Efficient Estimation of Call Blocking Probabilities in Cellular Mobile Telephony Networks with Customer Retries

Marco Ajmone Marsan, Giovanni De Carolis, Emilio Leonardi, Member, IEEE, Renato Lo Cigno, Associate Member, IEEE, and Michela Meo, Associate Member, IEEE

Abstract—A novel approximate technique is proposed for the estimation of call blocking probabilities in cellular mobile telephony networks where call blocking triggers customer retries. The approximate analysis technique is based on Markovian models with state spaces whose cardinalities are proportional to the maximum number of calls that can be simultaneously in progress within cells. The accuracy of the approximate technique is assessed by comparison against results of detailed simulation experiments, results of a previously proposed Markovian analysis approach, and upper and lower bounds to the call blocking probability. Numerical results show that the proposed approximate technique is very accurate, in spite of the remarkably small state spaces of the Markovian models.

Index Terms—Blocking probability, call reattempts, cellular networks, Markovian models, mobile telephony, repeated calls, retrial queues.

I. INTRODUCTION

Several techniques based on Markovian models were proposed for performance evaluation, design and planning of cellular mobile telephone networks [1]–[7], global system for mobile communications (GSM) is probably the most common example of such networks. In most of those techniques, the cells in the network are studied one by one with stochastic models that evolve according to a set of “driving processes” that are meant to provide a simplified description of the system dynamics in terms of some of the network characteristics, such as the user’s behavior and the cell’s dimension. In particular, it is often assumed that the call request time sequence is described by means of a Poisson process, and that the time for which a call requires service within a cell (the call dwell time) is exponentially distributed. The former assumption does not allow the description of the increase in call request rate due to repeated attempts (also called repeated calls, reattempts, or retrials) by users whose call request was refused due to the lack of available resources. Repeated calls can have quite a negative impact on the system performance and should therefore not be neglected in network design and planning.

A large number of papers dealing with customer retrials have been published in the last decades (see [8]–[10]). Most dealt with conventional wired networks. A comprehensive survey on the topic can be found in [8] and [9]. More recently, a few papers also studied the problem of customers retrials in the context of wireless networks [4], [7]. A wide body of analytical results has been derived for single-server retrial queues [7]–[9], while no general analytical solution has been obtained for multiserver retrial queues [8], [9]; in this case, results are available only for some special models (in [10], results for the two-server case are reported). Numerical approaches thus appear to be the only way to obtain results for the multiserver case (see [9] and references therein, and [4]). They rely on the numerical solution of the steady-state Kolmogorov equations of the continuous time Markov chain (CTMC) describing the system dynamics. The number of states of the CTMC model is, however, infinite if the user population is infinite and can be extremely large in the case of finite user population, so that the numerical computation of the model solution can become very expensive in terms of memory space and CPU time, or even impossible, in many cases of practical interest in the wireless context.

In this paper, an approximate modeling approach is proposed that tries to mitigate the impact of the detailed description of the call arrival process (including new requests as well as retrials) on the solution complexity. The proposed approach is validated against results obtained from a detailed discrete-event simulator of the cellular system with retrials. The comparison shows that, in spite of the approximate model simplicity, the approach yields very accurate predictions of the system performance. Thus, the approximate model presented in this paper can be successfully used for an accurate performance evaluation, design, and planning of cellular mobile telephone networks under assumptions that make the numerical solution of the exact CTMC prohibitive.

II. SYSTEM AND MODELING ASSUMPTIONS

A cellular mobile telephone network is considered, where each cell is served by a different base station. We focus on a particular cell that can handle up to \( N \) simultaneous active communications. Communications taking place in different cells are assumed not to interfere with each other.

In order to avoid handover failures, it is a common practice in cellular mobile telephone networks to favor handovers by re-
serving some channels for them. This means that if \( N_H \) channels are reserved for handovers in the cell we consider, new call requests are accepted when more than \( N_H \) channels are free, whereas handovers are accepted as long as some free channel exists.

In order to describe the repeated calls phenomenon, we must identify users whose communication request has been rejected or dropped by the system for lack of resources. We say that such users are blocked, separating users who are blocked because a new call has been rejected, or because a handover procedure failed. In both cases, blocked users frequently reattempt to establish a call, possibly with different rates.

The call arrival process into each cell is thus governed by four different dynamics, corresponding to:

1. new calls;
2. incoming handover requests;
3. retrials originated by blocked new calls; and
4. retrials originated by blocked handover requests.

The process of call departures from a cell is instead driven by three types of events:

1. completions of calls;
2. handover requests out of the cell; and
3. termination of the retrial sequence due to the customer impatience resulting from repeated blocking.

In the development of models of the cell dynamics we introduce the following assumptions:

- The aggregate process of new call requests from non-blocked users within a cell is Markovian; when the number of nonblocked users is large with respect to the number of channels \( N \), the sequence of new call request arrivals is described by a Poisson process with parameter \( \lambda \) (infinite population model); otherwise, the interarrival time of call requests by each individual, nonblocked user is exponentially distributed with rate \( \mathcal{X} \) (finite population model).
- The flow of incoming handover requests from other cells is modeled as a Poisson process whose rate is \( \lambda_h \) (\( \lambda_h \) is derived by balancing the incoming and outgoing handover flows, as explained below).
- The time between two successive retrials performed by each user blocked after a new call (handover) attempt is exponentially distributed with rate \( \lambda_{en} \) (\( \lambda_{en} \) is derived from the model parameters). An iterative procedure is used to balance the incoming and outgoing handover rates, assuming that the incoming handover rate at step \( j \) is equal to the outgoing handover rate at step \( (j-1) \). The iterative procedure is stopped when the relative precision is smaller than \( 10^{-6} \).

III. QUEUING NETWORK MODELS

The assumptions presented above allow the construction of a queuing network model of the system, as shown in Fig. 1. Both cases of infinite and finite population are represented in the figure. In the infinite population case the new call arrival process is represented by the dotted line, and queue \( d \) is not present, as well as all dashed lines. In the finite population case, queue \( d \) and the dashed lines represent the new call arrival process; the dotted line must be disregarded. The three queues named \( a \), \( b \), and \( c \) represent, respectively, the sets of active calls, of blocked new calls, and of blocked handovers. The fourth queue, named \( d \), that appears only in the finite population model, represents the set of inactive users in the cell.

Queue \( a \) has \( N \) servers but no waiting line. The arrival process into this queue is the superposition of four components: the arrival process of new call requests (in the infinite population model it is Poisson with rate \( \lambda \) while in the finite population model it is the output process from queue \( d \)), the arrival process of incoming handover requests which is Poisson with rate \( \lambda_h \), and the arrival processes due to retrials by users with blocked
new calls and handovers, i.e., the output processes of queues $b$ and $c$. When a new call arriving at queue $a$ finds no server available (less than $N_H + 1$ channels are free in the cell) it moves to queue $b$. When a handover request arriving at queue $a$ finds no server available (no channel is free in the cell), it moves to queue $c$. The service time at queue $a$ represents the active time of a call within the cell. Since a call leaves the cell because of either call completion or outgoing handover, in the case of exponential dwell times the service time is exponentially distributed with rate $\mu + \mu_h$.

Queues $b$ and $c$ have an infinite number of servers. Arrivals are due to requests generated for queue $a$ but rejected because of lack of available resources. The service time at queue $b$ models the time a blocked user waits before retrying a new call; hence, this time is exponentially distributed with mean $1/\lambda_{bn}$. Similarly, the service time at queue $c$ models the time a blocked user waits before retrying a handover; hence, this time is exponentially distributed with mean $1/\lambda_{cn}$. A blocked user which reattempts to enter queue $a$ and is rejected again decides to leave the system with probability $P_{in}$ for a new call and $P_{ih}$ for a handover request.

For the finite population case, the service time at queue $d$ models the time between the end of a call and a new request by a user within the considered cell; hence this time is exponentially distributed with mean $1/\lambda'$. The presence of blocking in the queuing network makes the model non-product-form, so that for the computation of the steady-state probability distribution it is necessary to resort to the continuous-time Markov chain (CTMC) generated by the queuing network model of the system. However, even this approach cannot be pursued because of the extremely large state space of the CTMC that does not allow the investigation of systems of significant size. Thus, the development of approximate CTMC models with manageable state spaces becomes mandatory for the investigation of cells with a large number of channels and/or users.

IV. APPROXIMATE MARKOVIAN MODELS

For the sake of clarity, we first present the approximate CTMC model that can handle the case $\lambda_{bn} = \lambda_{bh} = \lambda_n$, $P_{in} = P_{ih} = 0$, $N_H = 0$, exponentially distributed dwell times and infinite population that corresponds to the queuing network model depicted in Fig. 2. Then we introduce the generalizations necessary to cope with the general case.

A. First Approximate Model

In the CTMC associated with the queuing network in Fig. 1, the system state is represented by vector $(N_a, N_b, N_c)$, where $0 \leq N_a \leq N$ is the number of active calls in the cell (queue $a$), and $0 \leq N_b, N_c < \infty$ are the numbers of blocked new calls and handovers, respectively (queues $b$ and $c$).

The simplified configuration we are considering is such that blocked new calls and blocked handovers are indistinguishable; hence, we can aggregate all states with the same value of $N_B = N_b + N_c$. As a result, we obtain a CTMC with state $(N_a, N_B)$, where $N_a$ is defined as before, and $N_B$ is the total number of blocked users, either on new calls or handovers. This is equivalent to reducing the system to the simplified queuing network reported in Fig. 2, where queue $B$ collects all blocked calls, and where the rate at which call requests arrive at queue $a$ when the system is in state $(N_a, N_B)$, is $\lambda + \lambda_h + N_B\lambda_c$.

A model of this type was considered in the classical works on retrial systems (see [8] and [9] and references therein) focused on the performance evaluation of wired telephone networks, and more recently in [4] in the context of wireless systems, taking into account either an infinite or a finite user population. Naturally, the number of states of the CTMC model in [4] is infinite if the user population is infinite, and equal to $(N + 1)(M - N + 1)$ if the size of the user population is $M$ and the number of channels is $N$, so that the computation of the model solution can become very time consuming in the case of large user populations or large number of channels.

In order to obtain a finite CTMC model with small state space, it is possible to limit the number of simultaneously blocked users to a predefined maximum value, say $\mathcal{M}$ (thus truncating the state space, and disregarding states in which the number of blocked users would be larger than $\mathcal{M}$). However, this approach is not satisfactory because the model solution is rather sensitive to the maximum number of blocked users, especially when a small value is selected for $\mathcal{M}$ (as needed to obtain a small state space). The sensitivity of the results to the maximum number of blocked users is also visible when large values are selected for $\mathcal{M}$, especially in high traffic conditions, as we shall see later, with the further disadvantage of a marginal reduction in the state space size.

Contrary to reducing the state space by truncation, we propose an approximate CTMC model whose state is vector $(N_a, b)$ where $0 \leq N_a \leq N$ is the number of active calls in the cell, and $b \in \{0, 1\}$ is a Boolean variable indicating the presence of blocked calls (1 if there are blocked calls, 0 otherwise). The state space comprises only $2(N + 1)$ states. Intuitively, with this approximation we are hiding the detailed information carried by queue $B$ about the number of blocked users, keeping just a global indication about the existence of blocked users.

In particular, the lack of the information about the number of blocked users forces us to use approximations to describe two aspects: 1) the sequence of requests generated by blocked
users; 2) the instants in which queue B becomes empty, and correspondingly b is reset to 0.

The sequence of call requests generated by blocked users can be approximated with an interrupted Poisson process with rate $\lambda_B \lambda_b$, where $\lambda_B = E[N_B]$ represents the average number of blocked users in the system, provided queue B is not empty (to be estimated as described below).

A heuristic approach is necessary to compute the steady-state probability that, whenever a blocked user retrial is successful, other users with blocked calls exist in the cell (so that queue B does not become empty). We postulate that such probability is constant, equal to $P$, and evaluate such probability with an iterative procedure, as explained further on.

The state transition rate diagram of the proposed approximate CTMC model is shown in Fig. 3, where $\mu_t = \mu_0 + \mu_h$ is the total rate of call termination in the cell, due to either call completion or outgoing handover; $\lambda_t = \lambda + \lambda_b$ is the total rate of call arrival, due to new calls and incoming handovers. The arrival rate due to repeated calls is $\lambda_B \lambda_b$. In Fig. 3, we denote with $\alpha$ the rate of attempts which leave the blocked queue empty, and with $\beta$ the rate of those which leave the queue nonempty: $\alpha = (1 - P) \lambda_B \lambda_b$, and $\beta = P \lambda_B \lambda_b$. The transition rates are as follows:

- $(N_\alpha, 0) \Rightarrow (N_\alpha + 1, 0)$, $0 < N_\alpha < N$: these transitions happen whenever either a new call or a handover requests a channel and finds it free—the rate is $\lambda_t$.
- $(N_\alpha, 1) \Rightarrow (N_\alpha + 1, 1)$, $0 \leq N_\alpha < N$: these transitions happen whenever a new call, a handover request or a repeated call enters the system and some blocked users remain within the system—the rate is $\lambda_t + \frac{\beta}{\alpha} = \lambda_t + P \lambda_b \lambda_B$.
- $(N_\alpha, 1) \Rightarrow (N_\alpha + 1, 0)$, $0 \leq N_\alpha < N$: these transitions correspond to the acceptance of a repeated call leaving no blocked user in the system—the rate is $\alpha = (1 - P) \lambda_B \lambda_b$.
- $(N_\alpha, 1) \Rightarrow (N_\alpha, 0)$: this transition models the arrival of a request that cannot be satisfied, when no blocked user is in the system, hence the blocked flag is set to 1—the rate is $\lambda_b$.
- $(N_\alpha, x) \Rightarrow (N_\alpha - 1, x)$, $0 < N_\alpha \leq N$: $x = [0, 1]$: these transitions correspond to a user leaving the system for either call completion or outgoing handover — the rate is $N_\alpha \mu_t$.

In order to derive the values of $P$ and $\lambda_B$, let us go back to the queuing network of Fig. 2 and focus on the evolution of the number of blocked calls, i.e., of the number of customers in queue B. Customers enter the queue when new call requests or handovers from other cells are refused; customers leave the queue in correspondence of successful retrials performed by blocked users. For this $G/M/\infty$ queue, $P$ represents the steady-state probability that a departing customer leaves the queue nonempty. A well-known property of queuing systems with unitary increments and decrements states that this probability equals the steady-state probability that a customer entering the queue finds it nonempty. The steady-state probability that a customer entering the queue of blocked users finds it nonempty can be expressed as the ratio between the average flow of entering customers that find the queue nonempty, and the whole average flow of customers entering the queue. We use this consideration for our CTMC (see Fig. 3) and compute $P$ as the ratio:

$$P = \frac{\lambda B \Pi(N, 1)}{\lambda B \Pi(N, 0) + \Pi(N, 1)} = \frac{\Pi(N, 1)}{\Pi(N, 0) + \Pi(N, 1)}$$

where $\Pi(N, 0)$ and $\Pi(N, 1)$ are the steady-state probabilities that the system is in states $(N, 0)$ and $(N, 1)$, respectively.

$\lambda_B$ can be evaluated by assuming a geometric distribution of the number of customers in the queue of blocked users, and supposing that $1 - P$ represents the steady-state probability that the queue is empty. As a consequence, we get:

$$\lambda_B = \frac{P}{1 - P} = \frac{\Pi(N, 1)}{\Pi(N, 0)}$$

The reason for the introduction of a geometric assumption on the distribution of the number of customers in the queue lies in the lack of correlation between the number of blocked users and the state of the CTMC. The retrial queue in the model can be characterized as a single server queue, since we do not represent the modulation in the service rate due to variations of the number of queued customers but only consider its average. In addition, it is easy to recognize that, due to the lack of correlation between the blocking probability experienced by blocked calls in subsequent attempts, the interdeparture times from the retrial queue during busy periods are exponentially distributed. Thus, modeling the population distribution in the retrial queue similarly to the number of customers in an $M/M/1$ queue seems consistent with the other model approximations.

Similarly to $\lambda_b$, $P$ and $\lambda_B$ cannot be a priori derived from the model parameters; hence, an iterative procedure is necessary to compute their values. In the initial step of the iteration, the CTMC model is solved assuming $P(0) = 0$ and $E[N_B](0) = 1$; at each following step $j$, the values are updated on the basis of the results obtained for the state distribution at step $j - 1$. The procedure is iterated until the relative precision is smaller than $10^{-4}$. The convergence of $\lambda_B$ is a consequence of the convergence of $P$.

**B. Different Retrial Rates**

We now generalize the approximate CTMC model to a slightly more complex case, where the user population is still taken to be infinite, $P_m = P_{th} = 0$, $N_B = 0$ but $\lambda_n \neq \lambda_h$.

We develop two approximate CTMC models for this case. In the first one, the queues of blocked new calls and blocked handovers (queues b and c in Fig. 1) are separately described, while in the second model they are collectively considered. At the expense of twice the state space cardinality, the first model provides more accurate results.
As previously mentioned, the state of the exact model of the system is given by the vector \( \langle N_a, N_b, N_h \rangle \), where \( N_i \) is the number of users in queue \( i \) of the network reported in Fig. 1.

We first approximately describe the two queues of blocked users by means of two Boolean variables, which indicate the presence of some users in the queues. The state vector for the first CTMC model thus is \( \langle N_a, b, b \rangle \), where \( N_a \) is the number of active calls in the cell; \( b \) and \( b \), respectively, indicate the presence of blocked new calls (queue b) and the presence of blocked handovers (queue c). The behavior of each queue of blocked users is separately approximated by using two parameters: the average number of customers in the queue and the steady-state probability that a departing customer leaves the queue nonempty. We denote these parameters with \( \alpha_n \) and \( \beta_n \) for queue b, whose customers represent blocked new calls, and with \( \alpha_b \) and \( \beta_b \) for queue c, whose customers represent blocked handovers.

The resulting CTMC comprises \( 4(N+1) \) states; its state transition rate diagram is presented in Fig. 4. From state \( \langle N, 0, 0 \rangle \), in which the system is full (all channels are busy) and no user is blocked, the arrival of a new call with rate \( \lambda \) causes the transition to state \( \langle N, 1, 0 \rangle \) (\( b \) is set to 1 because the queue of blocked new calls has become nonempty). The arrival of a handover request, instead, makes the state move with rate \( \mu_h \) to \( \langle N, 0, 1 \rangle \) (\( b \) is set to 1). Similar transitions connect state \( \langle N, 1, 0 \rangle \) and \( \langle N, 0, 1 \rangle \) to state \( \langle N, 1, 1 \rangle \) where both flags are equal to 1. When a blocked user retrial is successful, and the corresponding queue becomes empty, the flag associated with the queue is reset. In the CTMC, these events correspond to transitions with rate \( \alpha_n = (1 - P_n)\lambda N_n \) for new calls, and \( \alpha_b = (1 - P_b)\lambda N_h \) for handovers. Successful retrials, which leave the queue nonempty, have rate \( \beta_n = P_n\lambda N_n \) for new calls, and \( \beta_b = P_b\lambda N_h \) for handovers.

\( P_n \) is iteratively derived from the CTMC steady-state solution as the ratio between the flow of entering customers which find \( b \) nonempty, and the total flow of customers entering the queue; the average number of customers is obtained assuming a geometric distribution.

\[
P_n = \frac{\Pi(N,1,0) + \Pi(N,1,1)}{\Pi(N,0,0) + \Pi(N,1,0) + \Pi(N,1,1) + \Pi(N,0,1)}
\]

Similarly, for blocked handovers we have

\[
N_h = \frac{P_h}{1 - P_h}.
\]

The second approximate model for this same configuration uses just one flag for the description of both queues of blocked users. Thus, the CTMC model state is \( \langle N_a, b \rangle \), where \( b \) is the Boolean variable which indicates that there are some blocked users, with no distinction between handovers and new calls. Note that, as in the case with \( \alpha_b \), we are using just one queue for all blocked users. However, while in that case we could exactly aggregate states due to the identical behavior of the two types of blocked users, in this case we approximate by averaging their behavior.

The parameters we need to estimate are now: \( P \), the steady-state probability that a customer departing from the blocked users queue leaves it nonempty; \( N_n \) and \( N_h \), the average numbers of blocked new calls and blocked handovers in the queue.

The queuing system we are referring to is similar to the one shown in Fig. 2. The CTMC state transition rate diagram is again as in Fig. 3, where now the values of the rates \( \alpha \) and \( \beta \) are \( \alpha = \Pi(N,1,0) + \Pi(N,1,1) \) and \( \beta = \Pi(N,0,1) + \Pi(N,1,1) \). Each class of blocked users contributes to the retrial rate with a term which is the product of the retrial rate per customer and the average number of customers. \( P \) is obtained with the same definition used before.
Some care is needed in the derivation of $\bar{N}_B$ and $\bar{N}_h$. Let us focus first on $\bar{N}_h$, the average number of blocked handovers. This quantity can be expressed as $\bar{N}_h = E[\lambda(B)] \cdot E[T_h]$, where $E[\lambda(B)]$ is the arrival rate of handovers at the blocked queue $B$, and $E[T_h]$ is the average time spent in the queue by a blocked handover $E[\lambda(B)] = \lambda_h (\Pi(N, 0) + \Pi(N, 1))$. $E[T_h]$ is given by the product of the average time between two successive attempts to establish a call, which is $1/\lambda_{th}$, and the average number of attempts $\bar{A} = 1/P_s$, where

$$P_s = \frac{\sum_{i=0}^{N-1} \Pi(i, 1)}{\sum_{i=0}^{N} \Pi(i, 1)}$$

is the probability that an attempt is successful. We finally get $E[T_h] = (1/\lambda_{th})\bar{A} = 1/(\lambda_{th}P_s)$ and

$$\bar{N}_h = \lambda_h (\Pi(N, 0) + \Pi(N, 1)) \frac{1}{\lambda_{th}} \frac{\sum_{i=0}^{N-1} \Pi(i, 1)}{\sum_{i=0}^{N} \Pi(i, 1)}.$$ 

Similarly, we can derive the average number of blocked new calls $\bar{N}_n$,

$$\bar{N}_n = \lambda (\Pi(N, 0) + \Pi(N, 1)) \frac{1}{\lambda_{bn}} \frac{\sum_{i=0}^{N-1} \Pi(i, 1)}{\sum_{i=0}^{N} \Pi(i, 1)}.$$ 

This model is less accurate than the previous one with two Boolean flags but simpler to solve, since its state space only comprises $2(N+1)$ states.

C. Impatient Users

We show next how the model with two flags can be enhanced to take into account the impatience of users. After any unsuccessful retrial, a blocked user may decide to definitively give up, and leave the system. In the CTMC state transition rate diagram shown in Fig. 5, the transitions which take into account the impatience of users are shown with thick lines.

Let us focus on the behavior of a blocked handover when the system is in state $(N, 0, 1)$ or $(N, 1, 1)$. Since no channel is available, a blocked handover which retries to access the system experiences a failure, and with probability $P_{th}$ decides to leave the system. The transitions from state $(N, 0, 1)$ to $(N, 0, 0)$ and from state $(N, 1, 1)$ to $(N, 1, 0)$ have rate $P_{th} \alpha = P_{th}(1-P_h)\lambda_{bh}N_h$, where $\lambda_{bh}N_h$ is the retrial rate and $(1-P_h)$ is the probability of leaving the queue empty.

Similar considerations apply to the transitions from state $(N, 1, 0)$ to $(N, 0, 0)$, and from state $(N, 1, 1)$ to $(N, 0, 1)$, which are due to the impatience of blocked new calls. Their rate is $P_{tn} \alpha = P_{tn}(1-P_n)\lambda_{bn}N_n$.

D. Finite Population

The modification of the CTMC models to account for the finiteness of the user population (say of size $M$) is quite simple, since it only impacts the new call arrival rate, which instead of being constant becomes a function of the system state as summarized below:

<table>
<thead>
<tr>
<th>State</th>
<th>Arrival rate</th>
<th>State</th>
<th>Arrival rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(i, 0, b_n)$</td>
<td>$(M-i)x'$</td>
<td>$(i, 1, b_n)$</td>
<td>$(M-i-N_n)x'$</td>
</tr>
</tbody>
</table>

Note that also for finite user population the solution of the exact CMTC does not appear feasible since the state space dimension is very large.
E. Channels Reserved to Handovers

Next, we briefly discuss how the CTMC models presented above can be modified to take into account that $N_H$ channels are reserved to handovers. We only discuss the introduction of this feature into the CTMC model with two flags—infinite population and patient users. Similar considerations apply also to the other CTMC models.

In Fig. 6, the transitions which account for the differences between this new model and the one for $N_H = 0$ (CTMC state transition rate diagram in Fig. 4) are shown with a thick line.

When a new call is blocked because less than $N_H$ channels are available, it enters the blocked queue. Transitions from states $(N_a, 0, b_H)$, with $N_a \geq N - N_H$, to states $(N_a + 1, b_H)$ have rate $\lambda$.

In the states with $N_a \geq N - N_H$, only handovers are accepted, so that the following transitions are possible: $(N_a, b_H, 0) \implies (N_a + 1, b_H, 0)$ with rate $\lambda_H$, $(N_a, b_H, 1) \implies (N_a + 1, b_H, 1)$ with rate $\lambda_H + \beta_H$; where $\beta_H$, as before, is equal to $P_h\lambda_h\bar{N}_H$.

F. General Dwell Time Distributions

The exact representation of arbitrary dwell time distributions with Markovian models is not possible; however, it is well known that phase-type (PH) distributions, obtained as combinations of negative exponentials, allow the match of any desired number of moments (provided they exist and are finite) of an arbitrary distribution. The drawback in the use of PH distributions is the explosion of the state space of the model. The extremely low complexity of the models that we presented before permits the use of PH distributions, albeit of low order. The extension of the Markovian model to accommodate PH distributions of dwell times is cumbersome, but straightforward; we skip a detailed description of the resulting model for the sake of brevity, since no further approximations are involved. We implemented the case of a hyperexponential distribution of order 2 ($H_2$) for dwell times, thus allowing the modeling of any (finite) coefficient of variation $C_v$ greater than one. Results obtained with this model will be presented in Section V-E, and compared to simulation results where several different dwell time distributions are used.

1) Approximation for High Values of $C_v$:

Although PH distributions allow arbitrarily high coefficients of variation, it is interesting to explore the possibility of developing a model that permits a low-complexity analysis of systems where dwell times exhibit large values of $C_v$.

Assume that the high coefficient of variation of dwell times is mainly due (as it is often the case) to two different classes of customers: almost static (slow) and rapidly moving (fast). Under this hypothesis, it may be reasonable to further assume that the dynamics of fast and slow customers are independent: the number of fast customers in the system when a new slow customer requires service is independent from its previous value. With this independence assumption, we can avoid keeping track of the number of fast users in the system state. In other words, we can assume that the state of the system can be described by the vector $(N_s, b_H)$, where $N_s$ is the number of slow users, and the two flags $b_H$ and $b_l$ indicate the presence of blocked users, as usual. This is a significant simplification with respect to the state vector $(N_f, N_s, b_H, b_l)$ used for the $H_2$ distribution (where $N_f$ is the number of fast users), and reduces the complexity of the model to the same level of exponentially distributed dwell times. The application of this approach requires that the dwell time of slow users is at least ten times larger than that of fast users.

We describe here the resulting model in the case when no channels are reserved to handovers and blocked users retry indefinitely to connect to the network. The extension to other cases is straightforward.

The state space of the CTMC model is identical to the one of the CTMC depicted in Fig. 4; however, transition rates are now...
different, since an arriving slow customer may be blocked also when \( N_s < N \), due to the presence of fast customers.

The probability that a slow customer experiences blocking when entering the system, and finds \( N_s \) slow customers already in service, equals the probability that \( N - N_s \) fast customers are being served at that instant. Under the assumptions introduced above, this probability is given by the Erlang-B formula, and is thus independent from the fast customers dwell time distribution. This consideration greatly increases the applicability of this simplified model.

The transition rate diagram in Fig. 7 represents this CTMC model. In comparison to the model in Fig. 4, transition rates are completely changed (see the description below); furthermore, transitions toward states with blocked users are now possible from any state, not only from states with \( N_s = N \) customers in service (additional transitions with respect to Fig. 4 are highlighted with thick lines).

The arrival rate of fast customers is \( \lambda_g (1 - P_b) \), where \( \lambda_g \) represents the global arrival rate of fast customers, taking into account also the retrials of blocked customers, and \( P_b \) is the average blocking probability experienced by users.

The service rate \( \mu_s \) refers only to slow customers. All arrival rates depend on the number of slow customers in service: the superscript “prime” (\( \alpha' \), \( \beta' \), and \( \gamma' \)) in the figure indicates this dependency, so that \( \lambda = \lambda(N_s) \), \( \alpha' = \alpha'(N_s) \), and \( \beta' = \beta'(N_s) \). Let \( E(N_s) \) denote the probability that \( N - N_s \) fast customers are in service, computed through the Erlang-B formula. Then the arrival rates in Fig. 7 are as follows:

\[
\begin{align*}
\lambda_{ns}(N_s) &= \lambda_g (1 - E(N_s)) \\
\lambda_{ns}(N_s) &= \lambda_{ns} E(N_s) \\
\lambda_{hs}(N_s) &= \lambda_{hs} E(N_s) \\
\alpha_{ns}'(N_s) &= (1 - P_h) E(N_s) \lambda_{bn} N_n \\
\alpha_{ns}'(N_s) &= (1 - P_h) E(N_s) \lambda_{bh} N_h \\
\beta_{ns}'(N_s) &= P_h E(N_s) \lambda_{bn} N_n \\
\beta_{ns}'(N_s) &= P_h E(N_s) \lambda_{bh} N_h.
\end{align*}
\]

The computation of \( P_n \) and \( P_h \) (the probabilities of blocked new calls and blocked handovers in the system, respectively) is slightly more complex in this case, since slow customers can experience blocking in any state. For \( P_n \), we have

\[
P_n = \frac{\sum_{i=0}^{N} \epsilon(i) [\Pi(i, 1, 0) + \Pi(i, 1, 1)]}{\sum_{i=0}^{N} \epsilon(i) [\Pi(i, 0, 0) + \Pi(i, 1, 0) + \Pi(i, 1, 1) + \Pi(i, 0, 1)]}
\]

and \( P_h \) is computed similarly, swapping the roles of the two flags.

G. Performance Indices

Many interesting performance indices can be computed from the steady-state solution of the approximate Markovian models. We show here how some of these indices can be derived for the approximate model presented above in Section IV-E. Similar expressions can be derived for the other models.

The most important performance measure for our systems is the probability that a call is blocked due to lack of available resources.

Exploiting the well-known “Poisson arrivals see time averages” (PASTA) theorem, we can derive the new call blocking probability, \( P[B_n] \), as well as the handover failure probability \( P[B_h] \):

\[
P[B_n] = \sum_{N_s=0}^{N} \sum_{b_n=0}^{N_b} \sum_{b_h=0}^{b_{bh}} \Pi(N_s, b_n, b_{bh})
\]

\[
P[B_h] = \sum_{b_n=0}^{N_b} \sum_{b_h=0}^{b_{bh}} \Pi(N, b_n, b_h).
\]
The probabilities that a retrial generated by a blocked new call or by a blocked handover fails are

\[
P[B_{rv}] = \frac{\sum_{N_a=N-a}^{N} \sum_{b_v=0,1} \Pi(N_a, 1, b_v)}{\sum_{N_a=0}^{N} \sum_{b_v=0,1} \Pi(N_a, 1, b_v)}
\]

\[
P[B_{rh}] = \frac{\sum_{b_h=0,1} \Pi(N_b, b_h, 1)}{\sum_{b_h=0,1} \Pi(N_b, b_h, 1)}
\]

Other interesting performance measures can be easily defined in similar ways.

### H. Complexity

Since the proposed models are based on simple CTMCs but require that some parameters (incoming handover flow and average number of blocked calls) are iteratively derived, the computational cost of the model solution consists in the repeated solution of a CTMC. The number of iterations depends on the accuracy required in the estimation of those parameters. For the results shown later in this paper, for example, the accuracy level was set to $10^{-4}$, and the number of iterations varied between a few units and 100.

The computational cost of the solution of a CTMC depends on the chosen solution technique and on the number of states of the chain. Employing typical direct methods (see [12]), the solution complexity is roughly $O(K^3)$, where $K$ is the number of states of the model.

Since the state space cardinality of the proposed models is very small $2(N + 1)$ for the one-flag model and $4(N + 1)$ for the two-flag model, where $N$ is the number of channels in the cell, for exponentially distributed dwell times; $(N + 1)(N + 2)$ and $2(N + 1)(N + 2)$, respectively, for hyperexponential dwell times], the solution complexity is always extremely low [recall that the number of states of the model in [4] with exponentially distributed dwell times is $(N + 1)(M - N + 1)$, where $M$ is the maximum number of blocked users].

### V. Results

In this section, we discuss the accuracy of the approximate CTMC models that were presented in the previous section, by comparing their results against results of exact CTMC models, simulation point estimates, and bounds. While the analytical and the simulation models adopt exactly the same assumptions as regards user mobility and call duration, they differ in the assumptions about the behavior of blocked users and the interactions between cells. In the simulation setup, a network comprising 19 cells is completely modeled, and, within each cell, blocked users and handovers are accounted for in detail, i.e., each blocked user tries to connect to the network with retrials separated by time intervals that are exponentially distributed with parameters $\mu_{bn}$ (for blocked new calls) or $\mu_{bh}$ (for blocked handovers), and handovers are performed with rate $\mu_h$ toward a neighboring cell chosen at random. These two aspects are crucial, since the main approximations introduced in the analytical model consist in studying just one cell in isolation, and in simplifying the description of the behavior of blocked users.

For all the reported simulation results, the relative width of confidence intervals is always below 10%, with confidence level 0.95.

The values of the system parameters used in the derivation of the performance curves presented in this section are given in Table I.

#### A. Identical Retrial Rates

Figs. 8–10 show curves of the blocking probability, which in this case, due to the fact that no channel is reserved to handovers, coincides with the probability that a new call or a handover request finds no free channel, versus the system traffic, for different values of the system parameters. All plots comprise five curves referring to different models, as well as simulation results.

- The solid line curve refers to a single-cell model in which the phenomenon of repeated calls is ignored. This amounts

| Table I: Parameter Values Used in the Numerical Results Figures |
|------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Figure number    | 8         | 9         | 10        | 11        | 12        | 13        | 14        | 15        | 16        |
| Number of channels | N         | 32        | 32        | 32        | 32        | 32        | 32        | 32        | 32        |
| New call arrival rate | $\lambda$ | 1         | 1         | 1         | 1         | 1         | 1         | 1         | 1         |
| Dwell time rate   | $\mu_a$  | 0.1$\mu$ | 0.1$\mu$ | 0.1$\mu$ | 0.1$\mu$ | 0.1$\mu$ | 0.1$\mu$ | 0.1$\mu$ | 0.1$\mu$ |
| Retrial rate, new calls | $\lambda_{bn}$ | 0.1 | 1 | 0.1 | 1 | 0.1 | 1 | 0.1 | 1 |
| Retrial rate, handovers | $\lambda_{bh}$ | 0.1 | 1 | 0.1 | 1 | 0.1 | 1 | 0.1 | 1 |
| Number of channels reserved to handovers | $N_h$ | 0 | 0 | 0 | 1 | 2 | 1 | 2 | 1 | 0 | 0 |
| Impatience probability, new calls | $P_{bn}$ | 0 | 0 | 0 | 0 | 0 | 0.6 | 0 | 0 | 0 |
| Impatience probability, handovers | $P_{bh}$ | 0 | 0 | 0 | 0 | 0 | 0.6 | 0 | 0 | 0 |
| Number of users | $M$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |

---

The probabilities that a retrial generated by a blocked new call or by a blocked handover fails are...
Fig. 8. Call blocking probability for identical retrial rates of blocked new calls and handovers ($\lambda_{bn} = \lambda_{bh} = \lambda_b$); analytical results, simulation estimates and bounds; $\mu_n/\mu = 0.1$.

Fig. 9. Call blocking probability for identical retrial rates of blocked new calls and handovers ($\lambda_{bn} = \lambda_{bh} = \lambda_b$); analytical results, simulation estimates and bounds; $\mu_n/\mu = 1$.

to setting $\lambda_{bn} = \lambda_{bh} = 0$, thus obtaining a lower bound to the blocking probability.

- The long-dashed line curve refers to a model in which $\lambda_{bn} = \lambda_{bh} \rightarrow \infty$, thus obtaining an upper bound to the blocking probability. It should be noted that for $\lambda_{bn} = \lambda_{bh} \rightarrow \infty$ the system behaves as a queuing system in which blocked requests queue to be satisfied as soon as a channel becomes available; in this case, an exact model for the single cell system (in which the correlation among adjacent cells is neglected) is provided by an $M/M/N$ queue, so that results can be obtained with the Erlang $C$ formula.

- The short-dashed line curve is obtained with the proposed approximate CTMC model.

- The two dot-dashed lines curves (marked TG–M) are obtained by adapting the model proposed by Tran-Gia and Mandjes in [4] to the infinite user population case, for coherence with all other models, and by limiting the maximum number of blocked users to 10 and 50, respectively.

- Square markers without line refer to simulation points.

In all figures, the external aggregate rate of call generation by nonblocked users is set to 1 ($\lambda = 1$), while the average duration of a call varies, so that different values of the traffic (defined as $\rho = \lambda/\mu$) are obtained.

Figs. 8–10 differ for the value of the average dwell time $\mu_n^{-1}$, while the upper and lower plots of the figures refer to different values of the retrial rate, $\lambda_{bn} = \lambda_{bh} = \lambda_b$, respectively, equal to 0.1 and 1. Note that, in single-cell models, changes in the value
of $\mu_h$ imply changes in the system load, even if $\rho$ remains the same.

In Fig. 8, $\mu_h/\mu$ is set to 0.1, so that, on the average, only 10% of all calls generated in the considered cell perform handovers. In Fig. 9, $\mu_h/\mu$ is set to 1; this value corresponds, on the average, to one handover request per call. In Fig. 10, instead, $\mu_h/\mu$ is set to 10. This value corresponds to an average number of handovers per call equal to 10.

First of all, remember that, by definition, both the single cell model without repeated calls and the $M/M/N$ model are insensitive to the values of $\lambda_{bh}$ and $\lambda_h$. Moreover, the $M/M/N$ model is also insensitive to the handover rate $\mu_h$. Instead, the single-cell model without retrials shows a decreasing blocking probability as the handover rate increases; this is due to losses that reduce the system load.

As expected, the curve obtained with the proposed approximate model lies between the two curves which represent upper and lower bounds for the system performance. The simulation results are practically coincident with the proposed approximate model results, showing a remarkable agreement in spite of the introduced approximations. It is quite interesting to notice that the single-cell model without retrials yields very optimistic results, especially in a high mobility scenario. In this case, if the network load is high, this model underestimates the blocking probability by more than one order of magnitude, independently of the retrial rate.

Results obtained with the model proposed by Tran-Gia and Mandjes in [4] show that a limited size of the blocked users
queue produces an underestimation of the blocking probability for high loads. The error can be appreciated even for rather large values (i.e., 50) of the maximum number of blocked users allowed in the model.

The upper bound offered by the $M/M/N$ model is tight only if the handover rate is low and the retrial rate is high (see the lower plot of Fig. 8), while in other cases the blocking probability is overestimated by a factor between two and four.

**B. Different Retrial Rates and Channels Reserved to Handovers**

In Figs. 11 and 12, we present curves of the blocking probability for new calls and curves of the handover failure probability, versus traffic, for different values of the system parameters, when the handover retrial rate is significantly higher than the retrial rate of users blocked because of a refused new call request. The upper and lower plots of the two figures refer to different numbers of channels reserved to handovers. In all plots, ten handovers per call are performed on average.

All plots comprise six curves (three for the new call blocking probability and three for the handover failure probability) referring to the two different approximate CTMC models presented for this case (one-flag model and two-flag model), as well as simulation results (the approach proposed by Tran-Gia and Mandjes in [4] just considers identical retrial rates, hence cannot be used for comparison in this case).
Simulation results practically coincide with the results of the proposed approximate model with two flags, again showing a remarkable accuracy in spite of the introduced approximations.

C. Impatient Users

In Fig. 13, we present curves of the blocking probability for new calls and curves of the handover failure probability, versus traffic, for different values of the system parameters, when blocked users may decide to terminate the retrial sequence after any repeated failure. The left and right plots of the two figures refer to different values for $\rho$. All plots comprise four curves (two for the new call blocking probability and two for the handover failure probability) referring to the two-flag model or simulation (the approach proposed by Tran-Gia and Mandjes in [4] does not consider impatience, hence cannot be used for comparison).

Also in this case, simulation results practically overlap with the results of the approximate two-flag model, once more proving its accuracy and robustness.

D. Finite Population

In Fig. 14, we present curves of the blocking probability for new calls and curves of the handover failure probability, versus traffic, for different values of the system parameters, when the number of users in the cell is finite. The upper and lower plots of the two figures refer to different values for $\rho$. The two plots comprise two curves for the new call blocking probability and two for the handover failure probability, reporting results obtained with either the two-flag model or simulation. Once more, simulation results practically coincide with the results of the approximate two-flag model.

Finally, Fig. 15 shows the effect of the population size on the blocking probability. In order to compare the results of our approximation to the model originally proposed in [4], we need to set $P_{in} = P_{ih} = 0$ and $\lambda_{in} = \lambda_{ih}$ (indeed, the model in [4] does not allow impatience probabilities and different retrial rates for new calls and handovers). The figure shows curves of the call blocking probability obtained with the two-flag approximate CTMC model and the model proposed in [4], for population size equal to 100, 200, 1000, and for infinite population. In order to limit the size of the state space generated by the model proposed in [4], the maximum number of elements in the retrial queue is limited to 20.

Numerical results clearly show that the results of the approximate two-flag model practically coincide with those of the model originally proposed in [4]. When the latter produces significantly lower estimates for the blocking probability this is due to the finite size of the queue of blocked users. In addition, as expected, we can see that smaller user population sizes lead to better system performance. It can be noted that for a given blocking probability, the amount of traffic that can be accepted varies significantly with the population size.
E. Generally Distributed Dwell Times

Blocking probability results for generally distributed dwell times are shown in Fig. 16, considering the case of identical retrial rates for new calls and handovers, an average of 10 handovers per call, and no impatience.

The curves in Fig. 16 compare the results obtained with two approximate analytical models to simulation results. The approximate analytical models are 1) the model described in Section IV-F1, and 2) the model resulting from a standard description of dwell times with a hyperexponential distribution of order 2, hence using two exponential phases in parallel. The results obtained with the analytical models are reported for values of the coefficient of variation of the dwell time—$C_v = 2, 10$—but the four curves almost overlap. Simulation results are reported for the exponential distribution, for hyperexponential distributions with $C_v = 2$ and 10, and for a Weibull distribution with $C_v = 5$. All curves provide almost identical results, as expected, thanks to the well-known property of several pure loss systems, whose performance only depends on the average service time, not on the distribution type.

F. Solution Complexity

It is interesting to observe that the CPU time required for the solution of the proposed approximate models is more than three orders of magnitude less than the time needed to obtain reliable simulation results. For example, each point of the curves in Fig. 14 is obtained in less than 1 s using the approximate analytical model, while more than 8000 s are necessary with the simulator.

The solution of the proposed approximate models is also much faster than the solution of the model in [4]: the average time required for the computation of one point of the curves in Fig. 15 referring to a population comprising 100 users is less than 1 s with the proposed approximate model, and about 1000 s with the model in [4]. This is due to the smaller state space of the former model (66 states instead of 693), and to a faster convergence: while the approximate model requires between 5 and 7 iterations to reach accuracy $10^{-4}$, the same is reached with about 60 iterations using the model in [4].

VI. CONCLUSION

In this paper, we have proposed a novel approach for the development of simple approximate CTMC models for the evaluation of call blocking probabilities in mobile cellular networks with customer retrials due to blocked new calls or failed handovers. The novel approach is based on the development of CTMCs whose state spaces grow only linearly with the number of calls that can be simultaneously in progress within a cell.

The results obtained with the approximate models were compared to exact analytical results, upper and lower bounds to the call blocking probability, and results of detailed simulation experiments. Approximate numerical results showed a remarkably good agreement with simulation estimates, and pointed out the weakness of traditional approaches that either do not take into account repeated calls, or give them priority over new calls.

The very limited size of the state space in the proposed models allows the extension of the modeling approach to system setups more complex than those analyzed here, for example, accounting for different service classes, different user mobility patterns, or nonstationary traffic loads resulting from variations of the number of users within the system.

REFERENCES


Marco Ajmone Marsan holds degrees in electronic engineering from Politecnico di Torino and University of California, Los Angeles (UCLA).

He is a Full Professor with the Electronics Department, Politecnico di Torino, Italy. From November 1975 to October 1987, he was with the Electronics Department, Politecnico di Torino, first as a Researcher, then as an Associate Professor. From November 1987 to October 1990, he was a Full Professor with the Computer Science Department, University of Milan, Italy. During the summers of 1980 and 1981, he was with the Research in Distributed Processing Group, Computer Science Department, UCLA. During the summer of 1998, he was an Erskine Fellow with the Computer Science Department, University of Canterbury, New Zealand. He has coauthored over 250 journal and conference papers in the areas of communications and computer science, as well as the two books Performance Models of Multiprocessor Systems published by the MIT Press, and Modeling with Generalized Stochastic Petri Nets published by John Wiley. His current interests are in the fields of performance evaluation of communication networks and their protocols.

Dr. Marsan received the best paper award at the Third International Conference on Distributed Computing Systems in Miami, FL, in 1982.

G. De Carolis, photograph and biography not available at the time of publication.

Emilio Leonardi (S’94–M’99) received the Dr. Ing degree in electronics engineering in 1991 and the Ph.D. degree in telecommunications engineering in 1995, both from Politecnico di Torino, Italy.

He is currently an Assistant Professor with the Electronics Department, Politecnico di Torino. In 1995, he spent one year with the Computer Science Department, University of California, Los Angeles (UCLA), where he was involved in the Supercomputer–SuperNet (SSN) project aimed at the design of an high capacity optical network architecture. During the summer of 1999, he joined the High-Speed Networks Research Group, Lucent Bell Labs, where he worked on the design of scheduling algorithms for high capacity switch architectures. His research interests are in the fields of all-optical networks, switching architectures, and queueing theory, wireless communications.

Renato Lo Cigno (A’92) received the Dr. Ing. degree in electronic engineering from Politecnico di Torino, Italy in 1988.

He is currently an Assistant Professor with the Electronics Department, Politecnico di Torino. Since 1988, he has been with the telecommunications research group, the Electronics Department, Politecnico di Torino, first under various research grants and contracts, then as a Research Engineer. From June 1998 to February 1999, he was with the Computer Science Department, University of California, Los Angeles (UCLA) as Visiting Scholar under Grant CNR 203.15.8. He is coauthor of about 50 journal and conference papers in the area of communication networks and systems. His current research interests are in performance evaluation of wired and wireless networks, modeling and simulation techniques, and flow and congestion control.

Michela Meo (S’94–A’95) received the Dr. Ing. degree in electronic engineering in 1993 and the Ph.D. degree in electronic and telecommunication engineering in 1997, both from Politecnico di Torino, Italy. During her Ph.D. studies, she was with the Computer Science Department, University of California, Santa Barbara (UCSB) for one year.

She is currently an Assistant Professor with the Electronics Department, Politecnico di Torino. Since 1997, she has been with the Telecommunication Research Group, Politecnico di Torino. Her research interests are in the field of performance evaluation of communication networks with a particular focus on wireless systems.