Dynamic Priority Queueing of Handover Calls in Wireless Networks: An Analytical Framework

Ariton E. Xhafa, *Member, IEEE,* and Ozan K. Tonguz, *Member, IEEE*

***Abstract*—In this paper, we present an analytical framework for dynamic priority queueing of handover calls in wireless networks. The framework employs a queueing discipline with two classes of priority for handover calls. Two queues, first priority and second priority, are employed for the two priority classes of handover calls. The priority of queued handover calls is not based only on the received signal strength, but also on the remaining time in the overlap region between two cells. We also incorporate a *priority transition* between handover calls in the queue; specifically, a second-priority handover call in the second-priority queue, based on certain criteria, can become a first-priority handover call and join the first-priority handover queue. In addition, the event that a handover call could finish its call while waiting in the queue is taken into account in the analysis. This event was not taken into consideration in previous related studies and, as a result, these previous studies overestimate handover failure probability.**

Our results also show that the predictions of the analytical framework developed in this paper are in very good agreement with simulation results. The developed analytical framework is comprehensive and can also cope with several priority schemes proposed by other researchers in the literature. For example, it is shown that, under certain conditions, the proposed framework converges to first-in–first-out queueing of handover calls. One can easily modify the proposed framework to incorporate priority schemes that use guard channels for handover calls. It is also shown that one could potentially use the framework developed in this paper in integrated voice/data networks, as well as for handover between different network types.

The proposed analytical framework is anticipated to be a very useful tool in evaluating performance of present and future wireless networks *employing dynamic priority queueing* for handovers and in designing more efficient handover algorithms.

***Index Terms*—Dynamic priority queueing, handover, wireless networks.**

I. Introduction

THE INCREASE of public interest and mass market for mobile communications and limited spectrum allocated by the Federal Communications Commission (FCC) is leading to smaller cell sizes in cellular networks. As a result, the number of mobile users crossing the cell boundaries is increasing; hence, the proliferation of handover calls. The way that handover calls are handled has a direct impact on the quality-of-service (QoS) provided to the mobile user (MU). Since dropping of a call in progress is less desirable than blocking a new call, various methods have been devised to prioritize handover calls over new calls [1]–[16]. An overview of priority schemes for handover calls was done by Posner and Guerin as early as 1985 in [17]. For later developments on handover priority schemes and work, the reader can refer to excellent references by Pollini in [7], Katzela and Naghshineh in [8], Tripathi *et al.* in [9], and Jabbari in [10].

To distinguish the work presented in this paper from the previous work done on handover priority schemes, we first provide a brief overview of the previous work on handover priority schemes.

One of the earliest analytical frameworks for guard channel method (GCM) was developed by Guerin in [3]. Guerin proposes a novel approach, where a certain number of channels is used exclusively for handover calls and *only queueing of originating (new) calls is investigated*. This approach, not only minimizes the handover blocking probability for the handover calls, but also increases the total carried traffic in the network. Simple closed-form expressions are provided for state probabilities; hence, the evaluation of the performance of the cellular network via these expressions is straightforward [3]. Daigle and Jain in [4] reconsider the approach proposed by Guerin in [3] and propose a novel and alternative analysis based on Neut's matrix approach. Hong and Rappaport in [5] develop an analytical framework for GCM with *first-in-first-out (FIFO) queueing of handover calls* and no queueing of originating (new) calls. Results show that the guard channel priority scheme with FIFO queueing of handover calls achieves smaller forced termination probability for handover calls compared to other schemes, thus reducing the number of dropped handover calls [5]. Chang *et al.* in [6] investigate a new cutoff priority scheme that allows finite queueing of both new and handover calls. In this approach, the handover and new calls are queued in two separate FIFO queues. In addition, Chang *et al.* in [6] consider reneging of new calls and dropping of queued calls as they move out of the handover area before the handover call is successful. Optimal cutoff parameters and appropriate queue sizes that minimize overall blocking probability are found numerically [6].

The aforementioned studies dealt with analytical frameworks for GCM and its variants, i.e., *with and without FIFO queueing*. While one could use FIFO queueing for new calls, this is not a good idea for handover calls. The reason for this is that mobile users move with different speeds, they stop at traffic lights, go into shopping malls, they accelerate and decelerate. Therefore, the handover requests need to be queued in such a way that the priority changes dynamically to account for the dynamics of the user motion. Therefore, a FIFO queueing scheme is not suitable for dealing with handover calls.

Manuscript received June 2, 2003; revised November 26, 2003.

O. K. Tonguz is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA (e-mail: tonguz@ece.cmu.edu; http://www.ece.cmu.edu/~tonguz).

A. E. Xhafa is with the Communications Systems Laboratory, Texas Instruments Inc., Dallas, TX 75243 USA (e-mail: axhafa@ti.com).

Digital Object Identifier 10.1109/JSAC.2004.826927

Tekinay and Jabbari in [11] study via simulations the performance of nonpreemptive priority queueing for handover calls, where, if a channel is released, then the handover call in the queue that has the lowest received signal strength gets served. It is shown that the proposed scheme, which is called measurement-based priority scheme (MBPS), outperforms FIFO queueing scheme under all traffic conditions [11]. However, the study in [11] does not take into account the dynamics of user motion. In [1] and [2], Ebersman and Tonguz investigate the dynamic queueing of handover calls using a signal prediction priority queueing (SPPQ) discipline, where the order of handover calls is not only based on the received signal strength (RSS), but also on the rate of change of RSS. The performance (i.e., new call blocking probability, forced termination probability, etc.) of a cellular system that uses SPPQ scheme is evaluated via extensive Monte Carlo simulations. The results show that SPPQ achieves smaller forced termination probability than FIFO queueing and MBPS, at the expense of slight increase in the new call blocking probability.

The simulation-based performance evaluation in [1], however, is quite cumbersome and time consuming. In addition, none of the studies mentioned above [3]–[6] provide an analytical approach (nor can the aforementioned studies be used) for evaluating the performance of personal communication systems (PCSs) employing *dynamic priority queueing [1] of handover calls*. To the best of authors' knowledge, a brief version of such a framework appears, for the first time, in [12]. In this paper, we describe the framework in detail and generalize it.

To employ dynamic priority queueing for handover calls, we propose a novel approach where two classes of priority for handover calls are considered and two queues, first-priority and second-priority queue are used for the two priority classes considered for handover calls. We also incorporate a *priority transition* between handover calls in the queue, specifically, a second-priority handover call in the second-priority queue can become a first-priority handover call and join the first-priority handover queue. In addition, the event that a handover call could finish its call while waiting in the queue is incorporated into the analysis. This event was not taken into account in previous studies [3]–[6] and our results show that this leads to overestimating the handover failure probability. Fantacci in [18] uses the generic channel holding time to describe the event that the handover call in a FIFO queue is still being served by its old base station until it gets service in the new cell, or is being dropped. However, *this event should be correlated to the channel holding time for handover calls and not the generic channel holding time, which is a function of new calls' and handover calls' channel holding times*. Recent field measurements have shown that the channel holding time for handover calls could be as low as 7 s [19], [20], as opposed to minutes, which is the generic channel holding time. Using generic channel holding time instead of channel holding time for handover calls leads to overestimating the handover failure probability. Our system model employs a two-dimensional (2-D) Markov chain approach and differs from the previously reported work as

- *two classes of priority for handover calls* are considered;
- the *transition time* required between these priorities is taken into account;
- *the event that a handover call could finish its call while waiting in the queue* is captured by our model.

We assume that call arrivals follow a Poisson process. While new call arrivals follow a Poisson process, the handover traffic is non-Poisson due to the blocking phenomenon in neighboring cells [21]. However, previous studies have shown that the *Poisson approximation* for handover traffic is a reasonable approximation when cells are identical, have the same statistical behavior, and the new call generation in the cell is a Poisson process [19], [21], [22]. In light of the aforementioned studies, in this paper, we approximate handover arrivals by a Poisson process. Results show that for Poisson arrivals and exponential channel holding time, the analytical framework developed in this paper is in very good agreement with the simulation results reported in [1]. It is also shown that FIFO queueing, which is widely used in handover priority schemes, is a special case of the analytical framework developed in this paper. Furthermore, it is shown that, under certain conditions, the proposed framework converges to the scheme proposed by Chang *et al.* in [6], which considers queueing of new and handover calls. The framework developed in this paper can easily be modified to incorporate priority schemes that use guard channels for handover calls. Therefore, GCM with or without FIFO queueing for handover calls, as well as dynamic priority queueing for handover calls can be analyzed via the proposed framework.

One can also use the generalized framework developed in this paper (see Section V) to analyze the handover performance of integrated voice/data networks. In this case, voice handover calls are assigned the first-priority, data handover calls are assigned the second-priority, and no priority transition occurs between the queues; hence, handover data calls in the queue are served only if a channel is released and no handover voice calls are in the first-priority queue. One could potentially use the framework developed in this paper to evaluate the handover performance between different network types (see Section V). Furthermore, the framework could be modified to incorporate handover calls whose priority is based not only on the RSS and the RSS's rate of change, but also on the bandwidth requirements, bit-error rate (BER), and other QoS requirements.

The remainder of this paper is organized as follows. In Section II, the problem under investigation is formulated. In Section III, we analyze and evaluate the system performance (i.e., blocking probability and handover dropping probability). In Section IV, we present comparisons between the numerical results based on the analytical framework developed in this paper and the simulation results reported in [1] and discuss the implications of these comparisons in Section V. Finally, conclusions are drawn in Section VI, while the necessary auxiliary material is relegated to the Appendices.

II. Problem Statement

In a real PCS environment, different MUs move with different speeds, they stop at traffic lights, go into shopping malls, they accelerate and decelerate. Therefore, the handover requests need to be queued in such a way that the priority changes dynamically

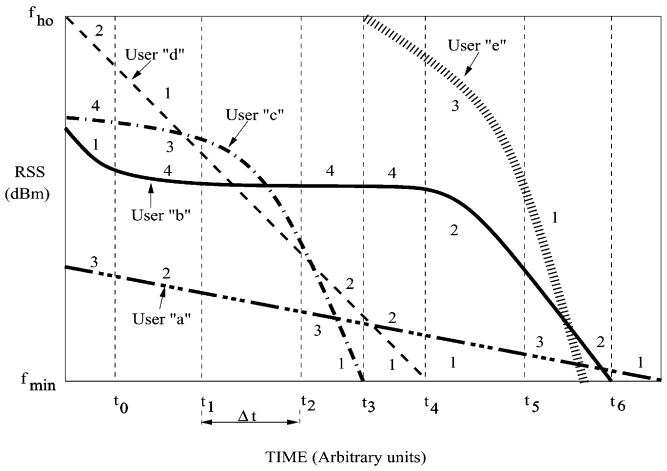

Fig. 1. Ordering of handover calls after [1].

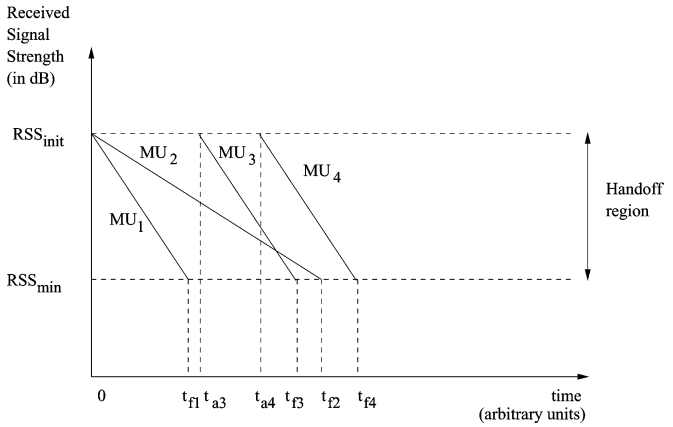

Fig. 2. Physical scenario under consideration.

to account for the dynamics of the user motion. This necessitates the use of dynamic priority queueing which, theoretically speaking, is a difficult problem [23].

In [1] and [2], the dynamic queueing of handover calls was not only based on the received signal strength (RSS), but also on the velocity at which a MU is moving; hence, the rate of change of RSS. The system performance was evaluated via Monte Carlo simulations. The main idea proposed in [1] is illustrated in Fig. 1. In a realistic scenario, RSS will undergo Rayleigh (short-term) and lognormal (long-term) fading; however, for illustration purposes RSS curves are given as in Fig. 1, essentially ignoring short-term fading. The numbers in the graphs represent relative handover ordering priority of each MU at a given time, while vertical dashed lines correspond to the intervals where the relative priorities could change for one or more MU [1]. For example, at time t_2 shown in Fig. 1, ordering of the users based on the RSS value from the highest to the lowest is as follows: user "b," user "c," user "d," and user "a." Assume that there are five separate RSS measurements made during Δt time interval, where $\Delta t = t_2 - t_1$. For each of these measurements, the RSS's rate of change is calculated as the change of RSS over the time interval it occurred. The RSS's rate of change at time t_2 is then found as the average of the five previously calculated RSS's rates of change. Therefore, one can estimate the remaining time in the queue for these users based on the RSS and the RSS's rate of change values at time t_2, as follows:

$$T_i|_{t_2} = \left.\frac{\mathrm{RSS}_i}{\Delta\mathrm{RSS}_i}\right|_{t_2} \tag{1}$$

where T_i, RSS_i, and $\Delta\mathrm{RSS}_i$ are the estimated remaining time in the queue, the RSS and the RSS's rate of change for user i at time t_2, respectively. Ordering of the users at time t_2 is done based on the estimated remaining time in the queue, e.g., the user with the lowest T_i has the highest priority. Hence, user priorities at time t_2 are as follows: user "b" has priority 4, user "a" has priority 3, user "d" has priority 2, and user "c" has priority 1, which is the highest priority. Observe that these priorities are drastically different from the priorities one obtains by merely measuring the RSS level at time t_2. The choice of the averaging time interval Δt is extremely important in determining priorities for the users at a given time [1].

While the study in [1] and [2] yielded accurate performance results, the Monte Carlo simulations are cumbersome and time consuming. In addition, the simulation-based performance evaluations do not provide physical insight into the impact of various key system parameters on the performance of the network. The main objective of this paper is to provide a generalized analytical framework for dynamic priority queueing, or equivalently, the signal prediction priority queueing scheme proposed in [1] and [2].

To simplify the problem, we assume there are only two classes of priority for handover calls, where priority depends on the waiting time in the queue. Fig. 2 depicts the physical scenario under investigation. Again, in a practical cellular system, RSS typically follow Rayleigh or lognormal distribution; however, for the sake of simplicity, RSS curves are given as straight lines in Fig. 2. The ordinate represents the received signal strength, while the abscissa represents the time. Since only two classes of priority for handover calls are considered, let their slopes be S_1 and S_2 for the first and the second class of priority, respectively. In Fig. 2, $\mathrm{RSS}_{\mathrm{init}}$ and $\mathrm{RSS}_{\mathrm{min}}$ depict the two critical levels employed in the two threshold level model used in the handover algorithm. When a mobile user moves out of the cell and/or when his received signal strength falls below $\mathrm{RSS}_{\mathrm{init}}$, the MU enters the handover region and "sends" a handover request. The call is dropped if no service is provided before the received signal strength from this user falls below $\mathrm{RSS}_{\mathrm{min}}$.

Let us consider the scenario shown in Fig. 2. Assume that at time $t = 0$ there are two handover requests, from MU_1 and from MU_2, and their slopes are S_1 and S_2, respectively. Let t_{f1} and t_{f2} be the times that MU_1 and MU_2 spent in the handover region. It should be noted that Fig. 2 merely shows two sample values for t_{f1} and t_{f2}. These times for the two priorities can be assumed to be random variables which, for simplicity, are assumed to follow an exponential distribution [2], [6]. Since t_{f1} is smaller than t_{f2} (or $S_1 > S_2$), MU_1 is assigned the first-priority, while MU_2 is assigned the second-priority handover request. Assume that a third handover request arrives from MU_3, which enters the region at time t_{a3}. We also assume that the rate of change for the received signal strength

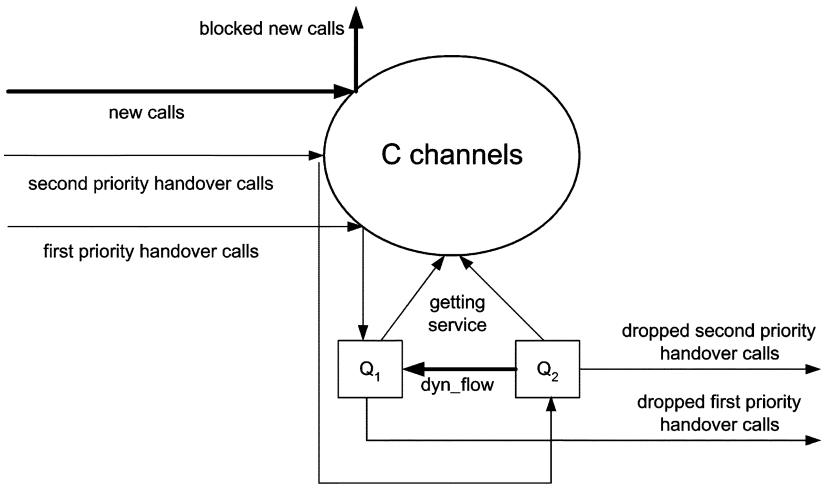

Fig. 3. Queueing model.

from MU_3 is S_1. Therefore, this request should be treated as a first-priority handover request. Let us consider another request coming from MU_4 at time t_{a4}, also having the rate of change for the received signal strength S_1. From Fig. 2, it is easy to observe that MU_2 will be out of the handover region before MU_4 is. Therefore, MU_2 must join the first-priority queue before MU_4 does. Hence, MU_2 should be in the first-priority queue at time $t = t_{a4}$.

It is clear that the total average time in the queue for a second-priority handover call should not change. Therefore, the mean of the transition time should be equal to the difference between the mean times in handover region of the second and first-priority, e.g., if mean transition time is t_t, then $t_t = t_{f2} - t_{f1}$. To make the analysis tractable, the distribution of the transition time is assumed to be exponential, although, in reality this is not the case.[1]

The described queueing model is shown in Fig. 3. We assume that there are C servers available (the number of channels per cell). The new and handover calls get service if there is a free server (i.e., a free channel). If all the servers are busy, then new calls are blocked, while handover calls are stored in the queues according to their priorities. We consider a finite storage for the queues (i.e., H_1, H_2 for first- and second-priority). If a handover request belonging to the first- (second-) priority queue finds H_1 (H_2) requests in the queue, this call is blocked; otherwise, it joins the queue which it belongs to. A handover call in the queue that does not get service before its waiting time is over, leaves the queue (i.e., the call is dropped). The priority of a handover request also depends on the waiting time in the queue; hence, there is a dynamic flow (dyn_flow) from queue Q_2 to queue Q_1, which manifests itself as the conversion of second-priority handover calls to first-priority handover calls.

Next, we present the analysis of this queueing scheme.

III. Analysis

Exact numerical calculation of handover performance of a cellular network is difficult due to large number of system states even for cellular networks that consist of few cells [22]. However, as shown in [22], the handover performance of a cell that is surrounded by a cluster of cells (i.e., there is handover traffic interaction between cells) and non-Poisson traffic is nearly identical to the handover performance of a single isolated cell when one assumes that the cells are identical, have the same statistical behavior and the traffic in the cells is Poisson. We will use the same assumptions in our analysis and consider the case when the distribution of channel holding time is exponential. This implies that one has to deal with an M/M/C/K queueing system [24]. Furthermore, we assume that the call duration time is a random variable that follows an exponential distribution with rate μ_M. We also assume that the mobiles are spread evenly over the area of the cell [5] and new calls follow a Poisson process with rate λ_n. Handover arrivals follow a Poisson process with rate $\lambda_h = \lambda_{h1} + \lambda_{h2}$, where λ_{h1} and λ_{h2} are the arrival rates for first and second-priority handover calls, respectively. Handover arrival rate λ_h can be found using flow equilibrium property [10]. Hence

$$\lambda_{h_{\text{out}}} = P_h(1 - P_B)\lambda_n + P_h(1 - P_H)\lambda_{h_{\text{in}}} \tag{2}$$

where P_h is the probability that a call in progress will experience a handover, $\lambda_{h_{\text{out}}}$ and $\lambda_{h_{\text{in}}}$ denote out-of-cell and into-cell handover rates, and P_B and P_H denote the new call blocking and handover failure probabilities, respectively. In equilibrium, $\lambda_{h_{\text{in}}} = \lambda_{h_{\text{out}}}$. Substituting λ_h for $\lambda_{h_{\text{out}}}$ and $\lambda_{h_{\text{in}}}$ in (2) and solving for λ_h, one gets

$$\lambda_h = \frac{P_h(1 - P_B)}{1 - P_h(1 - P_H)}\lambda_n. \tag{3}$$

The channel holding time T_H in a cell can be defined as the time interval between the time that a call starts occupying a channel and the time it releases the channel by either the completion of the call or a cell boundary crossing by a MU [5]. Channel holding time for a new/handover call is the minimum of the message duration time and the residual time of the new/handover call. Let us denote the channel holding times for new calls and handover calls as T_{Hn} (with mean $1/\mu_{Hn}$) and T_{Hh} (with mean $1/\mu_{Hh}$), respectively. Thus

$$T_{Hn} = \min(T_M, T_{rn}) \tag{4}$$

$$T_{Hh} = \min(T_M, T_{rh}) \tag{5}$$

[1] In fact, assuming that t_{f1} and t_{f2} are random variables and their distribution is exponential, then one can write $t_t = t_{f2} - t_{f1}$, and it can be easily shown that its distribution is not an exponential distribution.

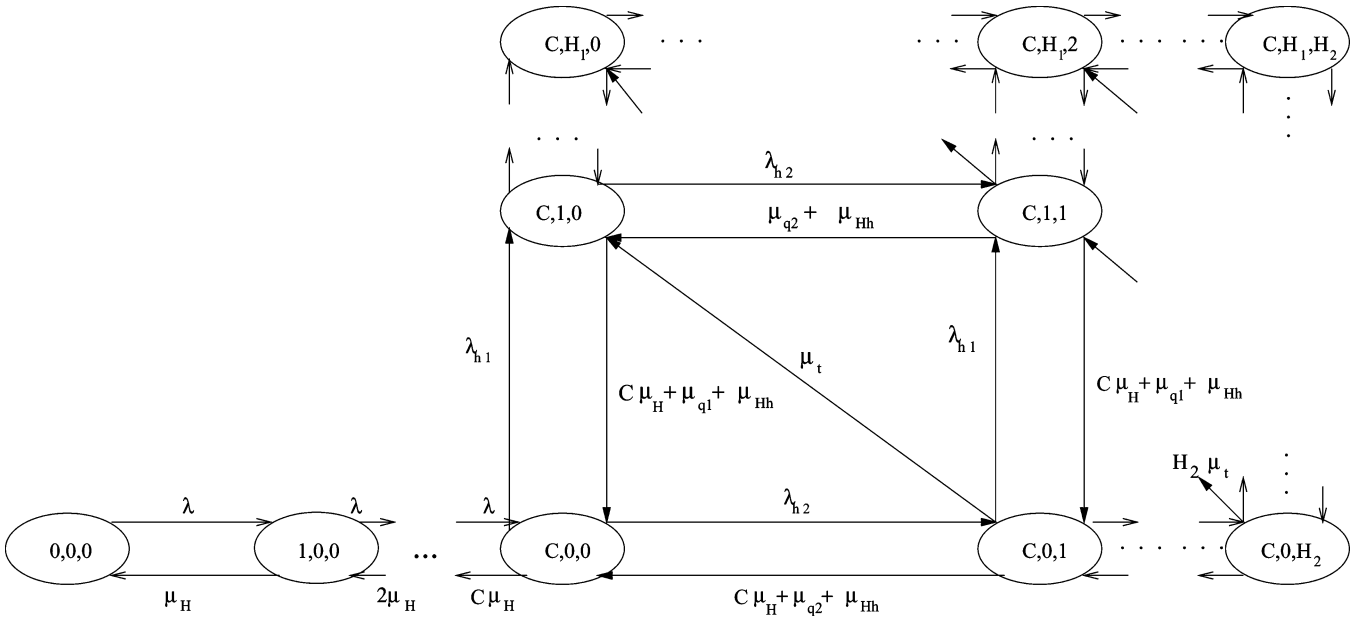

Fig. 4. Markov chain representation of the queueing model.

where T_M denotes the message duration, and T_{rn} and T_{rh} denote the residual times for new and handover calls, respectively. The call duration is assumed to follow an exponential distribution. For uniform speed distribution and uniform direction of movement, the residual times can be found as in [24] (refer to [24, Eqs. (47) and (48)]). Hence, the cumulative distribution function (cdf) for channel holding time is given as

$$F_{T_H}(t) = \frac{\lambda_n}{\lambda} F_{T_{Hn}}(t) + \frac{\lambda_h}{\lambda} F_{T_{Hh}}(t). \quad (6)$$

To make the analysis tractable, we approximate the distribution of T_H by an exponential distribution with mean $1/\mu_H$. One can write the following:

$$\int_0^\infty e^{-\mu_H t} dt = \int_0^\infty \left(1 - \frac{\lambda_n}{\lambda} F_{T_{Hn}}(t) - \frac{\lambda_h}{\lambda} F_{T_{Hh}}(t)\right) dt. \quad (7)$$

After some algebraic steps, one gets the parameter μ_H of the exponential distribution that is used to model the channel holding time distribution.

Let us now proceed with the performance analysis of the queueing system, which can be modeled as an M/M/C/K queue. In the case that there are free channels in the cell, new calls or handover calls are equally likely to get service. When all the channels are occupied, new calls are blocked whereas handover calls are queued in their respective queues according to their priorities. If the queues are full, handover call arrivals are dropped. Let us define $S_{k,m,n}$ as the state of the cell that has a total of k calls in progress, and m and n are first and second-priority handover calls in their respective queues. Fig. 4 shows the 2-D Markov chain. We have incorporated a *priority transition* between handover calls in the queue; specifically, a second-priority handover call in the second-priority queue can become a first-priority handover call and join the first-priority handover queue. This event happens when the remaining time in the handover area for the second-priority handover call is the same as the time that first-priority handover calls spend in the handover area. In addition, the event that a handover call could finish its call while waiting in the queue is incorporated in the analysis. This event, however, was not taken into account in previous studies [3]–[6] and our results show that this could lead to overestimating the handover failure probability (see Section IV). The study in [18] uses the generic channel holding time to describe the event that the handover call in the FIFO queue is still being served by its old base station until it gets service in the new cell, or is being dropped. However, *this event should be correlated to the channel holding time for handover calls and not the generic channel holding time which is a function of new calls and handover calls channel holding times*. Recent field measurements have shown that the channel holding time for handover calls could be as low as 7 s [19], [20], as opposed to minutes, which is the channel holding time. Using the average channel holding time $1/\mu_H$ instead of the average channel holding time for handover calls $1/\mu_{Hh}$ leads to overestimating the handover failure probability.

The steady-state probabilities $P_{k,m,n}$ that the cell is in state $S_{k,m,n}$ can be found by solving the system of linear equations consisting of the flow-equilibrium equations and the normalization condition $\sum_{k=0}^{C}\sum_{m=0}^{H_1}\sum_{n=0}^{H_2} P_{k,m,n} = 1$ (see Appendix I for details).

New call blocking occurs if a new call arrival finds C channels occupied; i.e., the cell is in state $S_{C,m,n}$, where $0 \le m \le H_1$ and $0 \le n \le H_2$. Therefore, the steady-state blocking probability for the new calls P_B can be expressed as

$$P_B = \sum_{m=0}^{H_1}\sum_{n=0}^{H_2} P_{C,m,n}. \quad (8)$$

Handover failure occurs if a handover call arrival finds all channels occupied and its respective queue full *or* the handover call arrival is queued in its respective queue; however, it is dropped before getting service because its waiting time in the queue is

over before the handover call gets served or finishes its service. The steady-state handover failure probability P_H is given as the sum of the handover failure probability for each class weighed by the probability that this call is a first or a second-priority handover call. Hence

$$P_H = \frac{\lambda_{h1}}{\lambda_h} P_{H/1} + \frac{\lambda_{h2}}{\lambda_h} P_{H/2} \tag{9}$$

where $P_{H/1}$ and $P_{H/2}$ are the conditional probabilities of the event that a first-priority and a second-priority handover call are dropped, respectively. The event that a first-priority handover call is blocked/dropped occurs if a first-priority handover arrival finds its queue full, *or* joins the queue, but its waiting time is over before the call gets served or finishes its service. Therefore, one has

$$P_{H/1} = \sum_{n=0}^{H_2} P_{C,H_1,n} + \sum_{m=0}^{H_1-1} \sum_{n=0}^{H_2} P_{H/1;m,n} P_{C,m,n} \tag{10}$$

where the first term describes the event that the first-priority handover queue is full, while the second term describes the event that the first-priority handover call is queued, but it is dropped before getting service because its waiting time is over before a channel is released. The term $P_{H/1;m,n}$ gives the probability of handover failure for a first-priority handover call in the queue given the handover call joined the queue as the $(m+1)$ call. This is found as

$$P_{H/1;m,n} = 1 - P_{SH/1;m,n} \tag{11}$$

where $P_{SH/1;m,n}$ is the probability of a successful handover request in the first-priority handover queue given it joined the queue as the $(m+1)$ call, i.e., the handover request gets served *or* finishes its call before it gets service or before its is dropped. After some algebra, one gets [12], [24]

$$P_{H/1;m,n} = \frac{(m+1)\mu_{q1}}{C\mu_H + m(\mu_{q1} + \mu_{Hh})}. \tag{12}$$

The event that a second-priority handover call is blocked/dropped occurs if a second-priority handover arrival finds its queue full, *or* joins the queue, but its waiting time is over before the call gets served or finishes its service. Therefore, one has

$$P_{H/2} = \sum_{n=0}^{H_1} P_{C,m,H_2} + \sum_{n=0}^{H_2-1} \sum_{m=0}^{H_1} P_{H/2;m,n} P_{C,m,n} \tag{13}$$

where the first term describes the event that the second-priority handover queue is full, while the second term describes the event that the second-priority handover call is queued, but it is dropped before getting service because its waiting time is over before a channel is released. The term $P_{H/2;m,n}$ gives the probability of handover failure for a second-priority handover call in the queue given the handover call joined the queue as the $(n+1)^{st}$ call and is given as

$$\begin{aligned} P_{H/2;m,n} = {} & 1 - Tr(S_{C,0,0}/S_{C,m,n}) \\ & - Pr[2^{\text{nd}} \text{ priority handover call finishes call/} \\ & \quad \text{it joined as}\}(n+1)^{\text{st}} \text{ call}] \end{aligned} \tag{14}$$

where $Tr(S_{C,0,0}/S_{C,m,n})$ is the transfer function that describes the event that the second-priority handover call that joined as the $(n+1)$ call in the queue reaches state $S_{C,0,0}$, thus gets service, while the second term describes the event that the second-priority handover call that joined as the $(n+1)$ call in the queue finishes its service before it reaches state $S_{C,0,0}$ *or* being dropped. The procedure to calculate $P_{H/2;m,n}$ is similar to the one followed for the calculation of $P_{H/1;m,n}$; however, it is slightly more complex than that of first-priority handover call procedure. For more details the reader is referred to Appendix II.

Next, we present numerical results obtained via (8) and (9) for the system performance when dynamic queueing is used and discuss their implications.

IV. Results

Before proceeding with the numerical results obtained via the framework developed in this paper, we investigate the impact of channel holding time for handover calls in the performance of the framework given in [6], where queueing of new and handover calls is considered. Our framework converges to the framework in [6] under the assumption that the transition rate is zero and no guard channels are used. The assumptions made in this case are as follows: each cell has 20 channels, total traffic in the cell is kept constant at 15 Erlangs, the channel holding time for the new calls follows an exponential distribution with mean 1.5 min, the channel holding time for handover calls follows an exponential distribution with mean 30 and 20 s (we consider two scenarios for the channel holding time for handover calls), the time that a handover call spent in the handover area is 10 s, while reneging time for the new calls is 20 s. Fig. 5 shows the overestimation of handover failure probability in percentage for the scheme developed in [6] versus the handover arrival rate given as the percentage of the total traffic. The overestimation percentage is calculated as

$$\text{Over Estim.} = \left| \frac{P_H(\text{new}) - P_H(\text{existing})}{P_H(\text{new})} \right| \times 100\%. \tag{15}$$

It is clear from Fig. 5 that as the channel holding time for handover calls decreases, the discrepancy between the new approach (i.e., the approach where the channel holding time for handover calls is used to describe the event that the handover call in the queue could finish its service before it is served in the new cell or before it is dropped) and the previous approach becomes more than 20% for handover traffic that is larger than 35%.

We also compared the results obtained for the handover failure probability for two scenarios: i) when channel holding time for handover calls is used and ii) when generic channel holding time is used to describe the event that a handover call in the queue could finish its service before it is served in the new cell or before it is being dropped [18]. Scenario ii) resulted in the handover failure probability being higher by 7.2% and 12.5% compared with the results obtained in scenario i), for 30 and 20 s channel holding times for handover calls, respectively. The simulation results for the same scenario showed an overestimation of 6.8% and 12.1%, respectively. These results clearly show that, in order to obtain accurate results, one has to use channel holding time for handover calls as opposed to the generic channel holding time.

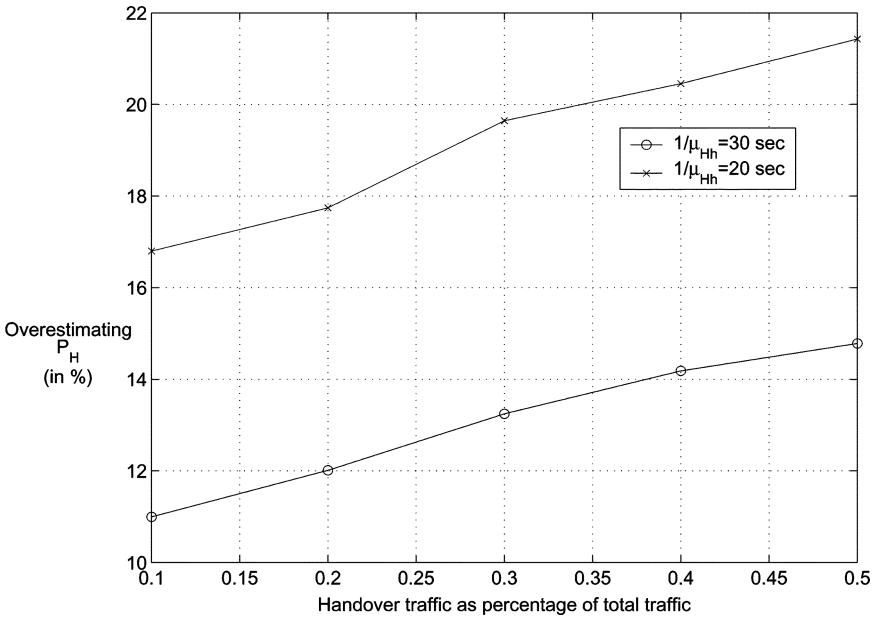

Fig. 5. Overestimation of handover failure probability versus handover traffic as percentage of total traffic.

TABLE I
STEPS FOLLOWED TO CALCULATE P_H AND P_B

1) Input parameters: C, λ_n, μ_{Hn}, μ_{Hh}, P_h, μ_{q1}, and μ_{q2}.
2) Calculate μ_H and μ_t.
3) Assume a value for λ_h.
4) Write down the equations for the Markov chain and solve for $P_{C,m,n}$.
5) Calculate P_B and P_H given by Eqns. (8) and (9).
6) Calculate new λ_h and compare it to the previous λ_h: if it does not converge, go to step 4 and use the new λ_h value computed.

To present numerical results for the system performance when dynamic priority queueing is used, we consider a cellular architecture whereby all cells have the same size. It is assumed that 30% of the handover calls are first-priority handover calls and that capacity for each queue is three. The algorithm used in calculating new call blocking and handover dropping probabilities is given in Table I.

Fig. 6 shows the comparison between the simulation results reported in [1] and analytical approach developed in this paper. The parameters such as message duration, probability of a handover, etc., were taken from the data given in [1]. In Fig. 6, the average channel holding time is 1 min, the average channel holding time for handover calls is 0.5 min, the average waiting times in the queue are 2 and 12 s for first and second-priority handover calls, respectively, probability of a call in progress experiencing a handover is 50%, and the cell has 30 channels. Comparing the blocking and handover failure probabilities, one can see that the agreement between the simulations in [1] and analytical results is very good (better than 96%).

In Fig. 7 the comparison is done for a different number of channels in the cell. In this case, it is assumed that there are 40 available channels. Here, the average channel holding time is 1 min, the average channel holding time for handover calls is 0.5 min, the average waiting times in the queue are 2 and 12 s for first and second-priority handover call, respectively, the probability of a call in progress experiencing a handover is 50%. Again, the agreement between the simulations in [1] and the analytical results is better than 96% for different traffic values. We did check the numerical results obtained from the proposed framework for different scenarios and an excellent agreement was observed between the results obtained from the developed framework and the simulations, however, due to space limitations, in this paper, we omit these results.

Next, we investigate special cases of the generalized framework developed in this paper.

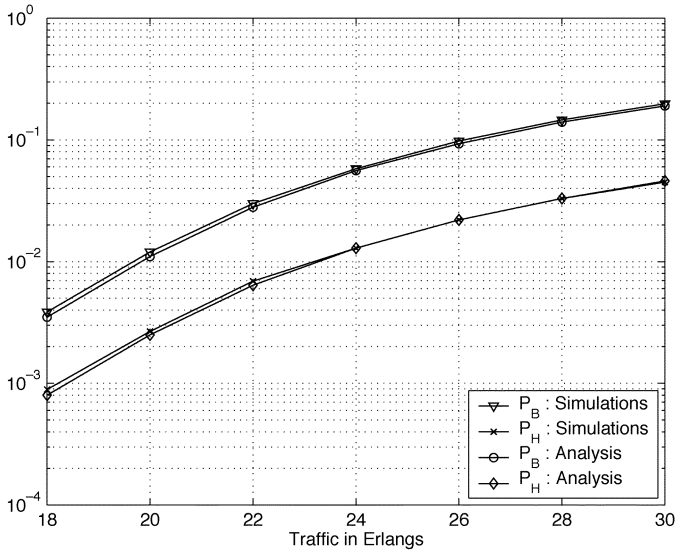

Fig. 6. Comparison of new call blocking and handover failure probabilities between the analytical approach and simulation results reported in [1]. In this case, the average channel holding time is 1 min, the average channel holding time for handover calls is 0.5 min, the average waiting time in the queue are 2 and 12 s for first and second-priority handover calls, probability of a call in progress experiencing a handover is 50%, and the cell has 30 channels.

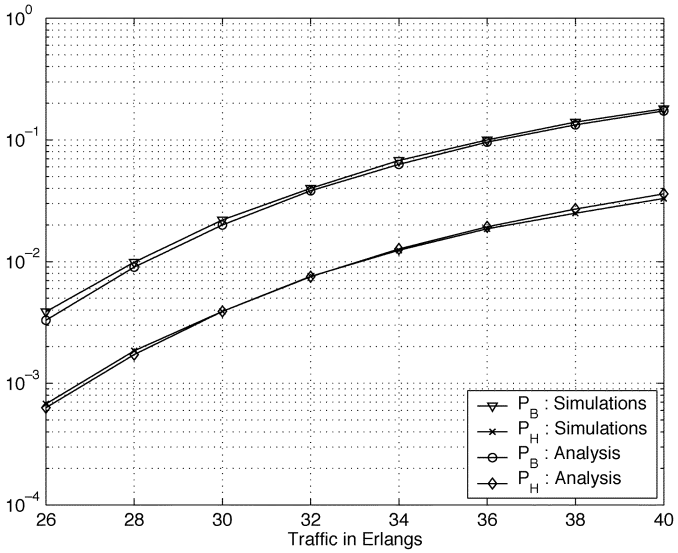

Fig. 7. Comparison of new call blocking and handover failure probabilities between the analytical approach and simulation results reported in [1]. Here, the average channel holding time is 1 min, the average channel holding time for handover calls is 0.5 min, the average waiting time in the queue are 2 and 12 s for first- and second-priority handover calls, the probability of a call in progress experiencing a handover is 50%, and the cell has 40 channels.

Case 1: Transition Rate Goes to Infinity ($\mu_t \rightarrow \infty$)

In this case, as soon as a second-priority handover call joins the second queue, it makes the transition to the first-priority queue, provided there is enough room in the first-priority handover queue. Hence, this implies that there is only one queue for handover calls, thus, the model converges to a FIFO queueing model. For example, assuming that the cell has 20 channels, the capacity of the FIFO queue is three, the channel holding time is 1 min, channel holding time for handover calls is 0.5 min, handover traffic makes 40% of the total traffic, waiting time in the queue is 10 s, using a FIFO queue one gets a new call blocking, and handover failure probability of 0.057 and 0.0139, respectively, while using the analytical framework developed in this paper, one gets a new call blocking and handover failure probability of 0.0565 and 0.0137, respectively. Thus, the results show that when the transition rate goes to infinity (i.e., $\mu_t \rightarrow \infty$) the model developed in this paper converges to a FIFO queueing model. This is an important sanity check that validates the accuracy of our framework.

Case 2: Transition Rate Goes to Zero ($\mu_t \rightarrow 0$)

In this case, since no transitions occurs, there are two separate queues for handover calls. If one lets $\lambda_n = \lambda_{h2}$, $\lambda_h = \lambda_{h1}$, and μ_{Hh} is not included in the second-priority queue (because new calls are not being served, while handover calls are served by their old base station even though they are queued in the new cell), then the framework developed in this paper reduces to the framework proposed by Chang *et al.* in [6] when no guard channel for handover calls are used. This shows that the developed analytical framework is very general and comprehensive in the sense that previous queueing approaches for handover calls reported by other researchers can also be handled by our framework as special cases.

Case 3: Arrival Rates for First-/Second-Priority is Zero ($\lambda_{h1} \rightarrow 0$ *or* $\lambda_{h2} \rightarrow 0$)

This implies that λ_{h1} (or λ_{h2}) is zero. Hence, the framework converges to a simple FIFO queueing case. Indeed, for the same scenario described in *Case 1*, we obtained the same results under the assumption that λ_{h1} (or λ_{h2}) is zero. Thus, once again, the excellent agreement in the results obtained via the framework developed in this paper and FIFO queueing, validates the accuracy of our framework, as well as showing its comprehensive and general nature.

It is important to mention here that the framework can easily be modified to take into account the use of guard channels for handover calls.

V. Discussion

In this paper, we developed an analytical framework for dynamic priority queueing of handover calls. The framework is based on a delay-dependent queueing discipline and employs a 2-D Markov chain to calculate the new call blocking and handover failure probabilities. Steady-state probabilities, new call blocking probability, and handover failure probability for first-priority handover calls can be obtained numerically once the Markov chain is solved. Handover failure probability for second-priority handover calls can be obtained numerically once the transfer functions given in (14) are obtained using Mason's formula (see Appendix II for details). In third-generation (3G) wireless networks, priority strategies may require more than two priorities for handover calls. Extension of the proposed framework to more than two priorities is straightforward, since the solutions to the Markov chain in this case is still numerically feasible. The transfer functions for lower priorities (equal or greater than two) can be calculated using Mason's formula, for which there exist subroutines in MATLAB.

Our numerical results (numerical evaluations of (8) and (9) given in Section III) are obtained in a few seconds as opposed to the many hours needed for the Monte Carlo simulations reported in [1]. Results show that for Poisson arrivals and exponential channel holding time, the analytical approximation developed in this paper is in very good agreement with the simulation results reported in [1]. Based on the above comparisons, one can see that the queueing model described in this paper gives a good approximation to the dynamic behavior of handover call requests in a PCS environment and is in good agreement with the simulation results reported in [1] for different scenarios described therein.

In order to develop the Markovian model, several time parameters are assumed to follow exponential distributions. Next, we address the validity of these assumptions.

- *Channel holding time distribution*: Recent field measurements have shown that the channel holding time follows a lognormal (or mixed lognormal, depending on the type of the environment) distribution as opposed to exponential distribution. The impact of lognormal and mixed lognormal distributed channel holding time on the handover performance of priority schemes such as DPQ, GCM, and GCM-FIFO is reported in [29] and [30]. Results show

that under the assumption of Poisson arrivals, the handover performance of these priority schemes is almost the same when one assumes a mixed lognormal distribution channel holding time as opposed to an exponential distribution. Therefore, the exponential distribution assumption for channel holding time provides good insights and good approximations for the handover performance of the aforementioned priority schemes.

- *Transition time between priority calls*: In the network simulator, we implement only one queue and when a channel is released, the handover call with the smallest remaining time in the queue is served first. Thus, the transition time does not exist in the network simulator as a separate parameter. This parameter is introduced in the analytical framework and the results show that we achieve good accuracy by representing the dynamics of the queue via the transition time between priority calls.
- *Queue waiting time*: The common assumption in the literature is that queue waiting time follows an exponential distribution. While one could assume that the queue waiting time follows other distributions such as, lognormal, truncated gaussian, etc., we believe that the impact (on the handover performance) of these queue waiting time distributions will be negligible.

Next, we illustrate two scenarios, where one could potentially use the approach developed in this paper to handle handover call arrivals.

A. Scenario 1: Integrated Voice/Data Wireless Networks

In integrated voice/data systems, handover voice calls are assigned first-priority and handover data calls are assigned the second-priority. Furthermore, queue waiting time for nonreal time services, such as e-mail, is not as crucial as the waiting time for voice and real time applications. This implies that there are two queues for handover calls and a first-priority handover call is always served if a channel is released. This also means that there is no transition from a second-priority handover call in the second-priority queue to the first-priority queue, which implies that a second-priority handover call can get served if a channel is released and there are no handover calls waiting in the first-priority queue. This model corresponds to *Case 2* in Section IV (i.e., $\mu_t \rightarrow 0$) and can potentially be studied using the approach developed in this paper.

B. Scenario 2: Handover Between Different Network Types

Future wireless networks will consist of a set of overlapping tiers, each with its own specific characteristics [25]–[27]. Therefore, fast lossless handover between different network types (i.e., "vertical handoff") will be crucial to the realization of seamless mobile multimedia networks. In general, high-data rate networks have smaller coverage areas than the lower data rate networks; e.g., the coverage area of a wireless local area network (WLAN) is much smaller than that of a cellular network. Consider the handover arrivals at a base station (BS) of a general packet radio service (GPRS) network. The queue waiting time for handover arrivals from a WLAN network is smaller than that of handover arrivals from other cells; e.g., cellular networks that use Global System for Mobile Communications (GSM) or code-division multiple-access (CDMA) technologies. Therefore, one could use the approach developed in this paper to handle handover calls in the queue by assigning first-priority to handover arrivals from the WLAN and second-priority to handover arrivals from the other GPRS cells.

In packet switched networks, arrivals (i.e., connection requests) have different rates and different resource demands from different traffic types [31]. Assuming that a channel is the smallest unit that a BS can allocate, different connection types may request different number of channels; hence, the Markov chain presented in this paper should be modified to take this issue into account. For example, the first connection request could demand one channel while the second connection request may demand five channels because it is a video streaming connection request. Since the arrivals come from different traffic types, the priority of the handover calls waiting in the queue should depend not only on the queue waiting time, but also on the traffic type; hence, the QoS requirements of queued calls as well. Therefore, the transition time between priority calls should be such that it takes into account not only the queue waiting time, but also the QoS requirements of queued calls. In addition, the most important performance metrics for packet switched networks are delay and throughput, as opposed to new call blocking and handover failure probabilities. Further research is needed to develop an analytical framework for handover priority schemes in wireless data networks.

VI. Conclusion

In this paper, we present an analytical framework for dynamic priority queueing for handover calls. To employ the dynamic priority queueing of handover calls, we proposed a novel approach where two classes of priority for handover calls are considered and two queues are used to distinguish between priority classes for handover calls. We also incorporate a *priority transition* between handover calls in the queue, specifically, a second-priority handover call in the second-priority queue can become a first-priority handover call and join the first-priority handover queue under certain conditions. In addition, the event that a handover call could finish its call while waiting in the queue is taken into account in the analysis. Our system model employs a 2-D Markov chain approach. Steady-state probabilities, new call blocking, and handover failure for first and second-priority handover calls can be obtained numerically. Results show that for Poisson arrivals and exponential channel holding time, the analytical approximation developed in this paper is in very good agreement with the simulation results reported in [1].

It is also shown that performance of other queueing schemes for handover calls reported by other researchers can be analyzed by the framework developed in this paper. In this sense, results clearly indicate that the analytical framework reported in this paper is much more comprehensive and general than other analytical approaches previously reported. For example, FIFO queueing, which is widely used in handover priority schemes, is a special case of the analytical framework developed in this paper. Furthermore, it is shown that under certain conditions,

the framework developed in this paper converges to the framework proposed by Chang *et al.* in [6], which considers queueing of new and handover calls. One can also modify the framework developed in this paper to incorporate other priority schemes that use guard channels for handover calls as well.

This work provides the basis of a new comprehensive analytical framework for dynamic queues for handover requests in a PCS environment. It is anticipated that such an analytical framework could be a very useful tool for assessing the performance of PCS employing dynamic priority queueing for handovers, and for designing more efficient handover algorithms for current and future wireless networks.

APPENDIX I
STEADY-STATE EQUATIONS FOR THE MARKOV CHAIN

In this appendix, we derive the steady-state equations of the system.

Fig. 3 shows the Markov-chain model of the system considered in this paper. If there is a free channel, no distinction between new and handover calls is made, provided the queues are empty. If the queues are not empty, then the channel is assigned to first-priority handover calls on a FIFO basis; if the first-priority queue (Q_1) becomes empty, this implies that there are no first-priority handover requests and, therefore, the next free channel is assigned to second-priority handover calls on a FIFO basis.

Let us define $S_{k,m,n}$ as the state of the cell that has a total of k calls in progress, and m and n are first and second-priority handover calls in their respective queues. The transition between states can be explained as follows.

- A transition from state $S_{k,0,0}$ to $S_{k+1,0,0}$ for $0 \le k < C$ occurs when a new call or handover call arrives, thus it occurs with rate λ.
- A transition from state $S_{k,0,0}$ to state $S_{k-1,0,0}$ for $0 < k \le C$ occurs if a call in progress finishes its service and releases the channel, thus occurs with rate $k\mu_H$.
- When all channels are busy, a transition to the next states occurs if there is a first or second-priority handover call arrival *and* the first or second-priority queue is not full. Hence, a transition from state $S_{C,m,n}$ to state $S_{C,m+1,n}$ occurs with rate λ_{h1}, while a transition from state $S_{C,m,n}$ to state $S_{C,m,n+1}$ occurs with rate λ_{h2}.
- A transition from state $S_{C,m,n}$ to state $S_{C,m-1,n}$ occurs if a channel is released *and* the first-priority handover call gets service *or* the first-priority handover call finishes its call while in the queue, *or* the waiting time in the queue for a handover call in first-priority is over before a channel is released, thus occurs with rate $C\mu_H + m(\mu_{q1} + \mu_{Hh})$, where $1/\mu_{q1}$ is the mean waiting time in the queue for a first-priority handover call which is assumed (for simplicity) to have a negative exponential distribution.
- A transition from state $S_{C,m,n}$ to state $S_{C,m,n-1}$ occurs if the waiting time for a second-priority handover call is over before a channel is released *or* the second-priority handover call finishes its call while in the queue, *or* a channel is released and a second-priority handover call gets served provided there is no handover call waiting in first-priority handover queue, thus it occurs with rate $n(\mu_{q2} + \mu_{Hh})$ or with rate $C\mu_H + n(\mu_{q2} + \mu_{Hh})$.
- A transition from state $S_{C,m,n}$ to state $S_{C,m+1,n-1}$ occurs if a second-priority handover call "becomes" a first-priority handover call, thus it occurs with rate $n\mu_t$.

Based on the above descriptions and Fig. 3, the steady-state equations describing this model, albeit tedious, are straightforward

$$(\lambda_n + \lambda_h)P_{k,m,n} = (k+1)\mu_H P_{k+1,m,n} \quad \text{for } 0 \le k < C, m = 0, \quad n = 0 \tag{16}$$

$$(\lambda_h + C\mu_H)\,P_{C,m,n} = (C\mu_H + (m+1)(\mu_{q1} + \mu_{Hh}))\,P_{C,m+1,n} + (\lambda_h + \lambda_n)P_{C-1,m,n} + (C\mu_H + (n+1)(\mu_{q2} + \mu_{Hh}))\,P_{C,m,n+1} \quad \text{for } m{=}0, \ n{=}0 \tag{17}$$

$$(\lambda_h + C\mu_H + m(\mu_{q1} + \mu_{Hh}))\,P_{C,m,n} = (C\mu_H + (m+1)(\mu_{q1} + \mu_{Hh}))\,P_{C,m+1,n} + (n+1)\mu_t P_{C,m-1,n+1} + (n+1)(\mu_{q2} + \mu_{Hh})P_{C,m,n+1} + \lambda_{h1}P_{C,m-1,n} \quad \text{for } 0 < m < H_1, \quad n = 0 \tag{18}$$

$$(C\mu_H + m(\mu_{q1} + \mu_{Hh}) + \lambda_{h2})\,P_{C,m,n} = \lambda_{h1}P_{C,m-1,n} + (n+1)\mu_t P_{C,m-1,n+1} + (n+1)(\mu_{q2} + \mu_{Hh})P_{C,m,n+1} \quad \text{for } m = H_1, \quad n = 0 \tag{19}$$

$$(C\mu_H + \lambda_h + n(\mu_t + \mu_{q2} + \mu_{Hh}))\,P_{C,m,n} = (C\mu_H + (n+1)(\mu_{q2} + \mu_{Hh}))\,P_{C,m,n+1} + (C\mu_H + (m+1)(\mu_{q1} + \mu_{Hh}))\,P_{C,m+1,n} + \lambda_{h2}P_{C,m,n-1} \quad \text{for } m = 0, \quad 0 < n < H_2 \tag{20}$$

$$(C\mu_H + \lambda_h + m(\mu_{q1} + \mu_{Hh}) + n(\mu_t + \mu_{q2} + \mu_{Hh}))\,P_{C,m,n} = \lambda_{h1}P_{C,m-1,n} + (n+1)\mu_t P_{C,m-1,n+1} + (C\mu_H + (m+1)(\mu_{q1} + \mu_{Hh}))\,P_{C,m+1,n} + \lambda_{h2}P_{C,m,n-1} + (n+1)(\mu_{q2} + \mu_{Hh})P_{C,m,n+1} \quad \text{for } 1 \le m < H_1, \quad 0 < n < H_2 \tag{21}$$

$$(C\mu_H + m(\mu_{q1} + \mu_{Hh}) + n(\mu_{q2} + \mu_{Hh}) + \lambda_{h2})\,P_{C,m,n} = \lambda_{h1}P_{C,m-1,n} + \lambda_{h2}P_{C,m,n-1} + (n+1)\mu_t P_{C,m-1,n+1} + (n+1)(\mu_{q2} + \mu_{Hh})P_{C,m,n+1} \quad \text{for } m = H_1, \quad 0 < n < H_2 \tag{22}$$

$$(C\mu_H + \lambda_{h1} + m(\mu_{q1} + \mu_{Hh}) + n(\mu_t + \mu_{q2} + \mu_{Hh}))\,P_{C,m,n} = (C\mu_H + (m+1)(\mu_{q1} + \mu_{Hh}))\,P_{C,m+1,n} + \lambda_{h1}P_{C,m-1,n} + \lambda_{h2}P_{C,m,n-1} \quad \text{for } 0 < m < H_1, \quad n = H_2 \tag{23}$$

$$(C\mu_H + \lambda_{h1} + n(\mu_t + \mu_{q2} + \mu_{Hh}))\,P_{C,m,n} = (C\mu_H + (m+1)(\mu_{q1} + \mu_{Hh}))\,P_{C,m+1,n} + \lambda_{h2}P_{C,m,n-1} \quad \text{for } m = 0, \quad n = H_2 \tag{24}$$

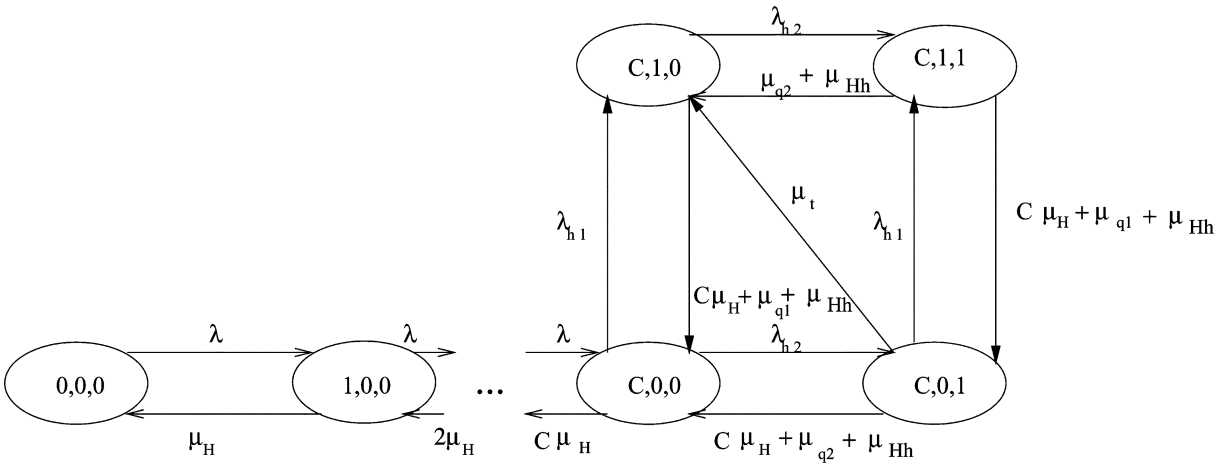

Fig. 8. Markov chain for calculating $Tr(S_{C,0,1}/S_{C,0,0})$.

$$(C\mu_H + m(\mu_{q1} + \mu_{Hh}) + n(\mu_{q2} + \mu_{Hh}))\, P_{C,m,n} = \lambda_{h1} P_{C,m-1,n} + \lambda_{h2} P_{C,m,n-1} \quad \text{for } m = H_1, \quad n = H_2. \tag{25}$$

The steady-state probabilities $P_{k,m,n}$ that the cell is in state $S_{k,m,n}$ can be found by solving the system of linear equations consisting of the flow-equilibrium equations and the normalization condition $\sum_{k=0}^{C} \sum_{m=0}^{H_1} \sum_{n=0}^{H_2} P_{k,m,n} = 1$.

Appendix II
Probability of Success for a Second Priority Handover Call in the Queue

In this Appendix, we derive the handover failure probability $P_{H/2;m,n}$ for a second-priority handover call in the queue given it joined the queue at the $(n+1)^{st}$ position. This is found as

$$P_{H/2;m,n} = 1 - P_{SH/2;m,n} \tag{26}$$

where $P_{SH/2;m,n}$ is the probability of a successful handover request in the second-priority handover queue given it joined the queue at the $(n+1)^{st}$ position, i.e., the handover request gets served. Hence

$$P_{H/2;m,n} = 1 - Pr[\text{handover call is successful/ it joins as } (n+1)^{\text{st}}\text{call}]. \tag{27}$$

The probability that the handover call in the queue is successful can be found as the sum of the probability that the handover call reaches state $S_{C,0,0}$ given that handover call joined the queue at the $(n+1)^{st}$ position and the probability that the handover call in the queue finishes its call before it gets service *or* before it is dropped. The probability that the handover call reaches state $S_{C,0,0}$ given the handover call joins the queue at the $(n+1)^{st}$ position can be found using signal-flow graph concepts and Mason's formula [12].

Calculating the probability that the handover call reaches state $S_{C,0,0}$ for second-priority handover calls, however, differs from the first-priority handover calls. The reason is that a second-priority handover call may get service as a first-priority call, *or* as a second-priority call. Then, this probability (which, in fact, is a transfer function) from a state $S_{C,m,n}$ will be sum of two transfer functions which represent the case that the call we are interested in, gets service as a first-priority call, *or* as a second-priority call. Also, a first-priority handover call arrival will affect the overall position of the call of interest in the queue.

To illustrate the idea, consider the case when the queue size is one for each priority and the cell is in state $S_{C,0,1}$ and we are after the probability that the handover call reaches state $S_{C,0,0}$, thus gets service (see Fig. 8).

To find the transition probabilities from state $S_{C,0,1}$ (without letting the call under consideration finish its call while waiting in the queue or being dropped) to other states, one can proceed as follows.

- A transition from state $S_{C,0,1}$ to state $S_{C,0,0}$ occurs if a channel is released and the handover call gets service. The reason why we look at the case that the handover call is *not dropped* or it *did not finish* its call while waiting in the queue, is because we are interested that the handover call under consideration reaches state $S_{C,0,0}$ and gets service. Therefore, any other handover call in the queues other than the call under consideration could be dropped or finish its call. However, during the same time, a first-priority handover call might arrive, *or* the handover call under consideration could be dropped *or* switch priorities. Since the random variables involved follow an exponential distribution and are independent, it can be shown that probability that a channel is released is given as

$$Pr\,(S_{C,0,0}/S_{C,0,1}) = \frac{C\mu_H}{C\mu_H + \mu_{q2} + \mu_{Hh} + \mu_t + \lambda_{h1}}. \tag{28}$$

- A transition from state $S_{C,0,1}$ to state $S_{C,1,1}$ implies that a first-priority handover call arrival occurs. Thus

$$Pr\,(S_{C,0,1}/S_{C,1,1}) = \frac{\lambda_{h1}}{C\mu_H + \mu_{q2} + \mu_{Hh} + \mu_t + \lambda_{h1}}. \tag{29}$$

- A transition from state $S_{C,0,1}$ to state $S_{C,1,0}$ implies that the second-priority handover call becomes first-priority handover call. Thus

$$Pr\,(S_{C,1,0}/S_{C,0,1}) = \frac{\mu_t}{C\mu_H + \mu_{q2} + \mu_{Hh} + \mu_t + \lambda_{h1}}. \tag{30}$$

- A transition from state $S_{C,1,1}$ to state $S_{C,0,1}$ occurs if a channel is released *and* the first-priority handover call gets

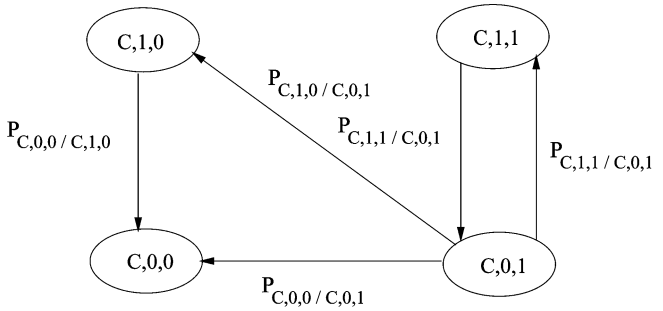

Fig. 9. Simplified graph for calculating $Tr(S_{C,0,1}/S_{C,0,0})$.

service, *or* the first-priority handover call is dropped before a channel is released, *or* the first-priority handover call finishes its service while in the queue. Thus, it occurs with probability

$$Pr\left(S_{C,1,1}/S_{C,0,1}\right)=\frac{C\mu_H+\mu_{q1}+\mu_{Hh}}{C\mu_H+\mu_{q1}+\mu_{q2}+2\mu_{Hh}}. \quad (31)$$

- One can proceed in the same way to find the transition probabilities from other states to some other state.

Based on the above explanation, Fig. 9 shows the simplified graph for the example under consideration. Applying Mason's formula to this graph

$$Tr\left(S_{C,0,0}/S_{C,0,1}\right)=\frac{\sum P_k\Delta_k}{\Delta} \quad (32)$$

where $Tr(S_{C,0,0}/S_{C,0,1})$ is the transfer function describing the event that the handover call under consideration reaches state $S_{C,0,0}$; P_k are all forward paths going from state $S_{C,0,1}$ to state $S_{C,0,0}$; and $\Delta = 1 -$ (sum of all single loops) $+$(sum of all two nontouching single loops) $-$(sum of all three nontouching single loops) $+$(sum of all four nontouching single loops)$\ldots$, whereas Δ_k-s are Δ-s which do not include loops touched by the kth forward path. For this particular example one has

$$\Delta=1-\frac{\lambda_{h1}}{C\mu_H+\mu_{q2}+\mu_{Hh}+\mu_t+\lambda_{h1}}\times\frac{C\mu_H+\mu_{q1}+\mu_{Hh}}{C\mu_H+\mu_{q1}+\mu_{q2}+2\mu_{Hh}} \quad (33)$$

and $\Delta_k=1$. For the forward paths, one has

$$P_1=\frac{C\mu_H}{C\mu_H+\mu_{q2}+\mu_{Hh}+\mu_t+\lambda_{h1}} \quad (34)$$

$$P_2=\frac{\mu_t}{C\mu_H+\mu_{q2}+\mu_{Hh}+\mu_t+\lambda_{h1}}\times\frac{C\mu_H}{C\mu_H+\mu_{q1}+\mu_{Hh}}. \quad (35)$$

One can substitute (33)–(35) into (32) to find the probability that the handover call that joined the queue at the first position gets service.

To find the probability that the handover call that joins the queue at the first position finishes its call before it gets service *or* before it is being dropped, consider the following.

- The handover call can finish its call while waiting in state $S_{C,0,1}$ with probability $\mu_{Hh}/(C\mu_H+\mu_{q2}+\mu_{Hh}+\mu_t+\lambda_{h1})$.
- The handover call can finish its call after becoming a first-priority handover call; thus it happens with probability $\mu_{Hh}/(C\mu_H+\mu_{q1}+\mu_{Hh})Pr(S_{C,1,0}/S_{C,0,1})$.
- The handover call can finish its call when the system is in state $S_{C,1,1}$; thus, it happens with probability $\mu_{Hh}/(C\mu_H+\mu_{q1}+\mu_{q2}+2\mu_{Hh})Pr(S_{C,1,1}/S_{C,0,1})$.
- From state $S_{C,1,1}$ it can go to state $S_{C,0,1}$ and the procedure is repeated.

After some algebraic manipulations one gets

$$Pr\left[hc\text{ finishes call/joined as }1st\right]=\frac{1}{\Delta}\left\{\frac{\mu_{Hh}}{C\mu_H+\mu_{q2}+\mu_{Hh}+\mu_t+\lambda_{h1}}+\frac{\mu_{Hh}}{C\mu_H+\mu_{q1}+\mu_{Hh}}Pr\left(S_{C,1,0}/S_{C,0,1}\right)+\frac{\mu_{Hh}}{C\mu_H+\mu_{q1}+\mu_{q2}+2\mu_{Hh}}Pr(S_{C,1,1}/S_{C,0,1})\right\}. \quad (36)$$

Hence, one can calculate $P_{H/2;0,1}$ as

$$P_{H/2;0,1}=1-Tr\left(S_{C,0,0}/S_{C,0,1}\right)-Pr\left[\text{handover call finishes call/it joined as 1st call}\right]. \quad (37)$$

One can proceed in a similar way to find the probability that the second-priority handover call will reach state $S_{C,0,0}$, thus gets service, given it joined the queue at the $(n+1)^{st}$ position.

ACKNOWLEDGMENT

The authors would like to thank the three anonymous reviewers whose comments helped to significantly improve this manuscript.

REFERENCES

[1] H. G. Ebersman and O. K. Tonguz, "Handoff ordering using signal prediction priority queueing in personal communication systems," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 20–35, Jan. 1999.
[2] H. G. Ebersman, "A novel handoff ordering scheme in mobile and personal communication systems: Signal prediction priority queueing," M.Sc. thesis, State Univ. New York, Buffalo, Nov. 1994.
[3] R. A. Guérin, "Queueing-blocking system with two arrival streams and guard channels," *IEEE Trans. Commun.*, vol. 36, pp. 153–163, Feb. 1988.
[4] J. N. Daigle and N. Jain, "A queueing system with two arrival streams and reserved servers with application to cellular telephones," in *Proc. IEEE INFOCOM'92*, 1992, pp. 2161–2167.
[5] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. VT-35, pp. 77–92, Aug. 1986.
[6] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping," *IEEE/ACM Trans. Networking*, vol. 2, pp. 166–175, Apr. 1994.
[7] G. P. Pollini, "Trends in handover design," *IEEE Commun. Mag.*, vol. 34, pp. 82–90, Mar. 1996.
[8] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *IEEE Pers. Commun.*, vol. 3, pp. 10–31, June 1996.
[9] N. Tripathi, J. H. Reed, and H. F. VanLandighan, "Handoff in cellular systems," *IEEE Pers. Commun.*, vol. 5, pp. 26–37, Dec. 1998.
[10] B. Jabbari, "Teletraffic aspects of evolving and next-generation wireless communication networks," *IEEE Pers. Commun.*, vol. 3, pp. 4–9, Dec. 1996.
[11] S. Tekinay and B. Jabbari, "A measurement-based prioritization scheme for handovers in mobile cellular networks," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 1343–1350, Oct. 1992.

[12] A. Xhafa and O. K. Tonguz, "Dynamic priority queueing of handoff requests in personal communication systems: An analytical framework for performance evaluation," in *Proc. IEEE Personal, Indoor, Mobile Radio Communications (PIMRC'99)*, Osaka, Japan, Sept. 1999, pp. 1346–1350.

[13] Y. B. Lin, S. Mohan, and A. Noerpel, "Queueing priority channel assignment strategies for PCS handoff and initial access," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 704–712, Aug. 1994.

[14] G. Senarath and D. Everitt, "Performance of handover priority and queueing systems under different handover request strategies for microcellular mobile communication systems," in *Proc. IEEE 45th Vehicular Technology Conf.*, Chicago, IL, 1995, pp. 897–901.

[15] A. J. Ransom, "Handoff consideration in microcellular systems planning," in *Proc. IEEE Personal, Indoor, Mobile Radio Communications (PIMRC'95)*, Sept. 1995, pp. 804–808.

[16] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE J. Select. Areas Commun.*, vol. 14, pp. 711–717, May 1996.

[17] E. C. Posner and R. Guerin, "Traffic policies in cellular radio that minimize blocking of handoff calls," in *Proc. 11th Teletraffic Congress (ITC 11)*, vol. 1, Kyoto, Japan, Sept. 1985, pp. 2.4B-2-1–2.4B-2-5.

[18] R. Fantacci, "Performance evaluation of prioritized handoff schemes in mobile cellular networks," *IEEE Trans. Veh. Technol.*, vol. 49, pp. 485–493, Mar. 2000.

[19] C. Jedrzycki and V. C. M. Leung, "Probability distribution of channel holding time in cellular telephony systems," in *Proc. IEEE Vehicular Technology Conf. (VTC'96)*, vol. 1, 1996, pp. 247–251.

[20] F. Barcelo and J. Jordan, "Channel holding time distribution in public telephony systems," *IEEE Trans. Veh. Technol.*, vol. 49, pp. 1615–1625, Sept. 2000.

[21] E. Chlebus and W. Ludwin, "Is handoff traffic really Poissonian?," in *Proc. 4th IEEE Int. Conf. Universal Personal Communications (ICUPC '95)*, Tokyo, Japan, Nov. 1955, pp. 348–353.

[22] P. V. Orlik and S. S. Rappaport, "On the handover arrival process in cellular communications," *ACM/Baltzer Wireless Networks*, vol. 7, no. 2, pp. 147–157, Mar./Apr. 2001.

[23] L. Kleinrock, *Queueing Systems Volume II: Computer Applications*. New York: Wiley, 1975, ch. 3.

[24] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures-Version 2a," College of Eng. Appl. Sci., State Univ. New York at Stony Brook, Tech. Rep. 773, June 1, 1999.

[25] S. Ohmori, Y. Yamao, and N. Nakajima, "The future generations of mobile communications based on broadband technologies," *IEEE Commun. Mag.*, pp. 134–142, Dec. 2000.

[26] L. Taylor, R. Titmus, and C. Lebre, "The challenges of seamless handover in future mobile multimedia networks," *IEEE Pers. Commun.*, vol. 6, pp. 32–37, Apr. 1999.

[27] K. Pahlavan *et al.*, "Handoff in hybrid mobile data networks," *IEEE Pers. Commun.*, vol. 7, pp. 34–46, Apr. 2000.

[28] Z. Gajić and M. Lelicć, *Modern Control System Engineering*. Englewood Cliffs, NJ: Prentice-Hall, 1996, ch. 2.

[29] A. E. Xhafa, "Analysis, design, and implementation of handover priority schemes in wireless networks," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, Oct. 2003.

[30] A. E. Xhafa and O. K. Tonguz, "Does mixed lognormal channel holding time affect the performance of guard channel scheme?," presented at the *IEEE Global Telecommunications Conf. (GLOBECOM'03)*, San Francisco, CA, Dec. 2003.

[31] T. Janevski, *Traffic Analysis and Design of Wireless IP Networks*. Norwood, MA: Artech House, 2003.

Ariton E. Xhafa (S'98–M'04) was born in Karbunar, Vlore, Albania. He received the B.S. degree in electrical and electronics engineering and physics from Eastern Mediterranean University, North Cyprus, Turkey, in 1997, the M.S. degree in electrical engineering from State University of New York at Buffalo, in 1999, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University (CMU), Pittsburgh, PA, in 2003.

From 2000 to 2001, he was a Visiting Researcher at CMU, doing research on design and performance analysis of communication systems. From 2002 to 2003, he was a Research Assistant in the Telecommunications Research Group, Center for Wireless and Broadband Networking, CMU, doing research on design and performance evaluation of wireless networks. From January 2004 to April 2004, he was a Research Associate at CMU, doing research on scheduling, handover management, and resource allocation in wireless networks. Currently, he is a Member of Technical Staff at the Communication Systems Laboratory, Texas Instruments Inc., Dallas. His current research interests include design and evaluation of multiple-access control (MAC) protocols, handover management, and QoS in wireless networks.

Dr. Xhafa is a Student Member of the IEEE Communications Society, the IEEE Vehicular Technology Society, and the IEEE Computer Society.

Ozan K. Tonguz (S'86–M'90) was born in Nicosia, Cyprus. He received the B.Sc. degree from the University of Essex, Colchester, U.K., and the M.Sc. and Ph.D. degrees in electrical engineering from Rutgers University, New Brunswick, NJ.

He is currently a tenured Full Professor in the Department of Electrical and Computer Engineering, Carnegie Mellon University (CMU), Pittsburgh, PA. Before joining CMU in August 2000, he was with the Electrical and Computer Engineering Department, State University of New York at Buffalo (SUNY). He joined SUNY in 1990 as an Assistant Professor, where he was granted early tenure and promoted to Associate Professor in 1995, and to Full Professor in 1998. Prior to joining academia, he was with Bell Communications Research (Bellcore), Red Bank, NJ, between 1988–1990, doing research in optical networks and communication systems. He has published in the areas of optical networks, wireless communications and networks, and high-speed networking. He is author or coauthor of more than 150 technical papers in IEEE journals and conference proceedings, and a book chapter (New York: Wiley, 1999). His contributions in optical networks and wireless networks are internationally acclaimed. He was also the architect of the "High Performance Waveform (HPW)" that was implemented in Harris RF Communications' AN/PRC-117f multiband man-pack tactical satellite radio. His industrial experience includes periods with Bell Communications Research, CTI, Inc., Harris RF Communications, Aria Wireless Systems, Clearwire Technologies, Nokia Networks, and Asea Brown Boveri (ABB). He currently serves as a consultant for several companies, law firms, and government agencies in the U.S. and Europe in the broad area of telecommunications and networking. He is also a Co-Director (Thrust Leader) of the Center for Wireless and Broadband Networking Research, CMU. His current research interests are in optical networks, wireless networks and communication systems, high-speed networking, and satellite communications.

Dr. Tonguz, in addition to serving on the Technical Program Committees of several IEEE conferences and symposia in the area of wireless communications and optical networks, currently serves or has served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS MAGAZINE and the JOURNAL OF LIGHTWAVE TECHNOLOGY. He was a Guest Editor of the special issue of the JOURNAL OF LIGHTWAVE TECHNOLOGY and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS on Multiwavelength Optical Networks and Technology, published in 1996.