis, CA 95616, USA
e-mail: ghosal@cs.ucdavis.edu

B. Mukherjee
e-mail: mukerje@cs.ucdavis.edu

V. Pandey
Blade Network Technologies, 2350 Mission College Blvd, Santa Clara, CA 95054, USA
e-mail: vijoy@bladenetwork.net

X. Wu
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
e-mail: xiaoxin.wu.@intel.com

1 Introduction

Future cellular networks will require improved quality of service (QoS), higher capacity, and a larger coverage area than existing networks. There are two key objectives in cellular network design—(1) to maximize the spectrum efficiency, and (2) to provide high Quality of Service (QoS) to users, i.e., minimize the handoff-call dropping probability and the new-call blocking probability, hereafter referred to as (call) dropping and (call) blocking probabilities, respectively.

1.1 Background

In order to increase the cellular network's capacity, we can employ a finer mesh of smaller cells (i.e., microcells) over areas with a large population of users in order to achieve higher channel reuse. On the other hand, to be able to cover a larger area and serve a large number of highly mobile hosts, we should increase the cell size.

To achieve the first objective, network designers employ efficient channel-allocation schemes based on Fixed Channel Allocation (FCA), Dynamic Channel Allocation (DCA), or hybrid schemes [1,2]. The key idea in these schemes is to achieve good load balancing of the total system capacity. To provide better QoS to users, we can employ a finer mesh of smaller cells (i.e., microcells) over areas with a large population of users in order to achieve higher channel reuse. On the other hand, to be able to cover a larger area and serve a large number of highly mobile hosts, we should increase the cell size.

A multi-tier cellular network architecture in which cells in a particular tier (or layer) are overlayed by larger cells at the next higher tier (or layer) [3–11] can satisfy these various, and sometimes conflicting, requirements. Throughout this paper, we have analyzed a two-tier cellular architecture (because it is the first and most natural extension beyond a flat single-tier network), but our results, observations and conclusions apply to multi-tier architectures as well.

1.2 Challenges

A typical two-tier architecture consists of a tier of microcells and a tier of macrocells such that a number of contiguous small microcells are overlayed by a large macrocell. Microcells are used to achieve higher capacity, while macrocells provide a larger coverage area and reduce overheads due to handoffs. In order to achieve these advantages, it is necessary to address three key engineering issues: (1) to develop algorithms to efficiently allocate channels to the different tiers, (2) to design effective new-call and handoff-call admission control algorithms, and (3) to design adaptive QoS control schemes for calls admitted into the system.

The goal of call-admission control in a two-tier cellular network is to determine which calls should be admitted and to subsequently assign an incoming new call or a handoff call to an appropriate tier. The goal is to minimize handoff traffic, improve QoS by minimizing call drops, while simultaneously taking advantage of the microcell layer by maximizing the number of users in the network.

Advances in accurate positioning technologies [12–15] and velocity-estimation methods [16,17] allow these attributes of a call to be determined with high accuracy. Furthermore, it is also possible to determine the call type and make reasonable predictions about the call duration and the resource requirements. A key question, which we have explored in this work, is how much benefit can be gained by exploiting these information in assigning a call to a tier.

1.3 Contributions of this paper

In this paper, we study the problem of call admission and handoff control for a multitier cellular network. We analyze both new call admission and handoff call management and control in a single unified framework, which, to the best of our knowledge, has not been done before. For this study, we propose call admission and handoff control algorithms that incorporate the attributes of the call—such as data calls or voice calls, long calls or short calls, fast or slow user mobility—to make call management decisions.

In the first part of this study, we consider a simple First-Come-First-Served (FCFS) admission scheme and study the call-assignment problem. We analyze the performance of a Dwell-time-based Call-Admission Control (DCAC) algorithm which assigns calls to tiers based on their predicted cell-dwelling times. In particular, calls with long holding time and/or high mobility which may generate a large number of handoffs in the microcell layer are assigned to the macrocell layer. We compare the performance of DCAC with a simpler Uniform Call-Admission Control (UCAC) algorithm which handles all calls uniformly.[1] This comparison is done with respect to the new-call blocking and handoff-call dropping probabilities across different channel partitions between the tiers (and hence different system capacities). Our results show that UCAC algorithm outperforms the DCAC algorithm for various traffic mixes in terms of user mobility and call-holding time. Furthermore, there is an optimal channel partition which minimizes the new-call blocking and handoff-call dropping probabilities for both admission control algorithms.

Once a new or handoff call has been assigned to a particular tier, an effective QoS control scheme would attempt to minimize the probability of a drop. This can be achieved by implementing a queuing mechanism which queues a handoff call if there are no channels available in the target cell. In effect, this provides a method of giving priority to handoff calls over new

[1] The DCAC and UCAC algorithms are described in Sect. 3.

calls. It has been shown that, for single-tier cellular networks, such queuing schemes can significantly reduce the handoff-call dropping probability [18].

In this paper, we extend the above studies by investigating the performance of different handoff-queuing strategies for a two-tier cellular network. We consider various queue architectures—queue at the microcell layer, queue at the macrocell layer, queue at both layers, and priority queues for different types of calls. We present a detailed performance comparison through analysis and simulation. Our results show that implementing a queue in only one of the tiers can improve the dropping probability significantly. Among the different queuing strategies studied, we find that implementing the queue in the macrocell layer results in a significant reduction in dropping probability for both tiers in the network while keeping the cost low and the operation of the system simple.

In Sect. 2, we describe the two-tier cellular network examined in this paper. Section 3 outlines the two call-admission control algorithms, DCAC and UCAC, and provides a description of the network model used for the study. The performance results are discussed in Sect. 3.4. In Sect. 4, we describe our call-handoff-control framework and enumerate the different queue architectures. In Sect. 4.1, we develop a mathematical model to analyze the performance of the various queue architectures. In Sect. 4.4, we discuss the analytical and simulation results. Finally, in Sect. 5, we conclude with a summary of our results and future research directions.

2 Two-tier cellular network architecture

We consider a two-tier cellular network based on the Global System for Mobile Communications (GSM) architecture [19].[2] We consider a homogeneous network in which each macrocell covers N microcells. We assume a total of C channels in the network. We allocate m channels to each microcell and c channels to each macrocell. Note that, $c = \lfloor (C - m \times R_m)/R_u \rfloor$, where R_u and R_m denote the macrocell and microcell reuse distance ratios, respectively. The total capacity of the network, denoted by κ, is defined to be the maximum number of calls the network can support at any given instant of time. The microcell layer can support a maximum of $M \times N \times m$ calls while the macrocell layer can support a maximum of $M \times c$ calls. Thus,

$$\kappa = M \times (Nm + c). \quad (1)$$

From the above equation, we see that the total network capacity can be increased by increasing m, i.e., by allocating more channels to the microcell layer. However, allocating all the channels to the microcell layer increases the handoff traffic when there are users with different mobility characteristics. The proper choice of m and c is referred to as the *frequency plan*, which is determined by many factors such as the reuse distances of the different layers, call-arrival rate, and characteristics of the calls in terms of call-holding time as well as user mobility. The frequency plan can be static, in which case the frequency assignments are based on long-term statistical averages of the above input parameters. In case of dynamic frequency plans, the assignment can change with, say, time of day. Given a frequency plan, it is necessary to develop a call-admission control algorithm that will determine which layer the call should be admitted in order to maximize the network's resource utilization and minimize the dropping probability. In the following section, we address this issue.

3 Call-admission control in two-tier cellular networks

The cell resources consumed by a call is proportional to the call's cell-dwelling time, namely, the time for which the call remains in the cell. The consumed resources are determined by two parameters—(1) call-holding time and (2) the user's cell-dwelling time. The latter depends on the velocity of the user and the cell size. In this study, we consider two different types of calls characterized by their call-holding times—voice calls have shorter call-holding time and data calls have longer call-holding time. We will refer to them as short and long calls, respectively. We also consider two types of user mobility—fast users and slow users, referred to as fast calls and slow calls, respectively. The four unique call types have different cell-dwelling times.

3.1 Uniform call-admission control (UCAC) algorithm

In this scheme, all calls are treated identically, where all incoming calls are admitted into the microcell layer

[2] Even though we state the GSM architecture for illustration purposes, we do not use any of its specifics to generate our results. Hence, our results are applicable to any multitier cellular network architecture.

first. The rationale behind this algorithm is that microcells offer greater channel reuse, and hence increase the capacity of a cellular network. Therefore, it would make sense to utilize this increased capacity and attempt to assign all calls to the microcell layer.

The key steps of the algorithm are as follows.

- A new call is first attempted in a microcell. If all channels are busy, the call overflows to the corresponding macrocell. If all the channels in the macrocell are also busy, the call is *blocked*.
- A handoff call is first attempted at the microcell layer. If all the channels are busy, the call overflows to the corresponding macrocell where it can get *dropped* if there is no available channel.
- When there is a handoff at the macrocell layer, the call is handed off to the appropriate microcell at the periphery of the adjacent macrocell.

3.2 Dwell-time-based call-admission control (DCAC) algorithm

For this algorithm, we assume that it is possible to determine both the type of the incoming call (short or long) and the mobility of the user (slow or fast). Calls with long holding time and/or fast mobility will generate more handoffs if they are assigned to the microcell layer. This can increase the new-call blocking and handoff-call dropping probabilities. On the other hand, if these calls are assigned to the macrocell layer, while the handoff rate will decrease, it will increase the cell-dwelling time. This, in turn, may lead to higher blocking and dropping probabilities. To investigate this tradeoff, we consider a call-admission algorithm where calls with long holding time and generated by fast users are always first attempted at the macrocell layer. The key steps of this algorithm are as follows:

- Slow new or handoff calls follow the UCAC algorithm (i.e., they always first attempt the microcell layer).
- Fast new calls always first attempt the macrocell layer. If no channel is available in the macrocell, the call overflows to the appropriate underlying microcell. If all channels in this microcell are also busy, the call is *blocked*.
- When there is fast-call handoff at the microcell layer, the first attempt is made at the macrocell corresponding to the target microcell. If the handoff is at the macrocell layer, the first attempt is made at the adjacent macrocells. In either case, if no channel is available at the target macrocell, the call overflows to the appropriate underlying microcell. If no channel is available in the microcell as well, then the call is *dropped*.

Given a two-tier cellular network with a limited number of channels and four types of calls, the problem is to analyze whether DCAC algorithm provides lower blocking and dropping probabilities compared to the UCAC algorithm. While studying this problem, we attempt to find an optimal orthogonal assignment[3] of the total available channels in the network, which provide the best operating region in terms of the lowest blocking and dropping probabilities.

3.3 Model assumptions

The arrivals of new fast and slow calls are drawn from a Poisson process with parameter λ (calls/s). We use F and $1 - F$ to denote the fraction of fast and slow calls in the network, respectively.

Mobility is modeled using an uniform fluid-flow approximation. Under this model, the rate at which a mobile crosses a microcell boundary is given by $\alpha = VL/\pi S$, where L is the perimeter of a microcell, V is the average velocity of the user, and S is the microcell area [20]. The cell-dwelling time of a mobile in a microcell is exponentially distributed with the mobility parameter α (cell crossings/s). In macrocells, the mobility parameter turns out to be $\alpha_u = \alpha/\sqrt{N}$.

When a mobile user moves from a cell (A) to a new cell (B), the resources in the old cell must be cleared before they can be reused by other new or handoff calls. These resource-allocation and clearance actions require processing which increases with the number of handoff. This causes a call's channel-holding time in a cell to increase. To incorporate this resource-management overhead, we model the BSC/MSC[4] signaling

[3] Orthogonal sharing is a partitioning of the total available channels without reuse *across* the tiers (i.e., microcells cannot reuse channels used by the macrocell layer and vice versa).

[4] BSC is the acronym for Base Station Controller and MSC is the acronym for Mobile Switching Center. These are network elements in the cellular network architecture and perform signaling and switching functions involved in call origination and delivery, handoff, and mobility management.

centers by a *handoff controller* which manages handoffs between a number of macrocells and their underlying microcells. This handoff controller is modeled as an infinite buffer queue with a single server with an exponentially distributed service time with parameter μ_c. This is shown in Fig. 1.

We assume that each handoff controller handles handoff requests from the M macrocells as well as their underlying microcells. A call requesting handoff occupies a channel in its old cell throughout the time it spends in the controller queue. Once the handoff request is serviced by the controller, the channel in the old cell is released for reuse by other (new and handoff) calls. The infinite-buffer approximation at the controller is used as a simplification to study the effects of the queuing delay on the call blocking and dropping probabilities in the network.

The total (new-call plus handoff-call) call-holding time is assumed to have an exponential distribution with a mean of $1/\mu$ seconds. The cell-dwelling time of a call in a microcell is therefore exponentially distributed with parameter $\alpha + \mu$. There are two ways in which a call exits a microcell: (1) there is a handoff with probability $\theta = \alpha/(\alpha + \mu)$ and (2) the call completes with probability $1 - \theta$. The above holds for a macrocell as well, with α replaced by α_u. For the above assumptions and equations, the reader is asked to replace μ with μ_v for short calls, μ with μ_d for long calls, α with α_s for slow calls, and α with α_f for fast calls.

We have used the following default parameter values for the analytical and simulation results shown later in this study: (1) reuse distance ratio = 3 for both macrocells and microcells (i.e., $R_u = R_m = 3$) and (2) total number of channels available (C) in the network is 60. The mean user mobility for a slow user is assumed to be 3.5 kmph (2.25 mph), which is the average walking speed of a human being. For a fast user, the mean mobility is assumed to be 35 mph (56 kmph), which is the average automobile speed in a city business area. The microcell radius is assumed to be 200 m, and $N = 12$. Therefore, $\alpha_s = 0.00367647$ (cell crossings/s) for slow users, and $\alpha_f = 0.05718954$ (cell crossings/s) for fast users. For short calls, a mean call-holding time of 2 min is used, while long calls are assumed to hold for 20 min, on average. Each simulation experiment was conducted for 10,000,000 new-call arrivals. The sensitivity of the results to variations in the default parameter values are discussed in Sect. 3.4.3.

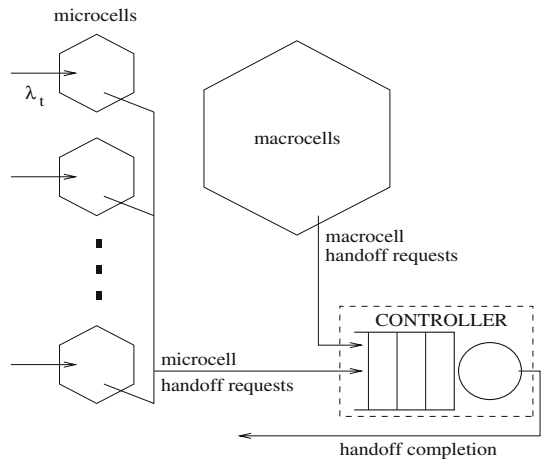


Fig. 1 Model of the network with handoff overheads

The comparison is based on the following performance metrics.

- *Blocking Probability:* Total blocking probability is defined as the ratio of the total new calls blocked to the total number of new-call attempts. This metric can be defined for the different call types, e.g., the short-call-blocking probability is defined to be the ratio of the number of new voice (short) calls blocked to the number of voice (short) call attempts.
- *Dropping Probability:* Total dropping probability is the ratio of the total number of calls that got dropped while attempting a handoff to the total number of successful new-call attempts. The dropping probability can be similarly defined for different call types and is a measure of the QoS provided by the network.

3.4 Performance comparison of call-admission control algorithms

We first consider slow users and single call type, i.e., either short or long calls.

3.4.1 Single call type

Figure 2 shows the total blocking and dropping probabilities under UCAC, for a call arrival rate of $\lambda = 0.035$ calls/s. As the number of channels in each microcell (m) is increased, the total capacity of the system (κ) increases due to more reuse at the microcell layer, as given by Eq 1. Since the offered load to the

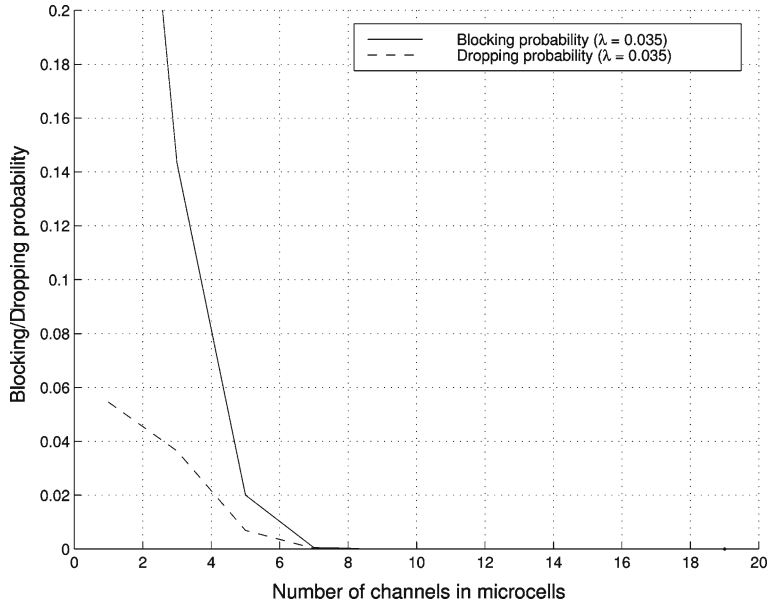


Fig. 2 Blocking and dropping probabilities for slow short calls

system is constant, when there are fewer channels in each microcell, most new calls overflow to the macrocell which gets saturated, thereby resulting in higher blocking and dropping. When more channels are allocated to each microcell, the total capacity of the microcell layer increases, as a result of which most calls are satisfied at the microcell layer. Few calls overflow to the macrocell layer, hence, fewer calls get blocked or dropped.

Furthermore, since the mobility is slow, the handoff arrival rate at the controller is not large enough to cause any significant delays. As a result, there is only a negligible increase in the channel holding time due to the handoff processing overhead.

Figure 3 shows the total number of handoffs as a function of the number of channels in a microcell.[5] When system capacity is low, many new-call attempts get blocked. So there are very few *active* calls in the system, and therefore there are few handoffs. Increasing m increases the system capacity which results in more number of active calls which generate more handoffs. When the network capacity becomes very large, the number of active calls in the system saturates, since the capacity is greater than the aggregate arrival rate. Thus, the number of handoffs generated also saturates, which is observed in Fig. 3. This leads to lower dropping and blocking probabilities as the capacity of the network increases in Fig. 2. In the moderate-capacity region, the increase in the number of handoffs is greater than the increase in capacity.

Figure 4 shows the blocking and dropping probabilities for short calls but with high user mobility (mean speed of 35 mph, and $\lambda = 0.06$ calls/s). As can be seen in the figure, there is an optimal orthogonal channel partition which provides the minimum blocking and dropping probabilities. This is a result of two opposing factors. As we allocate more channels to the microcell layer, the capacity of the network increases, which tends to decrease the blocking and dropping probabilities. Also, the cell-dwelling time of a call in a cell decreases, reducing the blocking and the dropping probabilities further. On the other hand, with more channels at the microcell layer, the number of handoffs increases, and therefore the rate of requests at the handoff controller increases, which results in higher waiting times at the controller. This causes the channel holding time in the "old" cell to increase, resulting in higher blocking probability. The cumulative result of these two factors results in the performance behavior shown in Fig. 4.

In the moderate- and high-capacity regions ($m = 3$ and above), the dropping probability is slightly larger than the blocking probability. This can be explained as follows. As the capacity of the network is increased,

[5] The figure shows the number of handoffs generated over the entire simulation period. Each experiment was run for 10 million new-call arrivals, and handoffs are generated by all calls which were successful.

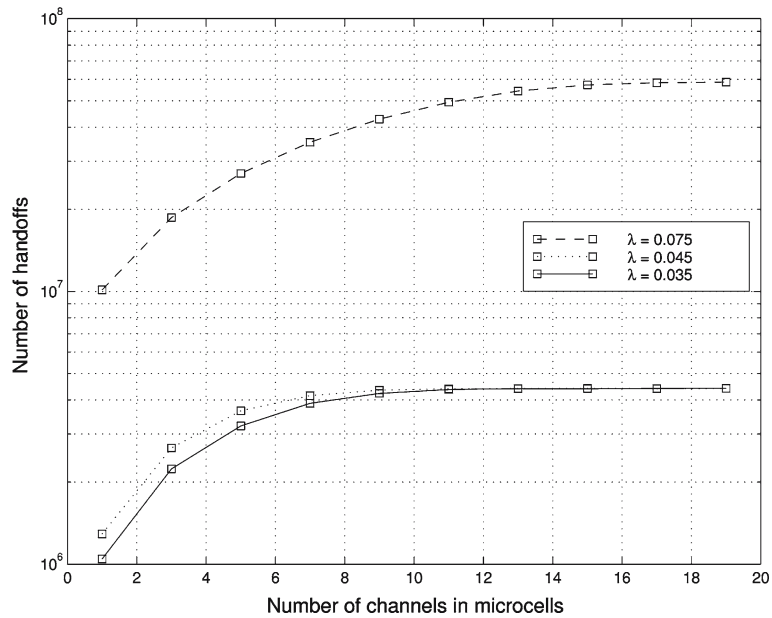


Fig. 3 Number of handoffs generated for slow short calls

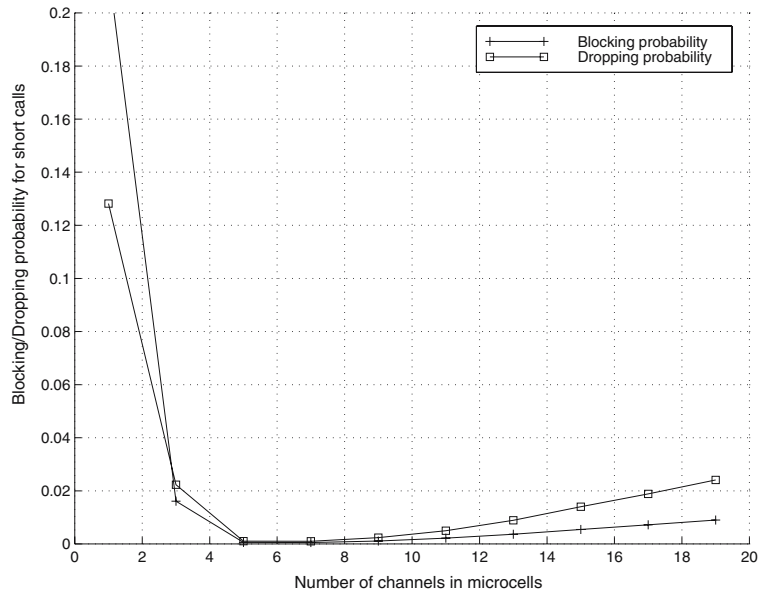


Fig. 4 Blocking and dropping probabilities for fast short calls (note effect of handoff overhead and fast mobility)

more calls are handled by the microcell layer. Since microcells are smaller in size, many more handoffs are generated when m is large. This increase in the number of handoffs has a 2-fold effect. First, the queuing time at the controller increases, increasing the holding time of a call in a cell, and hence increases the blocking probability. Therefore, fewer calls are successful in the large-capacity region (e.g., $m = 17$) as compared to the moderate-capacity region (e.g., at $m = 6$).

Second, the increase in the number of handoffs also implies an increase in the number of dropped calls. The rate of increase in the number of handoffs and the number of dropped calls, as m increases, is larger than the rate of increase in the number of new calls blocked. Hence, the dropping probability becomes larger than the blocking probability as capacity increases.

For smaller values of M ($M = 1, 2$, or 3), i.e., when the handoff controller serves handoff requests from a small number of macrocells and microcells, the results

obtained were similar to Fig. 2, and we did not observe the optimization shown in Fig. 4. These results are not included here to conserve space.

The three factors which affect the load at the handoff controller (and hence the delay) are: (1) the controller service time ($1/\mu_c$); (2) the average handoff rate (α); and (3) the number of macrocells (M) and their underlying microcells served by the controller. The average handoff rate depends on the user mobility and cannot be controlled. The controller service time can be reduced by employing faster processors at the BSC/MSC signaling centers and by adopting a hierarchical design where each handoff controller serves only a few macrocells and microcells. In Sect. 4, we will expand the function of the handoff controller to queue the handoff requests for a short period of time when there are no channels available.

3.4.2 Multiple call types

In Figs. 5–6, we compare the DCAC and UCAC algorithms with fast users generating short and long calls. The fraction of short (long) calls is 20% (80%), i.e., $F = 0.2$. The aggregate new-call arrival rate is assumed to be $\lambda = 0.06$ calls/s. Based on the results shown, we make the following observations.

The results show that, with multiple call types, there is an optimal orthogonal assignment of channels that minimizes the blocking and the dropping probabilities. The reasons are similar to those outlined in the case of the single call type with fast mobility.

Figure 6 shows the blocking and dropping probabilities for long calls. The blocking and dropping probabilities for short calls are very similar to the total probabilities shown in Fig. 5 as the percentage of short calls in the network is high. As mentioned before, the significantly longer cell-dwelling time of a long call in DCAC as compared to the UCAC adversely affects both long and short calls. This leads to higher blocking for short as well as long calls in the DCAC algorithm.

In general, we find that treating short and long calls identically, and letting them follow UCAC leads to better network performance. Moreover, since UCAC does not have to "guess" or determine the type of the call (long or short, fast or slow) before admitting it into the network, it is easier to implement.

3.4.3 Sensitivity of results to system parameters

In Sect. 3, we mentioned a list of default parameter values that have been used for the simulation experiments presented in this study. Though only a few results have been presented, we have carried out a detailed sensitivity analysis with respect to the other parameters. We observed the following.

The call-holding time, in the single-class traffic experiments, was varied from 90 to 1,200 s. As expected,

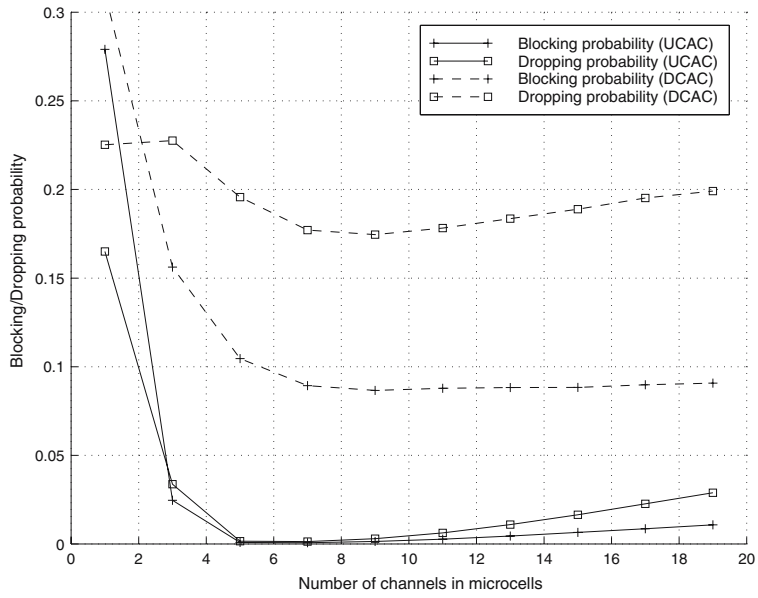


Fig. 5 Blocking and dropping probabilities for combined long and short calls and fast-moving users

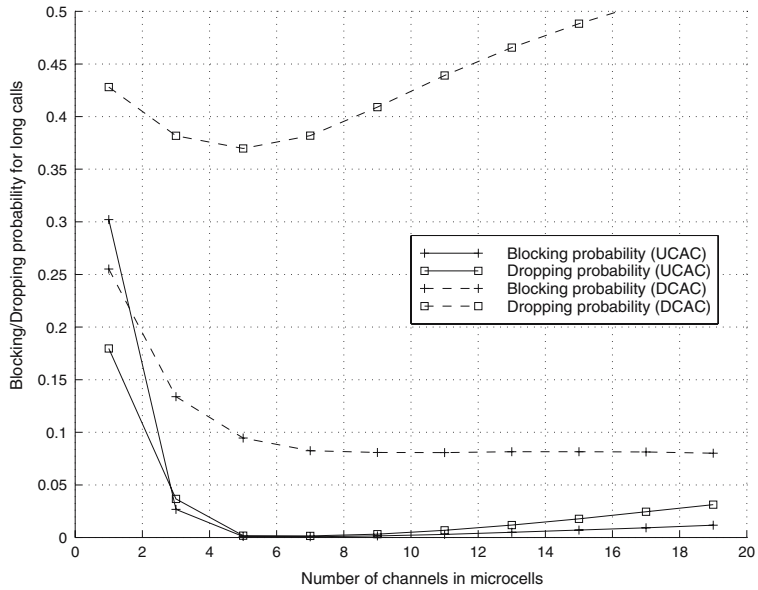


Fig. 6 Long-call blocking and dropping probabilities for fast-moving users

with an increase in holding time, both the blocking and dropping probabilities increased.

By varying the reuse distance ratios (R_u and R_m) between 2 and 4, and the total number of channels (C) between 50 and 150, we studied the effects of network capacity on the blocking and dropping probabilities. When handoff overheads were assumed to be absent, an increase in capacity always led to a decrease in blocking. When handoff overheads were significant, the blocking characteristics changed depending upon M. When M was small, the blocking and dropping probabilities decreased with an increase in capacity. On the other hand, a larger M, coupled with a larger capacity, increased the load at the handoff controller, thereby increasing the blocking and dropping probabilities.

By keeping the microcell radius constant and changing N from 7 to 50, we studied the effects of different macrocell sizes on blocking. An increase in the macrocell size implied an increase in coverage area, and hence an increase in the blocking and dropping probabilities.

4 Call handoff control using handoff queues

The goal of an effective call handoff control mechanism is to ensure that the percentage of dropped calls for users admitted into the network is minimized. In this section, we explore a queuing-based handoff control framework. In particular, when a handoff attempt is made to a target cell that does not have any idle channel, the call is queued for a period of time referred to as the *queue time*. A handoff call is dropped if it cannot get a channel before its queue time expires.

In the GSM-based network, this handoff queue can be deployed in the BSC and/or the MSC. The buffer in the BSC can queue intra-BSC handoffs while the buffer in the MSC can be used to queue inter-BSC handoffs. Each cell can have a unique queue for buffering handoff requests to the cell when there is no channel available.

To study the benefits of handoff queueing purely, we assume DCAC as our base admission control algorithm, such that our results do not incorporate the benefits gained via a better call admission control algorithm, and hence complicate the conclusions. We have modified DCAC to include this handoff queue as follows.

– On a handoff in any tier, if there is no free channel in the target cell, the handoff request is queued. The request remains in the queue as long as the mobile user is still "alive". During this time, the mobile user continues to send channel information to the BSC; thus, the handoff request in the queue is updated or removed as a result of change of state in the new target cell or due to termination of the call.

To avoid the "ping-pong" effect during handoff, when a channel is released in the target cell, the

mobile user will not switch to this channel immediately if the mobile user can still receive good signals from its original cell. It holds the channel for a period of time set for this operation. The mobile user performs the handoff only after this timeout or the signal from its original cell degrades below a certain level. Therefore, as an added advantage, handoff request queuing can also reduce unnecessary ping-pong-related handoff operations.

To investigate our handoff-queue-based QoS control framework, we have studied the following queuing architectures.

- *FIFOMIC—FIFO queue in microcell*: Each microcell has a queue that is serviced in a FIFO order. When all channels in a microcell are occupied, new calls are blocked while handoff calls are buffered in the queue. If a handoff call cannot get an idle channel before its queue time expires, it will overflow to the overlaying macrocell. If there is a free channel in the macrocell, it will acquire that channel. Otherwise, it is dropped.
- *FIFOMAC—FIFO queue in macrocell*: A queue is used only in the macrocells. If all channels in that macrocell are occupied, new calls are blocked and all handoff calls (slow and fast, short and long) will wait in the queue. A handoff call is dropped if it cannot get a channel before its queue time expires. As in the previous case, the queue is serviced in FIFO order.
- *PRIMAC—Priority queue in macrocell*: This model is the same as FIFOMIC except that a priority queue is used. Fast handoff calls in the queue have higher priority and when a channel is released, it is first assigned to a fast handoff call. Slow queued handoff calls can acquire a released channel when there are no fast handoff calls in the queue.
  Collectively, FIFOMAC and PRIMAC are referred to as *QMAC* in this work.
- *FIFOMIC-FIFOMAC—FIFO queue in both macrocell and microcell*: Queues are used in both macrocells and microcells. A handoff call in a microcell will overflow to the overlaying macrocell. If it cannot get a channel in the macrocell, it will stay in both the queue in the macrocell and the queue in microcell. It will use the tier which has the available channel first. A queued microcell handoff call will get out of both queues when it is assigned an idle channel or when its queue time expires.
- *FIFOMIC-PRIMAC—FIFO queue in microcell and priority queue in macrocell*: Similar to FIFOMIC–FIFOMAC in behavior, except that a priority queue is used at the macrocell tier.

In this paper, we develop an analytical model to calculate the blocking probabilities for different users under FIFOMIC and QMAC. We will use simulation to verify the accuracy of the analytical models.

4.1 Performance analysis

In this section, we first present the analytical model for two classes of users, namely, fast and slow users, for a single-tier cellular network. In Sect. 4.3, we use the results of the above model to analyze the performance for different queue architectures in a two-tier (macrocell/microcell) cellular network.

4.1.1 Modeling assumption

The analytical model developed in the following subsections is based on the following additional assumptions (see Sects. 2 and 3 for the base notations and assumptions).

We consider a single cell with c channels. The cell has a finite FIFO queue of size q to buffer handoff calls. The time that a call stays in a cell before it is handed-off to a new cell is referred to as the *cell-dwelling time*. The arrivals of new calls follow Poisson processes with rates λ_f and λ_s for fast and slow new users, respectively, while the arrivals of handoff calls are approximated as Poisson processes with rates λ_{hf} and λ_{hs} for fast and slow handoff users, respectively.

We approximate the dwelling time to be negative exponentially distributed with mean $1/\mu_d$ seconds. The queue time is assumed to be negative exponentially distributed with mean μ_q seconds, and both fast and slow users have the same average call holding time which is negatively exponentially distributed with mean $1/\mu$ seconds. The cell-dwelling times for both fast and slow users are negative exponentially distributed with means of $1/\mu_{df}$ and $1/\mu_{ds}$ seconds, respectively.

For simplicity and for fair comparison between the different types of users, we assume that both fast and slow users have the same queue time distribution, which is negatively distributed with mean $1/\mu_q$ seconds.

It is not very accurate to assume a fixed mean dwelling time in the cellular system with queue. The dwelling

time of a call in a cell depends on how long it stays in the queue, which changes with the traffic intensity. When dropping probability is low and/or the queue time is small, this effect can be ignored and the above assumption is reasonably accurate.

4.2 Two classes of users in a single-tier cellular network

We extend the analysis for a single class of users in a single-tier cellular network [18] to a cellular system with two types of users, namely, fast users and slow users.

We describe the state of the system by $s(i, j, q)$ where i, j, and q denote the number of fast users, slow users, and queued handoff calls in the target cell, respectively. Since c is the total number of available channels in the cell, when $i + j < c$, the state transition is the same as a normal two-dimensional Markov chain for two types of users without queue. If a new call arrives when $i + j = c$, it will be blocked. On the other hand, if a handoff call arrives when $i + j = c$, irrespective of whether it is a fast user or a slow user, it will be buffered in the queue and the queue length will increase by 1.

When $q > 1$, the queue length will decrease by 1 under the following three cases:

1. A queued user exits the queue because (a) its queue time expires or (b) a fast user releases a channel and the channel is occupied by a fast user which was at the front of the queue or (c) a slow user releases a channel and the channel is occupied by a slow user which was at the front of the queue. For these cases, the queue length is reduced by 1, and the number of the channels occupied by fast users and slow users in the cell remains the same.
2. A slow user releases the channel which is acquired by a fast user which was at the front of the queue. Thus, the number of channels occupied by fast users increases by 1, the number of channels occupied by slow users decreases by 1, and the number of users in the queue decreases by 1.
3. A fast user releases a channel which is acquired by a slow user which is at the front of the queue. Thus, the number of channels occupied by slow users increases by 1, the number of channels occupied by fast users decreases by 1, and the number of users in the queue decreases by 1.

In an equilibrium state, the probability that a user at the front of the queue is a fast user (or a slow user) depends on the relative arrival rates of the handoff calls. To a reasonable approximations, the probability that the user at the front of the queue is a fast user is equal to $\lambda_{hf}/(\lambda_{hf} + \lambda_{hs})$ and the probability that the user at the front of the queue is a slow user is equal to $\lambda_{hs}/(\lambda_{hf} + \lambda_{hs})$.

Based on the above discussions, we can enumerate the various state transitions and their corresponding transition rates as follows:

$$
\begin{aligned}
&s(i, j, 0) \rightarrow s(i + 1, j, 0) : \lambda_h + \lambda_{hf}, i + j < c;\\
&s(i, j, 0) \rightarrow s(i, j + 1, 0) : \lambda_l + \lambda_{hs}, i + j < c;\\
&s(i, j, 0) \rightarrow s(i - 1, j, 0) : i(\mu + \mu_{df}), i + j \leq c, i \geq 1;\\
&s(i, j, 0) \rightarrow s(i, j - 1, 0) : j(\mu + \mu_{ds}), i + j \leq c, j \geq 1;\\
&s(i, j, q) \rightarrow s(i, j, q + 1) : \lambda_{hf} + \lambda_{hs}, \quad i + j = c, q \geq 0;\\
&s(i, j, q) \rightarrow s(i, j, q - 1) : q\mu_q + \frac{\lambda_{hf} i(\mu + \mu_{df}) + \lambda_{hs} j(\mu + \mu_{ds})}{\lambda_{hf} + \lambda_{hs}}, \quad i + j = c, q \geq 1;
\end{aligned}
$$

$$
\begin{aligned}
&s(i, j, q) \rightarrow s(i + 1, j - 1, q - 1) : \frac{\lambda_{hf} j(\mu + \mu_{ds})}{\lambda_{hf} + \lambda_{hs}}, \quad i + j = c, q \geq 1;\\
&s(i, j, q) \rightarrow s(i - 1, j + 1, q - 1) : \frac{\lambda_{hs} i(\mu + \mu_{df})}{\lambda_{hf} + \lambda_{hs}}, \quad i + j = c, q \geq 1
\end{aligned}
$$

Figure 7 shows the state-transition diagram for a cell with two channels and a queue with a buffer size of 2. From the state-transition diagram, we can write down the flow-balance equations which can then be solved to obtain the probability for each state $p(i, j, q)$. The blocking probability for new calls P_{bn} is then given by $P_{bn} = \sum_{i+j=c} p(i, j, q)$.

Next, we find the dropping probability for handoff calls. We consider a tagged handoff call that arrives when the system is in state $s(i, j, q)$, where $i + j = c$ and $q \geq 0$. This state corresponds to the case in which there are already q handoff calls waiting in the queue. Note that, if $i + j < c$, then a handoff call immediately gets the channel. The tagged handoff call will not be blocked only if, before its queue time expires,

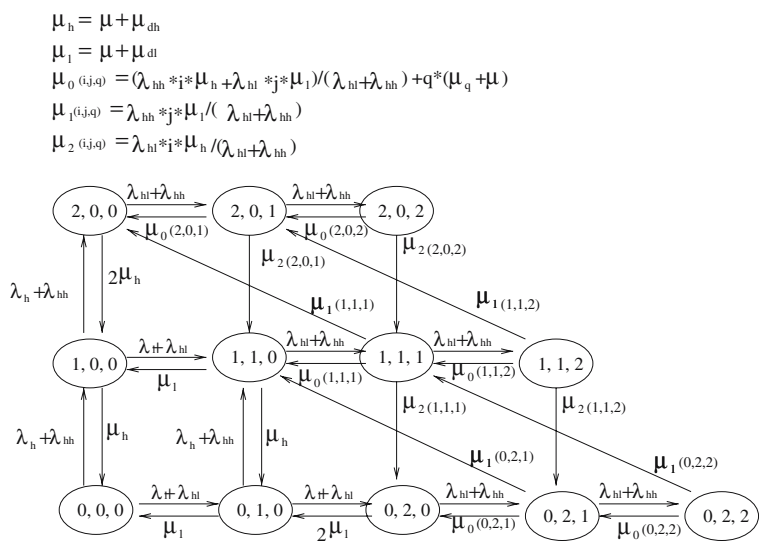


Fig. 7 State-transition diagram for a system with fast and slow users, two channels, and a queue size of 2

1. all the q queued users in front of it leave the queue either because their queue time expires or they successfully get a channel, *and*
2. it gets a channel.

Figure 8 shows all the state-transition paths following which the tagged handoff call (i.e., the handoff call coming at state $s(i, j, q)$) may finally get a channel. The queue time for the tagged handoff call must be longer than the time required for the system to change from state $s(i, j, q)$ to state $s(i', j', 0)$, where $i' + j' = c - 1$. Thus, the probability for the tagged handoff call to get a channel is the sum of the probabilities that the handoff call can get a channel following all the paths where each of these paths consists of multiple steps.

From each state in Fig. 8, the one-step transition can take the system to one of three possible states.[6] The transition probabilities to each of the three possible next steps can be calculated. For example, let $p[s(i-1, j+1, q-1)|s(i, j, q)]$ be the probability of transitioning from state $s(i, j, q)$ to state $s(i-1, j+1, q-1)$. This transition corresponds to the case that one of the i fast users releases a channel which is acquired by a slow user that is in the front of the queue. Let t_1 denote the time for a fast user to release a channel and t_2 denote the time for any other event to occur that causes a state change (i.e., a slow user releases the channel, or the queue time for any of the queued handoff calls in front of the tagged handoff call expires, or the queue time of the tagged handoff call expires). Define $p[t_1 < t_2]$ as the probability that $t_1 < t_2$. Due to the memoryless property of the negative exponential distribution, the density function of minimum time t_2 is given by:

$$f_{t_2}(t_2) = (q\mu_q + j(\mu + \mu_{ds}) + \mu_q) \times e^{-(q\mu_q + j(\mu+\mu_{ds})+\mu_q)t_2} \tag{2}$$

and the density function for t_1 is given by:

$$f_{t_1}(t_1) = i(\mu + \mu_{df})e^{-i(\mu+\mu_{df})t_1}. \tag{3}$$

The probability that a fast call will release a channel first is then given by:

$$\begin{aligned} p[t_1 < t_2] &= \int_0^\infty \int_{t_1}^\infty (q\mu_q + j(\mu + \mu_{df}) + \mu_q) \\ &\quad \times e^{-(q\mu_q + j(\mu+\mu_{ds})+\mu_q)t_2} i(\mu + \mu_{df}) \\ &\quad \times e^{-i(\mu+\mu_{ds})t_1} dt_2 dt_1 \\ &= \frac{i(\mu + \mu_{df})}{\mu_q + q\mu_q + i(\mu+\mu_{df}) + j(\mu+\mu_{ds})}. \end{aligned} \tag{4}$$

Given the probability that a slow handoff call is in the front of the queue as $\frac{\lambda_{hs}}{\lambda_{hs}+\lambda_{hf}}$, we have:

$$p[(i-1, j+1, q-1)|(i, j, q)] = \frac{\lambda_{hs}}{\lambda_{hf} + \lambda_{hs}} p[t_1 < t_2]. \tag{5}$$

[6] The states at which all c channels are occupied either by only fast users or by only slow users have only two next states.

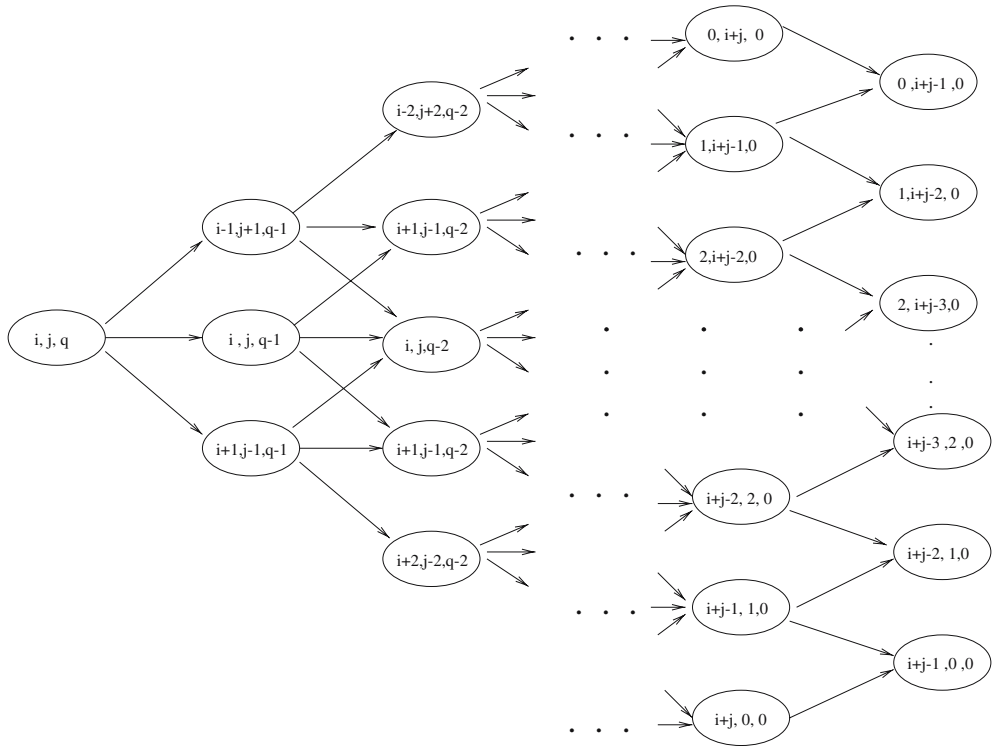


Fig. 8 State-transition paths from $s(i, j, q)$ under which a queued handoff call will not be blocked

Similarly,

$$p[s(i+1, j-1, q-1)|s(i, j, q)] = \frac{\lambda_{hf}}{\lambda_{hf}+\lambda_{hs}} \times \frac{j(\mu+\mu_{ds})}{\mu_q + q\mu_q + i(\mu+\mu_{df}) + j(\mu+\mu_{ds})}, \quad (6)$$

and

$$p[s(i, j, q-1)|s(i, j, q)] = \frac{\lambda_{hf}}{\lambda_{hf}+\lambda_{hs}} \frac{i(\mu+\mu_{df})}{\mu_q+q\mu_q+i(\mu+\mu_{df})+j(\mu+\mu_{ds})} + \frac{\lambda_{hs}}{\lambda_{hf}+\lambda_{hs}} \frac{j(\mu+\mu_{ds})}{\mu_q+q\mu_q+i(\mu+\mu_{df})+j(\mu+\mu_{ds})} + \frac{q\mu_q}{\mu_q + q\mu_q + i(\mu+\mu_{df}) + j(\mu+\mu_{ds})}. \quad (7)$$

Again, exploiting the memoryless property of the negative exponential distribution, we can use the same queue-time density function to find the probabilities for the following steps. Finally, at a state $(i, j, 0)$, the probabilities for this handoff call to get a released channel are equal to:

$$p[s(i, j-1, 0)|s(i, j, 0)] = \frac{j\mu_{ds}}{\mu_q + i\mu_{df} + j\mu_{ds}}, \quad (8)$$

or

$$P[s(i-1, j, 0)|s(i, j, 0)] = \frac{i\mu_{df}}{\mu_q + i\mu_{df} + j\mu_{ds}}. \quad (9)$$

Once we determine the probability for each step at different states, we can calculate the probability for each path to obtain the overall successful handoff probability. If we define $P_r(m)$ to be the probability following one of the paths that the tagged handoff call coming at state $s(i, j, q)$ will finally get a released channel, then $P_r(m)$ can be found by multiplying the probabilities of each step along this path. For example, one of the $P_r(m)$'s from state $s(i, j, q)$ to state $s(i-1, j, 0)$ following the central path of the tree in Fig. 8, $P_r(0)$ can be calculated as:

$$P_r(0) = p[s(i, j-1, 0)|s(i, j, 0)] \times \prod_{n=0}^{q-1} p[s(i, j, n)|s(i, j, n+1)]. \quad (10)$$

If there are M such paths following which a handoff call can finally get a channel, the blocking probability for a handoff call coming at state $s(i, j, q)$ is given by:

$$P_b(i, j, q) = 1 - \sum_{M} P_r(m). \quad (11)$$

The problem to determine $P_b(i, j, q)$ is a simple trellis tree problem and can be easily computed numerically. For a finite queue with length Q, the blocking probability for handoff calls is given by:

$$P_f = \sum_{q=0}^{Q-1} \sum_{i+j=c} P_b(i, j, q) + \sum_{i+j=c} p(i, j, Q), \quad (12)$$

where $p(i, j, Q)$ is the probability that all queue positions are occupied.

If P_c is the probability that a handoff call may finish when it stays in the queue, the dropping probability P_{bh} is modified as:

$$P_{bh} = P_h(1 - P_c).$$

From [18], we have:

$$P_c = \frac{\mu}{\mu + \mu_q}.$$

The handoff rates are calculated recursively as follows:

$$\lambda_{hf} = \sum_{i=0}^{c} i \sum_{j=0}^{c-i} \mu_{df}\, p(i, j, 0) + \sum_{q=1}^{q=Q} \sum_{i+j=c} i\mu_{df}\, p(i, j, q), \quad (13)$$

$$\lambda_{hs} = \sum_{j=0}^{c} j \sum_{i=0}^{c-j} \mu_{ds}\, p(i, j, 0) + \sum_{q=1}^{q=Q} \sum_{i+j=c} j\mu_{ds}\, p(i, j, q). \quad (14)$$

4.3 Analysis of two-tier cellular network with queue

Using the above results, we can analyze the performance of FIFOMIC and QMAC. In [18], it is shown that t_d, the dwelling time of a mobile user in a cell, is given by $t_d = \frac{\pi S}{E[v]L}$, where S and L are the area and the perimeter of the cell, respectively, and $E[v]$ is the mean speed of the user. If we approximate both macrocells and microcells to be circular in shape and one macrocell covers (or is as large as) N microcells, and define μ_{ds} and μ_{ds-ma} to be the mean departure rates for slow users in a microcell and in a macrocell, respectively, we find that $\mu_{ds-ma} = \mu_{ds}/\sqrt{N}$.

In FIFOMIC, if the blocking/dropping probabilities for the new calls and handoff calls in a microcell are P_{n-mi} and P_{h-mi} (obtained by using the method in [18]), respectively, then the overflows to the macrocell tier of these two sources follow a Poisson process with average rates of $P_{n-mi}\lambda_s$ and $P_{h-mi}\lambda_{hs}$, respectively. The problem to find the blocking probabilities in a macrocell is then a simple two-dimensional Markov problem with known new-call and handoff-call arrival rates. If the handoff-call and new-call dropping/blocking probabilities in the macrocells are P_{h-ma} and P_{n-ma}, while P_{b-hf}, P_{b-nf}, P_{b-hs}, and P_{b-ns} are the dropping/blocking probabilities for fast handoff calls, fast new calls, slow handoff calls, and slow new calls, respectively, we get $P_{b-hh} = P_{h-ma}$; $P_{b-nh} = P_{n-ma}$; $P_{b-hl} = P_{h-mi} \cdot P_{h-ma}$; and $P_{b-nl} = P_{n-mi} \cdot P_{n-ma}$.

In FIFOMAC, the analytical model in the microcell tier is a simple $M/M/m$ model. The analytical model with two classes of users in a single tier can be used to analyze the blocking/dropping probabilities in the macrocell tier with a small modification. After obtaining the blocking/dropping probabilities in the both tiers, all the blocking/dropping probabilities for fast and slow users can be calculated.

4.4 Comparison of queue architectures for QoS improvement

In this subsection, we present a performance comparison between the various handoff queuing schemes, using the mathematical models developed above, as well as simulation. The performance metrics that we have used to compare the various queue strategies are the new-call blocking and the handoff-call dropping probabilities, as defined in Sect. 3.

4.4.1 Benefits of queuing handoff calls in a single-tier network

Figure 9 shows the new-call blocking and the handoff-call dropping probabilities for single-tier network with two classes of users as a function of the arrival rate of new calls for fast users λ_f.[7] Since the arrival rate of new calls from fast users is the same as the arrival rate of new calls from slow users, increasing λ_f also implies increasing λ_s. The results are shown for

[7] The arrival rate is normalized by $c\mu$, which is the *instantaneous* capacity of each cell.

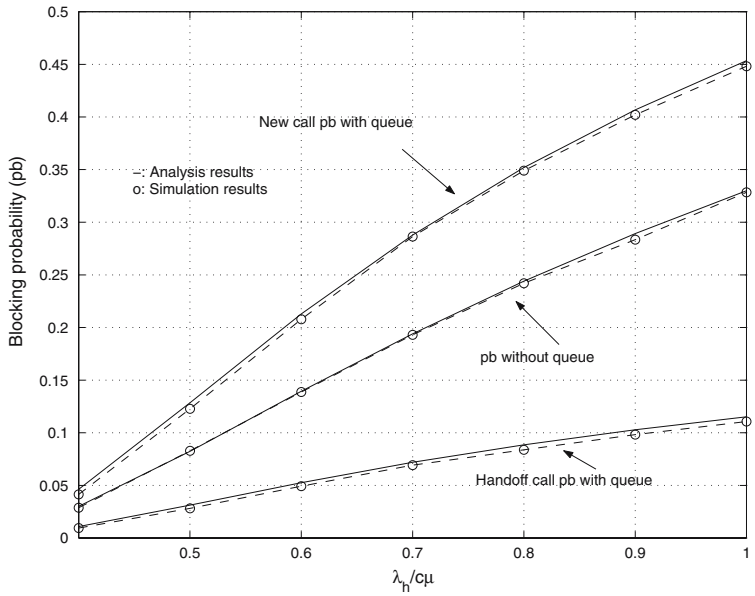


Fig. 9 New-call blocking and handoff-call dropping probabilities as functions of the offered load of fast users $\lambda_f/c\mu$

two types of networks—one which has simple FIFO queues for queuing handoff calls and the other without any queue. Note that, when there is no queue, the handoff dropping and the new-call blocking probabilities are the same and equal to the overall call-blocking probability. From the results shown in the figure, we make the following observations.

The results obtained from the analytical and simulation models are very close. This is expected since, for the single-tier network, there is no difference between the simulation and the analytical models. The key observation, however, is that the iterative method for calculating the actual handoff rates converges to the same fixed point for both the simulation and the analytical models. We found that, at low loads, this iterative scheme converges quickly. However, at very high load, such as $\lambda/c\mu = 0.9$, it took up to 40 iterations before handoff rates between successive iterations converged to a very small difference.

Comparing the results for the two networks, we observe that, when a queue is used to buffer handoff calls, the dropping probability decreases while the new-call blocking probability increases. As expected, increasing the arrival rate of fast users results in higher blocking probabilities.

Figure 10 quantifies the benefit of queuing handoff call as a function of the channel utilization. The figure plots the decrease in the dropping probability related to the case in which handoff calls are not queued. The plot also shows the increase in the new-call blocking probability. From the figure, we observe that the reduction in dropping probability is more than the increase in the new-call blocking probability. As the load increases, the arrival rate of handoff calls also increases which gets preference over the new calls which leads to lower blocking probability. However, as the load increases, there is an upper bound on the arrival rate of handoff calls to the cell. For the uniform mobility model considered in this paper, the arrival rate of handoff calls reaches a maximum when all the channels in neighboring cells are fully utilized. Beyond this point, increasing the arrival rate of new calls increases the new-call blocking probability, which becomes dominant, and the difference between the new-call blocking probability in the case of the network with queue and the overall blocking probability in the case of the network without queue decreases.

In Fig. 11, we show the effect of mean queue time on the new-call blocking and handoff dropping probabilities. As the mean queue time of a handoff call is increased, there is a higher probability that it will get a channel and hence the dropping probability decreases. In a real network, the queue time is related to the size of the overlapping zone between adjacent cells. The queue time will be larger if the overlapping area is larger. However, this implies that the frequency reuse distance will be smaller, resulting in lower spectrum efficiency per unit coverage area.

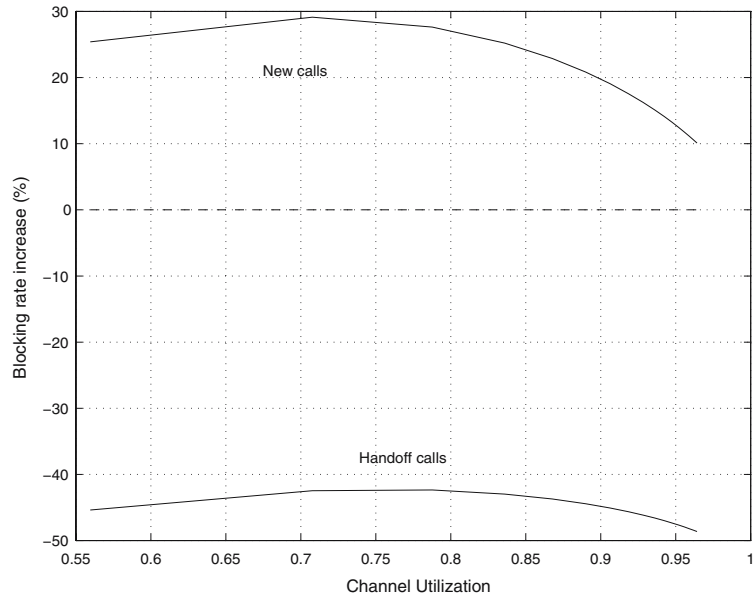


Fig. 10 Decrease (increase) in dropping (new-call blocking) probability as a function of the channel utilization relative to the case of a network with no queue

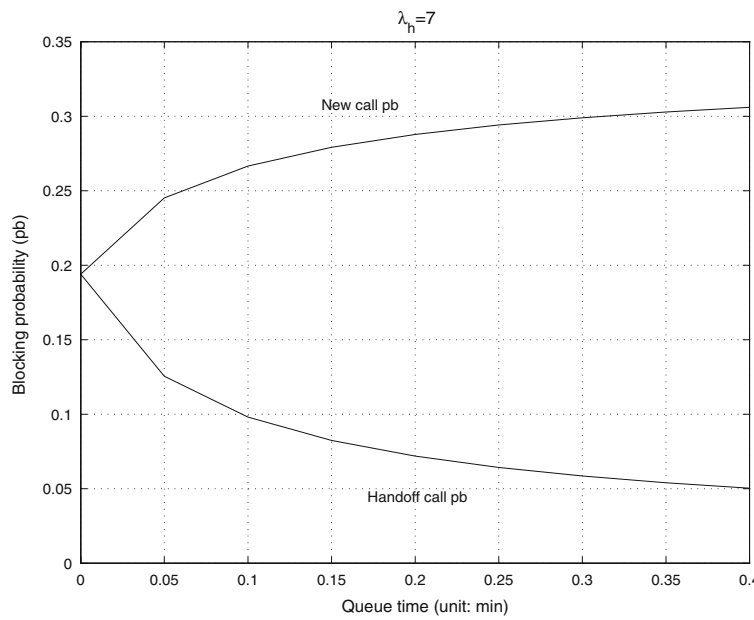


Fig. 11 Handoff dropping and new-call blocking probabilities as functions of queue time

In Fig. 12, we show the handoff dropping and new-call blocking probabilities as functions of the queue length. We find that, in a cell with 10 channels, a queue length of 3 is long enough to guarantee that the handoff calls can have the maximum benefit from the queue. This result is important because longer queue length implies more radio resources to manage and control the larger queue length. Furthermore, a larger queue also implies higher management overhead at the BSC.

Table 1 shows the sensitivity of the results to the number of channels assigned to each cell. The maximum queue length refers to the size of the queue beyond which adding more buffer results in less than 1% improvement in dropping probability. Beyond these queue lengths, the dropping probabilities will improve by less than 1% when adding one more queue position.

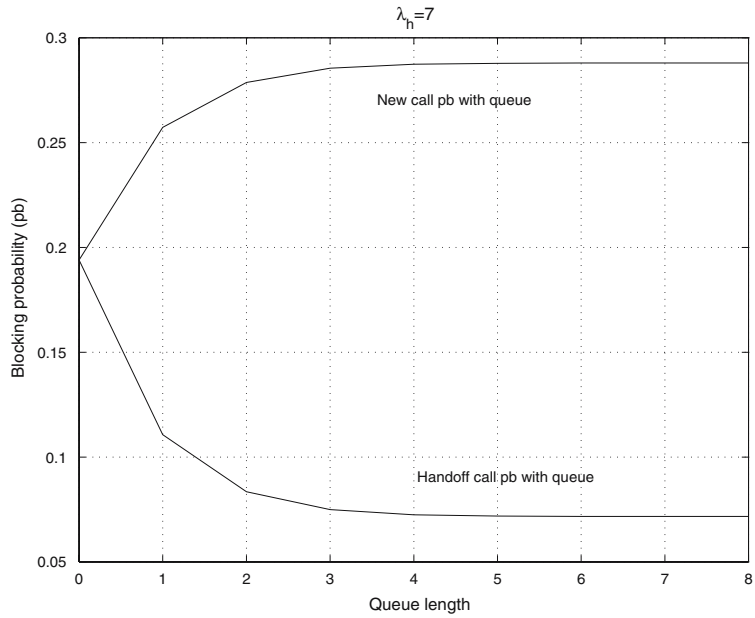


Fig. 12 Handoff dropping and new-call blocking probabilities as functions of queue length

Table 1 Minimum queue length required to guarantee 1% dropping probability

Channels per cell	10	20	30	40	50
Minimum queue length	4	5	6	6	7

4.4.2 Comparison of FIFOMIC and FIFOMAC

Now, we analyze the benefit of the queues in a two-tier network with fast and slow users based on the analytical model developed in Sect. 4.1. Figure 13 shows the handoff dropping and the new-call blocking probabilities for slow users. Figure 14 shows the handoff dropping and new-call blocking probabilities for fast users. The following are the key observations from the plots.

For slow users, there is a reduction in the handoff dropping probability for FIFOMIC and QMAC. This reduction comes at the expense of an increase in the new-call blocking probability. The reduction in dropping is similar for both FIFOMIC and FIFOMAC, with FIFOMIC performing marginally better. In FIFOMIC, handoff calls of slow users are queued in the microcell tier which significantly reduces the dropping probability. In FIFOMAC, since there is no queue in the microcell tier, handoff calls of slow users that cannot find a channel will overflow to the macrocell tier where they contend for the queue with handoff calls for fast users. As a result, the dropping probability of slow users is lower in FIFOMIC.

In FIFOMIC, since there is no queue at the macrocell tier, the new-call blocking and handoff dropping probabilities for fast users are the same (as shown in Fig. 14). The blocking probability is marginally lower than that for the network without any queue, as the overflow of slow handoff and new calls is smaller. In FIFOMAC, the dropping probability of fast users is lower due to the benefit of queuing handoff calls in the macrocell tier.

4.4.3 Overall comparison

Since PRIMAC, FIFOMIC–FIFOMAC, and FIFOMIC–PRIMAC are analytically intractable, we have used simulations to compare the performance of all the queuing strategies. Figure 15 shows the dropping probability for slow users for all the models, while Fig. 16 shows the dropping probability for fast users. Based on the figures, we make the following observations.

The dropping probabilities for FIFOMIC and FIFOMAC have the same trends as those observed from analysis. However, in absolute terms, there are large differences. The key factor for this discrepancy is due to approximation in the analytical model that the overflow traffic follows a Poisson process.

When priority queuing is used in the macrocell tier as in PRIMAC, the dropping probability for fast users is reduced, as fast handoff calls have the highest priority

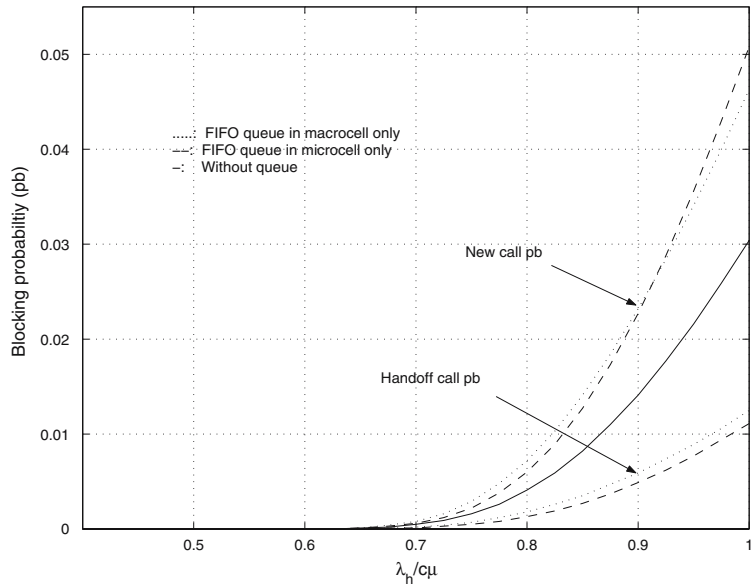


Fig. 13 Handoff dropping and new-call blocking probabilities as functions of the offered load for slow users for FIFOMIC and FIFOMAC

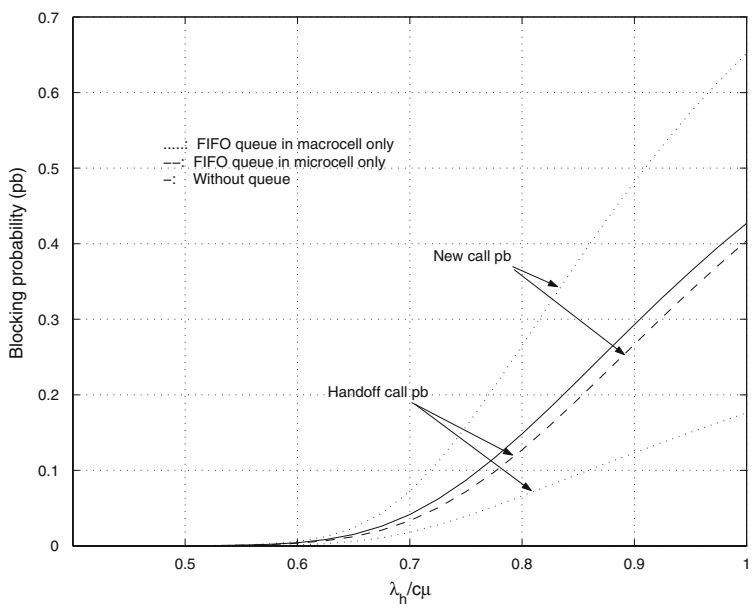


Fig. 14 Handoff dropping and new-call blocking probabilities as functions of the offered load for fast users for FIFOMIC and FIFOMAC

over slow handoff calls and new calls. Since FIFOMIC–PRIMAC has a queue in both tiers, including a priority queue in the macrocell tier, it provides the best separation of slow and fast calls between the two layers; i.e., calls from slow users remain in the microcell tier and calls from fast users get preference to channels in the macrocell layer. As a result, queues in both tiers are effectively used, and this model has the lowest dropping probability for both fast and slow users. Finally, it should be noted that, in PRIMAC, the dropping probability for slow users is greater than that in the model without any queues, as the slow users' overflow traffic is treated like new calls.

Although we find that FIFOMIC–PRIMAC has the best performance, it is only slightly better than either FIFOMAC or PRIMAC. Considering also other factors such as the cost of implementation and the complexity of the control protocols, FIFOMIC–PRIMAC may not be a good choice. More importantly, when queues are deployed in both tiers, the queue control protocol has

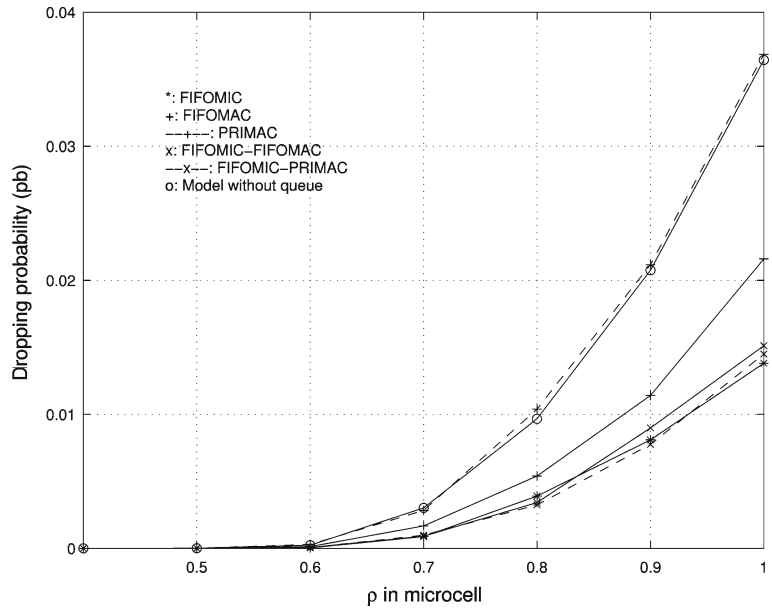


Fig. 15 Comparison of handoff dropping probabilities for all queuing architectures for slow users

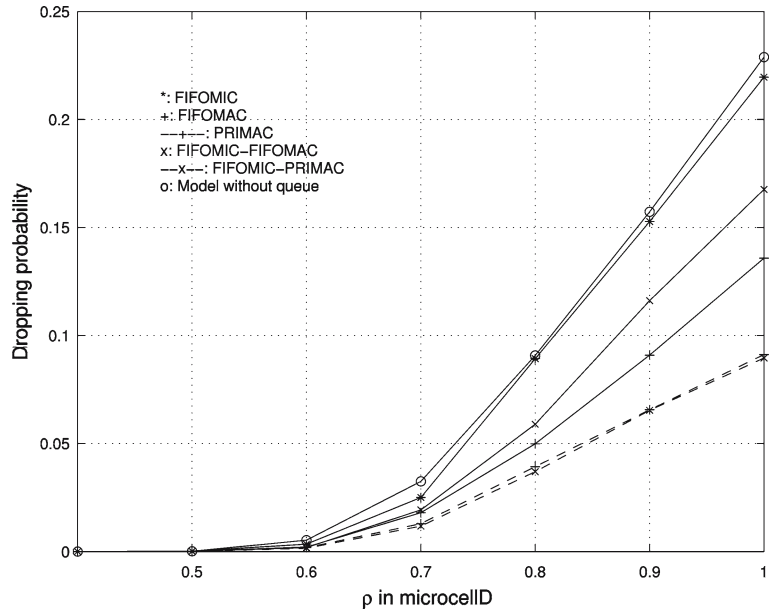


Fig. 16 Comparison of handoff dropping probabilities for all queuing architectures for fast users

to occupy extra radio resources in both tiers, which, in turn, further reduces the number of usable channels in the system.

5 Conclusion

A multi-tier cellular network architecture has been proposed as a solution to two key limitations of cellular networks: (1) a lack of spectrum and capacity and (2) management of high-speed users, who generate very small cell-dwelling times and who are more prone to call drops. In this work, we have studied a call-admission and handoff control framework for multi-tier cellular networks, using a two-tier cellular network as representative for multi-tier networks. We have analyzed both new-call admission and handoff-call control in a single unified framework.

We first presented a call-admission control algorithm which takes into account the cell-dwelling time

in assigning a call to the different tiers. This algorithm, known as the Dwell-time-based Call-Admission Control (DCAC) algorithm is taken as our base CAC algorithm. We compared the performance of DCAC against a simpler Uniform Call-Admission Control (UCAC) algorithm which handles all calls and users uniformly, without differentiating them as high-speed (fast) or low-speed (slow) users. For both algorithms, we studied the impact of channel partitioning between the tiers on the new-call blocking and handoff-call dropping probabilities. Our results showed that the simpler UCAC algorithm outperformed the DCAC algorithm for various mixes of user mobility and call characteristics, as the advantage (in terms of lower dropping and blocking probability) of increased resource availability via greater channel re-use at the lower tier far outweighs the advantage due to reduction in handoff traffic by placing longer calls into the higher tier. Additionally, by keeping longer calls in the higher tier, we hinder the ability of the network to assign overflow calls to the upper tier, hence increasing the dropping and blocking probabilities further. Moreover, there is an optimal fixed channel partition of the total spectrum among the two tiers which minimizes the blocking and dropping probabilities for both the CAC algorithms. An open problem for future research is how can one analyze these results to develop heuristics which will guide network architects in implementing a CAC scheme which will dynamically choose either DCAC or UCAC based on whether the network performance or handoff manageability is more important, given the current network operating conditions.

Once a call-admission control framework admits a new call into the network, the job of a call-handoff-control mechanism is to ensure that this admitted call is not dropped due to a blocked handoff. We have studied, through analysis and simulation, various handoff queuing strategies which attempt to improve the dropping probability of calls in the network. Our results show that implementing a queue only in one of the tiers can improve the dropping probability significantly. Among the different queuing strategies studied, we find that implementing the queue in the macrocell tier results in a significant reduction in dropping probability for both the tiers in the network while keeping the cost low and the operation of the system simple.

A major concern of our call-admission control and handoff queuing framework was ease of implementation while being effective in network resource management. After comparing various proposals, we show that the simpler UCAC scheme is more efficient than the more "intelligent" DCAC scheme for new-call admission control, while implementing a single macrocell layer queue provides significant improvement in handoff dropping.

An open problem for future research is to extend our models to incorporate more realistic network architectures which are multi-tiered and carry heterogeneous voice and data traffic, where the voice traffic is packetized.

Acknowledgements We gratefully acknowledge the Editors and the reviewers of our paper for their constructive criticisms which helped to improve the paper significantly.

References

1. Scheibenbogen, M., Clausen, S., & Guntsch, A. (1999). Dynamic channel allocation in hierarchical cellular systems. In *Proc., vehicular technology conference*, Vol. 4, pp. 2418–2422.
2. Sung, C. W., & Shum, K. W. (1999). Channel assignment and layer selection in hierarchical cellular system with fuzzy control. In *Proc., vehicular technology conference*, Vol. 4, pp. 2433–2477.
3. Greenstein, C.-L. I. L., & Gitlin, R. D. (1993). A microcell/macrocell cellular architecture for low- and high-mobility wireless users. *IEEE Journal on Selected Areas in Communications, 11*(6), 885–890.
4. Anpalagan, A. S., & Katzela, I. (1999). Overlaid cellular system design with cell selection criteria for mobile wireless users. In *Proc., Canadian conference on electrical computer engineering*, May 1999, Vol. 1, pp. 24–28.
5. Chang, C., Chang, C.-J., & Luo, K. R. (1999). Analysis of a hierarchical cellular system with reneging and dropping for waiting new and handoff calls. *IEEE Transactions on Vehicular Technology, 8*(4), 1080–1091.
6. Lagrange, X., & Godlewski, P. (1999). Performance of a hierarchical cellular network with mobility-dependent handover strategies. In *Proc., vehicular technology conference*, April 1999, Vol. 3, pp. 1868–1872.
7. Cimone, G., Weerakoon, D. D., & Aghvami, A. H. (1999). Performance eveluation of a two layer hierarchical cellular system with variable mobility user using multiple class applications. In *Proc., IEEE vehicular technology conference*, September 1999, pp. 2835–2839.
8. Ho, C. J., Lea, C. T., & Stuber, G. L. (2001). Call admission control in the microcell/macrocell overlaying system. *IEEE Transactions on Vehicular Technology, 50*(4), 992–1003.
9. Choi, C. H., Kim, M. I., Kim, T. J., & Kim, S. J. (2001). A call admission control mechanism using mpp and 2-tier cell structure for mobile multimedia computing. In *Proc., conference on computer communications and networks*, pp. 581–584.

10. Ekici, E., & Ersoy, C. (2001). Multi-tier cellular network dimensioning. *Wireless Networks, 7*(4), 401–411.
11. Begain, K., Rozsa, G., Pfening, A., & Telek, M. (2002). Performance analysis of gsm networks with intelligent under lay-overlay. In *7th symposium on computers and communications*, pp. 135–141.
12. Dru, M.-A., & Saada, S. (2001). Location-based mobile services: The essentials. Tech. Rep., Alcatel Telecommunications Review, 1st Quarter 2001.
13. Nokia, http://www.nokia.com, *Mobile Location Services*, 2001.
14. Akyildiz, I. F., & Wang, W. (2002). A dynamic location management scheme for next-generation multitier pcs systems. *IEEE Transactions on Wireless Communications, 1*(1), 178–189.
15. Murphy, J. B., & Morgan, S. L. (1993). A satellite-based position location system for global data collection and messaging. In *Proc., MILCOM*, October 1993, Vol. 2.3, pp. 374–378.
16. Yeung, K. L., & Nanda, S. (1996). Channel management in microcell/macrocell cellular radio systems. *IEEE Transactions on Vehicular Technology, 45*(4), 601–612.
17. Jabbari, B., & Fuhrmann, W. (1997). Teletraffic modelling and analysis of flexible hierarchical cellular networks with speed sensitive handoff strategy. *IEEE Journal on Selected Areas in Communications, 15*(8), 1539–1548.
18. Lin, Y.-B., Mohan, S. & Noerpel, A. (1994). Queueing priority channel assignment strategies for pcs hand-off and initial access. *IEEE Transactions on Vehicular Technology, 43*(3), 704–712.
19. Redl, S. M., Weber, M. K., & Oliphant, M. W. (1995). *An introduction to GSM*. Artech House Mobile Communications Series. Artech House.
20. Thomas, R., Gilbert, H., & Mazziotto, G. (1988). Influence of the movement of the mobile stations on the performance of a radio cellular network. In *Proc., third nordic seminar on digital land mobile radio communications*, September 1988.

Vijoy Pandey received a B.Tech.(Hons.) degree from Indian Institute of Technology, Kharagpur, in 1995, and an M.S. degree from University of California, Davis in 1997. He is currently pursuing his Ph.D. degree at the above institution, while working in the Ethernet Switching Group at Blade Network Technologies in Santa Clara, California. At Davis, he was nominated for the Professors for the Future Fellowship Award in 1999. His research interests include architectures and protocols for next generation wireless networks, and intelligent packet switching for secure wired and wireless local area networks. He can be reached at: vijoy@bladenetwork.net

Dipak Ghosal received his B.Tech degree in Electrical Engineering from Indian Institute of Technology, Kanpur, India, in 1983, MS degree in Computer Science from Indian Institute of Science, Bangalore, India, in 1985, and Ph.D degree in Computer Science from University of Louisiana, Lafayette, USA, in 1988. From 1988 to 1990 he was a Research Associate at the Institute for Advanced Computer Studies at University of Maryland (UMIACS) at College Park, USA. From 1990 to 1996 he was a Member of Technical Staff at Bell Communications Research (Bellcore) at Red Bank, USA. Currently, he is with the faculty of the Computer Science Department at the University of California at Davis, USA. His research interests are in the areas of IP telephony, peer-to-peer systems, mobile and ad hoc networks, and performance evaluation of communication systems. He can be reached at: ghosal@cs.ucdavis.edu

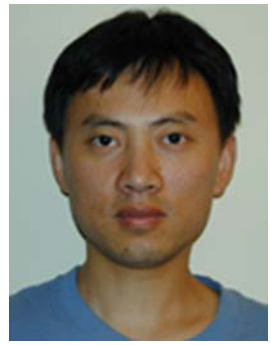
Biswanarh Mukherjee (S'82–M'87–F'07) received the B.Tech. (Hons) degree from Indian Institute of Technology, Kharagpur (India) in 1980 and the Ph.D. degree from University of Washington, Seattle, in June 1987. At Washington, he held a GTE Teaching Fellowship and a General Electric Foundation Fellowship. In July 1987, he joined the University of California, Davis, where he has been Professor of Computer Science since July 1995 (and currently holds the Child Family Endowed Chair Professorship), and served as Chairman of the Department of Computer Science during September 1997 to June 2000. He is winner of the 2004 Distinguished Graduate Mentoring Award at UC Davis. Two Ph.D. Dissertations (by Dr. Laxman Sahasrabuddhe and Dr. Keyao Zhu), which were supervised by Professor Mukherjee, were winners of the 2000 and 2004 UC Davis College of Engineering Distinguished Dissertation Awards. To date, he has graduated nearly 25 Ph.D. students, with almost the same number of MS students. Currently, he supervises the research of nearly 20 scholars, mainly Ph.D. students and including visiting research scientists in his laboratory.

Mukherjee is co-winner of paper awards presented at the 1991 and the 1994 National Computer Security Conferences. He serves or has served on the editorial boards of the IEEE/ACM

Transactions on Networking, IEEE Network, ACM/Baltzer Wireless Information Networks (WINET), Journal of High-Speed Networks, Photonic Network Communications, Optical Network Magazine, and Optical Switching and Networking. He served as Editor-at-Large for optical networking and communications for the IEEE Communications Society; as the Technical Program Chair of the IEEE INFOCOM '96 conference; and as Chairman of the IEEE Communication Society's Optical Networking Technical Committee (ONTC) during 2003-05.

Mukherjee is author of the textbook "Optical WDM Networks" published by Springer in January 2006. Earlier, he authored the textbook "Optical Communication Networks" published by McGraw-Hill in 1997, a book which received the Association of American Publishers, Inc.'s 1997 Honorable Mention in Computer Science. He is a Member of the Board of Directors of IPLocks, Inc., a Silicon Valley startup company. He has consulted for and served on the Technical Advisory Board (TAB) of a number of startup companies in optical networking. His current TAB appointments include: Teknovus, Intelligent Fiber Optic Systems, and LookAhead Decisions Inc. (LDI). He is a Fellow of the IEEE.

Mukherjee's research interests include lightwave networks, network security, and wireless networks. His e-mail address is: mukherje@cs.ucdavis.edu.



Xiaoxin Wu received his bachelor degree from Beijing university of Posts and telecommunications in 1990, with the major of wireless communications. In 2001, he received his Ph.D degree from Department of Electrical and Computer Engineering at University of California, Davis. His thesis title is "Achieving Quality of Service in Integrated Wireless Networks". Since 2002, he has been working with Department of Computer Science, Purdue University, as a postdoctoral research associate. His research involves in broad areas in wireless networks and integrated networks. Research topics include algorithm, protocol, and architecture design for improving network performance, security, and privacy. His major publication list can be found at http://www.cs.purdue.edu/homes/wu/HTML/research.html. In 2006, he joined Intel Communication Beijing Lab, working on network and security issues in WiMAX and digital health. Contact: xiaoxin.wu@intel.com.