

# A Call-Admission Control Algorithm (CACCA) for Providing Guaranteed QoS in Cellular Networks

Satya Kovvuri, Dipak Ghosal\*, Biswanath Mukherjee\*, and Dilip Sarkar  
 Department of Computer Science, University of Miami, Coral Gables, FL 33124

\*University of California, Davis, CA 95616

e-mails: {jyothi,sarkar}@cs.miami.edu, {ghosal, mukherjee}@cs.ucdavis.edu

*Abstract—*

Future broadband wireless access systems plan to integrate various classes of MTs (Mobile Terminals), each class with a different type of *quality of service* (QoS) requirement. When the load on a wireless network is high, the guarantee of QoS for each class of MTs is a challenging task. This study considers two classes of MT — *profiled* MTs and *non-profiled* or *regular* MTs. It is assumed that *profiled* users require a guaranteed QoS. The measure of QoS is the probability of forced termination of a call that was allowed to access the network. Two previous handoff prioritization schemes — (i) *pre-request scheme* and (ii) *guard channel scheme* — decrease handoff failure (and hence forced termination). In this work we compare and contrast both the schemes through extensive simulation; it is found that neither method can guarantee a desired level of QoS for the *profiled* MTs. We then propose a novel Call-Admission Control Algorithm (CACCA) that can maintain any desired level of QoS, while successful call completion rate is very high. In the proposed algorithm, the new call arrival rate is estimated continuously, and when the estimated arrival rate is higher than a predetermined level, some new calls are blocked irrespective of the availability of channels. The objective of this pre-blocking of calls is to maintain the wireless-network-system's observed new call arrival rate at no more than a predetermined rate. We show that the proposed method can guarantee any desired level of QoS for *profiled* users.

## I. INTRODUCTION

With increasing demand for mobile computing services and limited available bandwidth, wireless networks increase the number of simultaneous users in the network systems by reducing cell size. It is projected that in the future wireless networks will adopt micro/pico-cellular architecture [1]. However, smaller cell size naturally increases the number of handoffs a MT is expected to make. As new call arrival rate or *load* increases, so does the probability of handoff failure; this phenomenon, combined with the large number of handoffs before completion of a call, increases the forced termination probability of calls.

Future broadband wireless access systems plan to integrate various classes of MTs, each class with a different type of *quality of service* (QoS) requirement. This study considers two classes of MT — *profiled* MTs [2], and *non-profiled* or *regular* MTs; measure of QoS for the *profiled* MTs is the probability of forced termination of a call that was allowed to access the network. When the load on a wireless network is high, the guarantee of QoS for each class of MTs is a challenging task. Here it is assumed that the trajectory of *profiled* MTs is known, that is, the sequence of cells a MT crosses during the life of a call is known (say, from a

profile database).

Two generic handoff prioritization schemes are (i) advanced request for a channel — *pre-request scheme* [3], and (ii) reserving a number of channels for only handoff calls — *guard channel scheme* [3], [4], [5]. In general, handoff prioritization schemes decrease handoff failure (and hence forced termination), but increase call blocking. In this work, we compare and contrast both the schemes through extensive simulation; it is found that neither method can guarantee a desired QoS for the *profiled* MTs. Thus, novel methods are necessary for guarantee of QoS to *profiled* MTs.

We then propose a novel method that can maintain any desired level of QoS, while successful call completion rate is very high. In the proposed method, actual new call arrival rate is estimated continuously, and when the estimated new call arrival rate increases beyond a predetermined level, some new calls are blocked irrespective of the availability of channels. The objective of this pre-blocking of calls is to maintain the wireless-network-system's observed new call arrival rate at no more than a predetermined rate.

## II. PROFILED USERS AND CHANNEL ASSIGNMENT SCHEMES

We consider only Mobile Terminals (MTs)— they are moving at a certain speed and hence have the possibility of crossing over from one cell to another. Furthermore, MTs in this study are divided into two categories - (i) *Profiled MTs* who subscribe to a premium Quality of Service (QoS) and (ii) *Nonprofiled MTs*.

### A. Profiled Users

A user profile includes mobility patterns and services accessed. It is recorded against time of day, and is kept in the Home Location Register (HLR) database of the network for all the *profiled* users [6]. It is assumed that spatial and temporal information (the place and the time of travel) of the *profiled* users can be obtained from this database. A Base Station Controller (BSC) can use this database to perform profile-based channel allocation decisions, if necessary. The network utilizes profile information in the database for significantly improving QoS of the *profiled* MTs by reducing the forced termination probability. Any MT that is not included in the category of *profiled* MTs is in the category of *nonprofiled* MTs. We assume that the network's knowledge of the profile information is "perfect"; it knows a mobile's trajectory in time and space.

Two generic channel assignment schemes — the guard channel scheme [9], [3], [4], [5] and the channel pre-request scheme [3]— have been proposed. They can reduce force termination probability and improve QoS. Here we extensively study them to answer two questions that were not answered in previous studies: (i) How much QoS can they improve for the profiled users *and not all users*? (ii) Can they provide any desired level of QoS to the profiled users under variable loads? Next a brief description of each scheme is provided for clarity and completeness.

### B. Guard Channel Scheme

In this method, a number of wireless channels, say  $G$  out of total  $C$  channels, called the “guard channels”, are exclusively reserved for handoff calls of profiled users. The remaining channels, called the “normal channels”, are shared among all types of calls. By “all types of calls” we mean the new calls, the handoff calls of profiled users, and the handoff calls of nonprofiled users. New calls and nonprofiled handoff calls are accepted as long as a channel other than the guard channels is available. Profiled handoff calls are accepted until all the channels in the cell are occupied.

### C. Channel Pre-Request Scheme

Along the cell boundary, there exists an area where channels of more than one cell can service a MT. The channel pre-request scheme exploits this fact as described next. In the channel pre-request scheme, the neighboring cell to which a profiled user is moving into next can be obtained from the HLR database much before the user leaves the current cell. This information can then be used to pre-request a channel from a neighboring cell certain time (called *the reservation period*) prior to leaving the current cell. The reservation period may start at anytime if the profiled user is in the region of overlap between cells. It is assumed that trajectories of nonprofiled users are not available and hence channels for them cannot be pre-requested. By increasing the reservation period, the probability of the profiled user being forced terminated can be decreased.

## III. SIMULATION MODEL

Because of space limitation we omit details about our simulation model. We used 49 hexagonal cells with a wraparound topology. This topology is used, since it eliminates the boundary effect keeping exactly six neighbors for each cell [7]. We use a static channel allocation scheme for the cells, and hence, the number of channels allocated to a cell does not change during the simulation process. All cells receive the same number of channels, and the reuse distance is not a parameter in our study.

For our work, mobility of MTs is modeled using simple Brownian-motion or random-walk approximation. In this model, a MT moves to any of the current cell’s neighbors with equal probability -  $1/6$  for the hexagonal layout. It is assumed that both profiled and nonprofiled MTs are taking random-walk from cell to cell. However, since the trajectory of each profiled MT is known, the channels for it can be requested before it actually moves to the next cell.

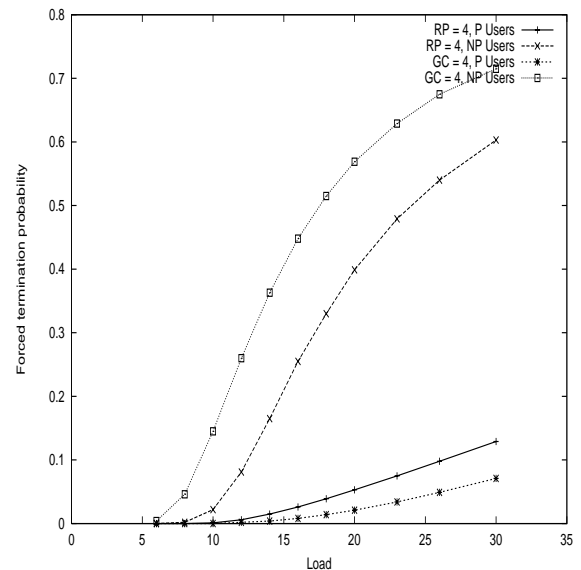


Fig. 1. Variations of  $P_{ft}$  between profiled and non-profiled MTs using the two schemes.

Residence time of a MT (profiled or nonprofiled) in a cell is exponentially distributed with mean  $1/\eta$  sec, that is, all MTs are mobile. In addition, we assume a fraction  $f$  of the users to be profiled and the other  $(1 - f)$  to be nonprofiled. And new call arrivals follow a Poisson distribution with  $\lambda$  calls/sec and the call holding time or total duration of the call follows an exponential distribution with mean  $1/\mu$  sec. Note that these are the most commonly used assumption for simulation and analysis. The *load* of a cell is the ratio of call arrival rate to call completion rate,  $\rho = \lambda/\mu$  Erlangs/cell.

Unless stated otherwise, we make the following assumptions. We simulate 300,000 calls using 20 channels per cell, 50% profiled users, 50% nonprofiled users, 120 seconds mean call holding time, 12 seconds call residence time, 4 seconds reservation period, and 4 guard channels.

### A. Performance Metrics

Three important metrics are used to assess the behavior of the cellular system modeled here. We study them for profiled MTs, nonprofiled MTs, and the whole system.

- *Total Blocking Probability*,  $P_b$ , is defined as the ratio of total number of calls blocked even before entering the cell to the total number of new-call attempts made.
- *Forced Termination Probability (FTP)*,  $P_{ft}$ , is defined as the ratio of the number of calls forced to terminate due to failed handoff to the number of mobile calls that successfully entered the network.
- *Successful Call Completion Rate*,  $SCCR$ , is defined as the number of calls that are successfully served to completion per unit time by each cell.

## IV. RESULTS AND DISCUSSIONS

Figure 1 shows that the forced termination probability for profiled users and nonprofiled users is increasing with

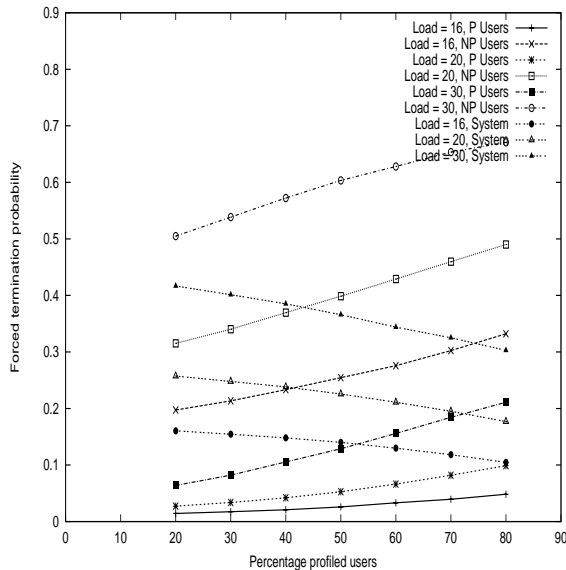


Fig. 2. Comparison of  $P_{ft}$  for different types of calls in the channel pre-request scheme.

an increasing load in both the schemes. This can be attributed to higher call arrival rate and hence higher channel occupancy.

It is found that  $SCCR$  increases in the channel pre-request scheme as the percentage of profiled users changes from 20% to 80%. The reason for this growth can be explained as follows: As the percentage of profiled users increase, the number of users who can get a channel successfully from the neighbor increases. The probability of forced termination for both profiled and nonprofiled users however, is found to increase (see Figure 2). The actual growth in  $P_{ft}$  is much higher for the nonprofiled users than for the profiled users. For instance, at load 16 with 20% profiled users,  $P_{ft}$  for profiled users is 0.01433 and that of nonprofiled users is 0.197216. For the same load, if the percentage of profiled users is increased to 80%, the corresponding forced termination probabilities are 0.0482161 and 0.332106. Net effect of this different growth rate is very interesting. Let us compute the weighted average of the forced termination probabilities for the whole system, using the equation,

$$P_{ft}(system) = P_{ft}(profiled) \times f + P_{ft}(nonprofiled) \times (1-f).$$

The values are 0.160532 and 0.104811 with 20% and 80% profiled users, respectively. We observed that  $P_{ft}$  of the system has decreased with an increase in the percentage of profiled users. This naturally increases the  $SCCR$ . Successful call completion rate with varying percentage of profiled users in guard channel scheme exhibits the same behavior.

In the channel pre-request scheme, when the effect of increasing the reservation period for a constant load is considered, we observe that with an increase in reservation time the  $P_{ft}$  of profiled users decreases drastically while that of nonprofiled users increases (see Figure 3). This is due to the fact that with an increase in the reservation period,

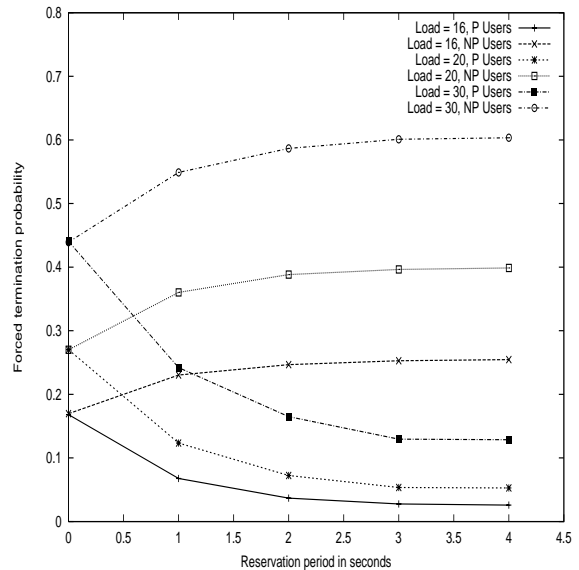


Fig. 3. Impact of reservation period on  $P_{ft}$  of calls in the channel pre-request scheme.

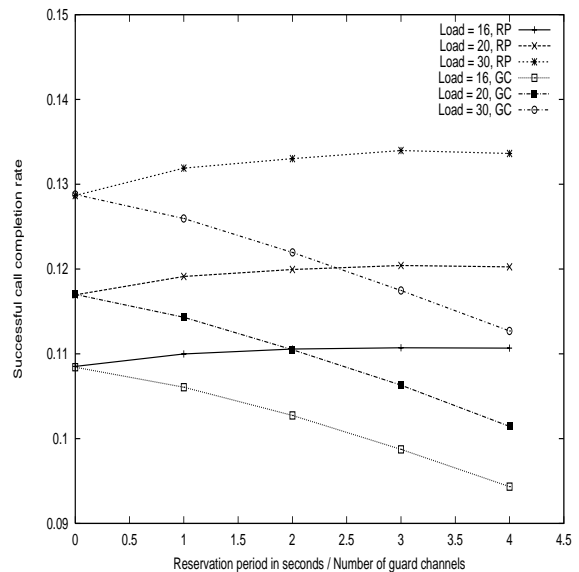


Fig. 4. Reduction of  $SCCR$  due to increase in the number of guard channels.

the profiled users get a larger time interval to be handoffed successfully to the neighbor. Hence, their probability of getting a channel from the neighbor increases. This naturally decreases the probability of nonprofiled users getting a channel from the cell. The observations for the guard channel scheme are very similar to that in the case of the channel pre-request scheme. In the guard channel scheme, however, it is observed that  $SCCR$  decreases continuously with an increase in the number of guard channels (see Fig. 4).

From the above discussion, it is evident that the guard channel scheme decreases the forced termination probability of the profiled handoff calls, but it also decreases the successful call completion rate of the system. However, the

advantage of this scheme is that it can assure better QoS to the profiled users if maximum load is known in advance.

The channel pre-request scheme has the advantage of making efficient use of the channels in the system. However, when the load goes beyond a predetermined value, it does not guarantee QoS for the profiled users. For system utilization, channel pre-request scheme is more desirable.

The schemes discussed above do not always guarantee desired QoS to profiled users while utilizing the system efficiently. In order to achieve both, we propose a **novel approach that can be used with the guard channel scheme as well as the channel pre-request scheme.**

## V. THE PROPOSED CALL-ADMISSION CONTROL ALGORITHM (CAC)

The proposed algorithm works by controlling the *load observed* by the system, irrespective of the *actual load* of the system. An acceptable maximum load  $\rho_m$  is determined either by simulations or by an analytical method. The value of  $\rho_m$  is specified before the system is operational. Note that, for any load above  $\rho_m$ ,  $P_{ft}$  of profiled users exceeds desired  $P_{ft}$  and hence fails to meet QoS for the profiled users.

During the operation of the system, the arrival rate and hence the expected load is estimated. If the estimated load is no more than  $\rho_m$ , attempts are made to allocate channels for all incoming new calls. Otherwise, the load is greater than  $\rho_m$ , and attempts are made to allocate channels for only a fraction,  $f_r$ , of the incoming new calls. The fraction  $f_r$  is calculated as  $f_r = \rho_m/\rho_o$  where  $\rho_o$  is the estimated load.

There are many possible ways for estimating the load  $\rho_o$ . The one used here assumes that information about the arrival times of the most recent  $N$  new calls is maintained by the system. Here  $N$  is called the ‘‘sample size’’ for estimation of load. If the arrival time of the first new call is  $t_1$  and the arrival time of the last new call is  $t_N$ , then  $\frac{N-1}{t_N-t_1}$  gives the estimated arrival rate  $\lambda$ . The value of  $\lambda/\mu$  where  $\mu$  is the service rate of the system, gives the estimated load  $\rho_o$ .

### A. Results and Observations

For the results reported here, we assume that the maximum allowed  $P_{ft}$  for profiled users is 2%. From our simulation, it is observed that in the original channel pre-request scheme, a load of 14 Erlangs/cell with a reservation period of four seconds satisfies this value of  $P_{ft}$ , while in the original guard channel scheme, a load of 14 Erlangs/cell with three guard channels guarantees  $P_{ft}$  of profiled users to be below 2%. So we assume that a load of 14 Erlangs/cell is the acceptable maximum load in the proposed method. Default sample size is 50.

Figure 5 shows the variation of forced termination probability with increasing load, when the CACA is used with the pre-request scheme. Forced termination probability increases for profiled as well as for nonprofiled users up to a certain load and then becomes almost constant. At low load, the system accepts almost all new calls. However,

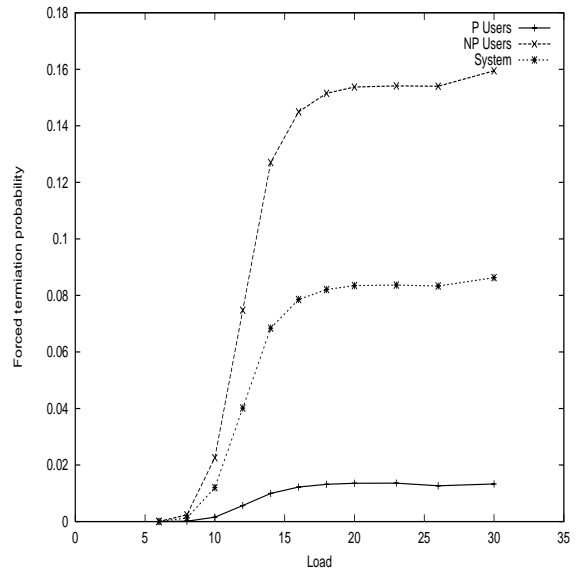


Fig. 5.  $P_{ft}$  in the system when the CACA is incorporated into the channel pre-request scheme.

the network rejects some new calls if the estimated load is greater than 14 Erlangs/cell. This call pre-rejection always keeps the observed load in the system to be about 14 Erlangs/cell. This stabilizes the forced termination probability of all types of users at a load greater than 14 Erlangs/cell. Observe that the  $P_{ft}$  for profiled users remains below 2% throughout, though the load increases. We tested even with such (unreasonably) high load as 300 Erlangs/cell, and the performance was as good! When the CACA is incorporated in the channel pre-request scheme,  $SCCR$ , initially increases up to load 14 Erlangs/cell and then becomes almost constant. When the estimated load is below 14 Erlangs/cell, the successful call completion rate increases with load, since no call or very few calls are rejected without any attempt to allocate channel. The reason for  $SCCR$  becoming almost constant above load 14 Erlangs/cell is: In order to keep  $P_{ft}$  of profiled users below 2%, new calls are not admitted into the system when the estimated load goes above 14 Erlangs/cell and so,  $P_b$  increases and  $(1 - P_b)$  decreases. But since the time taken for new call arrivals at high load is lower than the time taken for new call arrival at low load for the same number of calls, the number of calls successfully served per unit time is almost constant for any offered load above 14 Erlangs/cell.

When we tried to look at the effect of incorporating the CACA in the guard channel scheme, we observed that  $P_{ft}$  and  $SCCR$  exhibited the same behavior as in the case of incorporating the proposed method in the channel pre-request scheme. However,  $SCCR$  in guard channel scheme is lower than that of channel pre-request scheme when the CACA is used. This can be attributed to the improper utilization of guard channels in the that scheme.

When CACA is incorporated in channel pre-request scheme, the blocking probability of the system increases faster with load. Over load 14 Erlangs/cell, we observe an

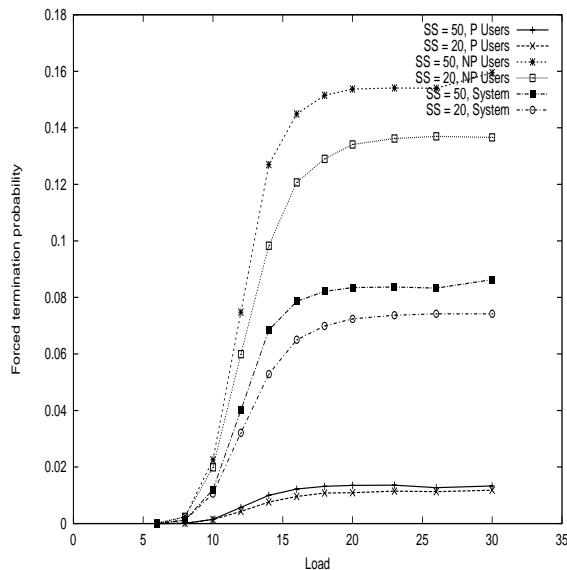


Fig. 6. Decrease in the  $P_{ft}$  with decrease in the sample size.

increase in the blocking probability of the system using the proposed method, when compared to the blocking probability of the system using the original channel pre-request scheme. This is expected due to the rejection of a fraction of new calls, even if a channel is available, when the load is above 14 Erlangs/cell.

We now discuss the effect of varying the sample size, on the system parameters. A lower sample size increases the blocking probability in the overall system. This naturally decreases the forced termination probability of all types of calls and the overall system (see Figure 6).

The reason for this is that, when the blocking probability increases at the same load, there is less competition for channels within the system; this decreases the  $P_{ft}$  of the calls within the system. However, we find that the  $SCCR$  with a smaller sample size is lower than that with a larger sample size. The reason for this is the increase in the  $P_b$ .

From the above discussion, it is evident that the CACA can guarantee Quality of Service to the profiled users. Moreover, it maintains the successful call completion rate of the system to be close to the maximum value.

## VI. CONCLUSION

In this paper, we evaluated two channel allocation schemes. We showed the advantages and disadvantages of using each of them. We then proposed a method called CACA (call-admission control algorithm), which guarantees QoS to profiled users at any load, that is, it can satisfy the target forced termination probability for profiled users at any load. It provides better service to profiled users while maintaining a very good successful call completion rate.

The system using the CACA works just like the system using the original scheme. The only differences are:

- When the estimated load  $\rho_o$  is below a certain value  $\rho_m$ , the proposed system accepts all new calls.

- When the estimated load  $\rho_o$  goes above the value  $\rho_m$ , it accepts the new calls selectively.

There are a number of questions that arose in the course of our study and are topics for the our current study.

## REFERENCES

- [1] A. S. Acampora, and M. Naghshineh, "Control and quality of service provisioning in high speed microcellular networks," *IEEE Personal Communications*, vol. 1, no. 2, pp. 36-43, 1994.
- [2] A. N. Rudrapatna, and C. Giardina, "Channel occupancy and network utilization implications for a profile-driven resource allocation scheme in cellular networks," in *Proceedings of Globecom*, pp. 2000-2011, November 1998.
- [3] Y. B. Lin, Seshadri Mohan, and A. Noerpel, "Queueing priority channel assignment strategies for PCS hand-off and initial access," *IEEE Transactions on Vehicular Technology*, vol. 43, pp. 704-712, no. 3, August 1994.
- [4] D. Hong, and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. VT-35, no. 3, pp. 77-92, August 1986.
- [5] Y. C. Kim, D. E. Lee, B. J. Lee, Y. S. Kim, and B. Mukherjee, "Dynamic channel reservation based on mobility in wireless ATM network," *IEEE Communications Magazine*, pp. 47-51, Nov. 1999.
- [6] V. Pandey, D. Ghosal, and B. Mukherjee, "Quantifying the benefits of exploiting user profiles in cellular networks," in *INFOCOM 2000*.
- [7] L. O. Guerrero, and A. H. Aghvami, "A prioritized handoff dynamic channel allocation strategy for PCS," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 4, pp. 1203-1215, July 1999.
- [8] O. T. W. Yu, and V. C. M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 7, pp. 1208-1225, September 1997.
- [9] R. Ramjee, D. Twosley, and R. Nagarajan, "On optimal call admission control in cellular networks," *ACM/Baltzer Wireless Networks (WINET)*, vol. 3, pp. 29-41, March 1997.