Approximate Analytical Model for Queued Handover Requests in Cellular Networks

Vicente Casares-Giner, Jorge Martinez-Bauset
Instituto Universitario de Tecnologas de la Informacin y Comunicaciones
Universitat Politcnica de Valncia
Email: vcasares@itaca.upv.es, jmartinez@itaca.upv.es

***Abstract*—An approximate analytical model to analyse the performance of the handover process in cellular networks is proposed, where new and handover calls that arrive when insufficient free resources are available are queued instead of being lost. The approximation is based on the aggregation of states of the double infinite continuous-time Markov chain that models the system, and exhibits an excellent accuracy and low computational cost. The approximate model might be of interest to the next-generation of 5G mobile networks that must be engineered to achieve high QoS and extremely low latencies.**

***Index Terms*—Guard Channel Algorithm (GCA); handover; priority; forced termination; Quasi Birth Death (QBD) process.**

I. INTRODUCTION

Handover algorithms are of paramount importance in wireless cellular networks, and suitable analytical models are needed to evaluate their performance. From the user equipment (UE) point of view, it is less desirable the interruption of a call in progress than the blocking of a new one. The most commonly deployed strategy to achieve this Quality of Service (QoS) objective has been to assign higher priority to calls in progress than to the newly arriving ones.

One of the most celebrated prioritization schemes is the Guard Channel Algorithm (GCA) [1]. Let C be the total number of Resource Units (RU) available at an LTE eNodeB (eNB). The meaning of a unit of resource depends on the specific implementation of the radio interface. Let $C_h \leq C$ the number of guard RUs. Then, new and handover calls are admitted when the number of free RUs is larger than C_h. However, when the number of free RUs is equal to C_h, or less, only handover calls are admitted, while new calls are blocked. Clearly, when all C RU are occupied, both new and handover requested calls are blocked. Note that there is no prioritization when $C_h = 0$. It is worth pointing out that the GCA has been proposed in other wireless networks, such as trunking systems in which interconnect calls have priority over dispatch calls [2].

The problem of prioritization of handover calls over new one has been commonly treated in the context of admission control in cellular networks. Table I summarizes the four main schemes that have received attention in the literature. In the *loss-loss* scheme, both new and handover requests that arrive when not enough free RU are available, will terminate being lost.

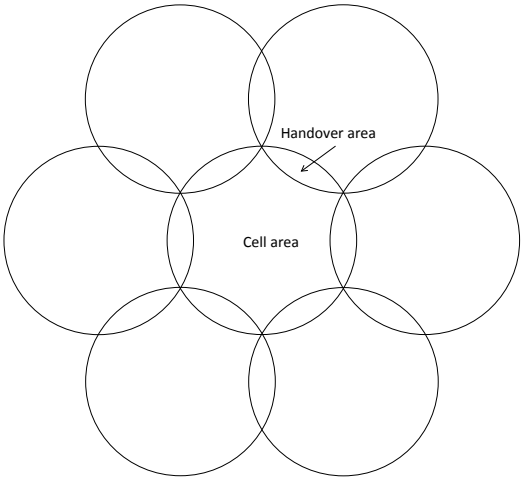

Fig. 1. Cell area and handover area.

In the *loss-delay* scheme, new calls might be blocked, but instead of blocking handover requests, they are placed in a waiting queue of capacity Q_h until enough RU become available. The time handover requests are maintained in the queue is mainly a function of two parameters: i) the residence time of the UE in the handover area, i.e., the overlapping area between the current serving cell and the new one; ii) the speed of the UE. Please refer to Fig. 1. This scheme has been analyzed in [3] with the GCA and the first-in-first-out (FIFO) service queue discipline for the queued handover requests. A later study in [4] extended the work in [3] by considering that a call might terminate while waiting in queue.

In the third scheme, the *delay-loss* one, handover requests might be blocked, but instead of blocking new requests, they are placed in a waiting queue of capacity Q_f, until enough free RU become available. In this case, the time new requests are

TABLE I
TREATMENT OF CALLS OR SESSIONS.

Scheme	New, (f)	handovers, (h)	Queue size, Q_*
1	Loss	Loss	$(Q_f = 0, Q_h = 0)$
2	Loss	Delay	$(Q_f = 0, Q_h > 0)$
3	Delay	Loss	$(Q_f > 0, Q_h = 0)$
4	Delay	Delay	$(Q_f > 0, Q_h > 0)$

 ISBN: 978-1-61208-736-8

maintained in the queue is mainly a function on the residual sojourn time in the non-overlapping area of the cell. Such scheme has been analysed in [1], and revisited in [5], where the system was modeled as a Quasi Birth and Death (QBD) Markov process and basic principles of the M/G/1 system were used to analyse the queuing model [6].

Finally, in the *delay-delay* scheme, both, new and handover requests, are queued when insufficient RU are available upon arrival. An exhaustive analysis of this fourth model is provided in [4] when both Q_f and Q_h are finite.

The interest of the *delay-loss* scheme is based on the fact that, in general, UEs spend a short time in the handover area. This might be due to small overlapping areas, UEs speed or both. In any case, the time spent by a UE in the handover area might be negligible when compared to the time spent in the non-overlapping area of the cell. However, the next-generation of 5G cloud-enabled services are being engineered to achieve high QoS and latencies as small as 1 ms. In such network operation scenarios, it is gaining a renewed interest the performance evaluation of handover schemes that might place handover requests in a queue for short periods of time.

An approximate analytical model to evaluate the *delay-delay* scheme is proposed, where the time evolution of the number of new and handover calls in the system is modeled by a double infinite continuous-time Markov chain (CTMC), that has the form of a QBD process. The QBD process turns out to be non-homogeneous in both dimensions, and therefore its solution is computationally expensive. The approximate analytical model is based on the aggregation of states of the CTMC, and exhibits an excellent accuracy and low computational cost. The original QBD process is in this way converted into an equivalent finite QBD, for which different efficient solution algorithms have been proposed.

Although in this paper a brief description of the approximation method and a preliminary study of the accuracy achieved is given, our interest is to extend the study and apply it to the analysis of the handover procedure in 5G cellular networks. In particular, to analyse the impact that different network features, such as size of the cell overlapping areas, UEs speed, density of small cells, etc, might have on the QoS perceived by the UEs, when both new and handover requests might get queued at eNB.

The paper is structured as follows. Section II defines the Markovian hypothesis for the queue models. Section III deals with a detailed qualitative description of the scheme analyzed in our work. The development of the analytical approach to determine the stationary distribution of the CTMC and the main performance parameters are presented in Sections IV and V, respectively. A cellular scenario is presented in Section VI, and the corresponding results are reported in Section VII. Conclusions and future work in progress are reported in Section VIII.

II. MAKOVIAN HYPOTHESIS

For model tractability, it is assumed that new and handover requests arrive following a Poisson process with rates λ_f and λ_h. respectively. This modeling approach has been widely debated and accepted in the literature [7], [8], [9]. In the same way, it is assumed that call or session duration, cell residence time, and residence time in the handover area, are exponentially distributed random variables with rates μ_M, μ_R and μ_F, respectively. Due to the memory-less property, the RU holding times are again exponentially distributed, with rates $\mu_H = \mu_M + \mu_R$ when the UE resides in the cell area, and $\mu_Q = \mu_M + \mu_F$ when it resides in the handover area. Obviously $\mu_R < \mu_F$ so $\mu_H < \mu_Q$. Note that in most of previous works it has been assumed that $\mu_Q \approx \mu_F$. Then, the call was not able to terminate while residing in the handover area. This limitation was overcome in the model proposed in [4], where, for the first time and to our best knowledge, the authors consider that the call can finish in the handover area, i.e., $\mu_Q = \mu_M + \mu_F$. Finally, note that according to [10] and [11], the mean residence time in the handover area, $1/\mu_F$, can be around 5-10 seconds, which is much shorter than the cell residence time $1/\mu_R$, or to the call duration $1/\mu_M$, that can be around 2 minutes on average.

III. MODEL OF THE DELAY-DELAY SCHEME

In this Section, we describe the model proposed for the scheme 4 of Table I. As described before, let C the total number of RUs of the eNB and C_h the number of guard RUs. Sessions or calls occupy a singe RU in the eNB while being served. Two types of requests are offered to the eNB, new (fresh) and handover calls requests. A new call originated in a given cell is admitted if more than C_h free RUs are found upon arrival. Otherwise, the fresh call is placed in a queue of infinite capacity, and remains in that queue while residing in the cell area. No impatience is assumed for the calls in the queue. A handover request is admitted if at least one free RU is found upon arrival. Otherwise, it joins a queue of infinite capacity and remains in that queue until the UE abandons the handover area, or until the call ends, whichever occurs first.

The system is modelled as a 2-D Markov process of infinite size in both dimensions, as shown in Fig. 2. The system state is defined by the tuple (i, j), $i, j = 0, 1, 2, \ldots \infty$, where $\min(i, C)$ define the number of calls in progress in the cell, $\max(i - C, 0)$ the number of handover requests in the queue, and j the number of new (fresh) calls in the queue.

Our model is an extension of the one studied in [1] and [5] in two aspects. First, queued fresh calls are allowed to leave the queue when they leave the cell area. In this case, the set up request will be rejected, and it will not be transferred to any neighboring cell. Although the subscriber might retry the call after a random period, this behavior has not been considered in the current model. Second, different to the treatment in [5], we allow that a handover request that joins the queue remains in it until the call in progress ends, or until the UE leaves the handover area, whichever occurs first.

Parallel to [5], when visiting state (i, j) we say that the process is at phase i and at level j. Two key observations. First, for any phase $i > C_s = C - C_h$, we realize that transition between phases are independent of the arrival rate of new calls

 ISBN: 978-1-61208-736-8

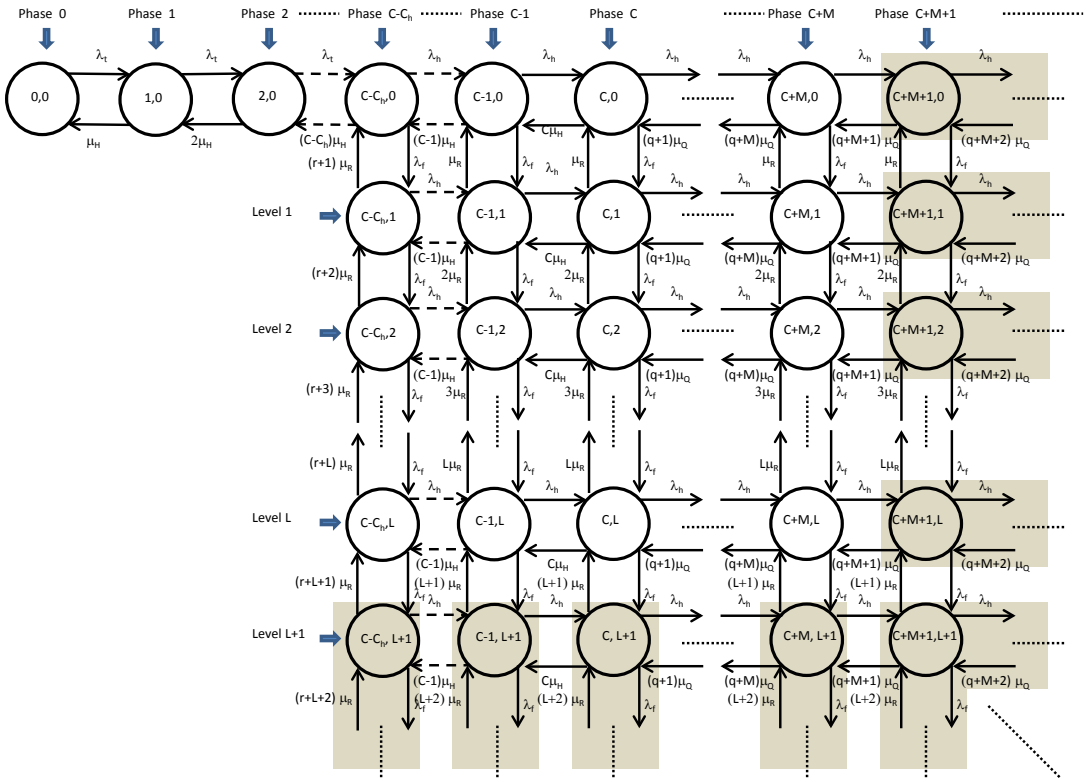

Fig. 2. State transition diagram of the 2D CTMC for the delay-delay model.

λ_f. Second, for any level $j > 0$ we realize that transitions between levels are independent of the arrival rate of handover requests λ_h. Also, note that for $i \geq C$, the number of handover requests in the handover queue increases by one unit when a transition from phase i to phase $i+1$ occurs. In the same way, for $j \geq 0$, a transition from level j to level $j+1$ makes the number of new requests in its queue to increase by one unit. Finally, note also that a transition form level j to level $j-1$ that happens when the system is in phase $C - C_h$, leads to the reduction of the number of new calls in its queue by one unit.

In the next Section, we describe the approximate model. As mentioned before, is based on the aggregation of sets of states of the 2D Markov process [12].

IV. State aggregation approach.

First, we focus on the set of states at level $j \geq 0$, and which phases meet $i \geq C$, i.e., states in which the handover queue is not empty. This infinite set of states recall us a similar set of states in a $M/M/\infty$ queue. This infinite set of states recall us a similar set of states in a $M/M/\infty$ queue. Then, we evaluate the mean value of the first passage time from state $C+M+1$ to state $C+M$ and approximate this infinite set of states by a single (aggregated) state for which the service (exiting) rate equals the inverse of the mentioned mean value.

Second, we focus on the set of states at a fixed phase $i \geq C_s$, and which levels meet $j > 0$, i.e., states in which the queue of new calls is not empty. The aggregation procedure turns out to be similar to the one described above, with the exception that two sets of states must be taken into account. The first one is the set with phase $i = C_s = C - C_h$, while the second is with phase $i > C_s$. Clearly, aggregating a set of states into a single one is an approximation, where only the first moment of the first passage time is being taken into account.

A. First passage time from phase $C+M+1$ to phase $C+M$

An upper and lower bound for the mean value of the first passage time from phase $C+M+1$ to phase $C+M$, $M > 0$, is derived. This mean value is denoted as $\bar{t}_{ph}(\lambda_h, \mu_Q, q, M)$ where $q = C\mu_H/\mu_Q$. Please refer to Table II for details. The procedure is as follows. From Fig. 2 the set of states with phase $i \geq C$ and level $j = 0$, define a birth-death process, as

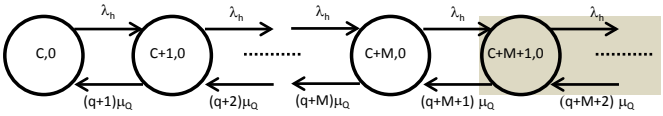

Fig. 3. Part of the CTMC of Fig. 2 for phases $i \geq C$ and level $j = 0$.

TABLE II
MAIN PARAMETERS

Capacity (RUs)	Definition
$C = C_s + C_h$	# of RUs in the cell.
C_s	# of RUs shared (fresh or new and handovers).
C_h	# of RUs reserved for handovers.
Rate (e.r.v.)	**Definition**
λ_f	Rate of offered fresh (new) calls
λ_h	Rate of offered handover calls
$\lambda_t = \lambda_f + \lambda_h$	Total arrival rate
γ_f	Admitted fresh (new) calls
γ_h	Admitted handover calls
$\gamma_t = \gamma_f + \gamma_h$	Total rate of admitted calls.
μ_M	Call (message) departure rate
μ_R	UE residence time in the cell area
$\mu_H = \mu_M + \mu_R$	RU occupancy in the cell
μ_F	UE residence time in the handover area
$\mu_Q = \mu_M + \mu_F$	RU occupancy in the handover area
$q = C\mu_H/\mu_Q$	$q_f = \lfloor q \rfloor, \quad q_c = \lceil q \rceil$
$r = C_s\mu_H/\mu_R$	$r_f = \lfloor r \rfloor, \quad r_c = \lceil r \rceil$
Erlangs	**Traffic**
$A_{os} = \lambda_f/\mu_M$	Offered new traffic -session, call or message-.
$A_f = \lambda_f/\mu_H$	Offered traffic at cell level -RU-.
$A_h = \lambda_h/\mu_H$	Handover requested traffic at cell level -RU-
$A_t = A_f + A_h$	Total offered traffic at cell level -RU-.
$A_R = \lambda_f/\mu_R$	"Unattended" new traffic in the queue.
$A_Q = \lambda_h/\mu_Q$	Ongoing handover traffic in the queue.
$A_F = \lambda_h/\mu_F$	"Unattended" handover traffic in the queue.
Probability	**Definition**
$P_{sc} = \mu_M/\mu_H$	Call ends in the cell.
$P_{hd} = \mu_R/\mu_H$	Request for a handover.
$P_{sh} = \mu_M/\mu_Q$	Call ends in the handover area.
P_B	Blocking of new (fresh) calls.
P_{fh}	Failure of handover request.
P_{FT}	Forced termination.
P_{NC}	Unencumbered call.

shown in Fig. 3. Omitting the level sub-index in Fig. 3, the transition rates and the steady state probabilities are given by,

$$\begin{aligned} &\lambda_i = \lambda_h; \quad i \geq C \\ &\mu_i = (q + i - C)\mu_Q; \quad i \geq C + 1. \end{aligned} \tag{1}$$

$$P_{C+i} = \begin{cases} P_C = \left[1 + \sum_{k=1}^{\infty} \frac{A_Q^k}{\prod_{n=1}^{k}(q+n)}\right]^{-1}; & i = 0 \\ \frac{A_Q^i}{\prod_{n=1}^{i}(q+n)} P_C; & i \geq 1 \end{cases} \tag{2}$$

with $A_Q = \lambda_h/\mu_Q$.

Following Appendix A in [13], we can derive the mean value of the first passage time from phase $C + M$ to phase $C + M + 1$. Its mean value, denoted as $\overline{\tau}_{ph}(\lambda_h, \mu_Q, q, M)$, can be written as,

$$\overline{\tau}_{ph}(\lambda_h, \mu_Q, q, M)\lambda_h = \left[1 + \sum_{k=1}^{M} \frac{A_Q^k}{\prod_{n=1}^{k}(q+M+n)}\right]\left[\frac{A_Q^M}{\prod_{n=1}^{M}(q+M+n)}\right]^{-1} \tag{3}$$

Finally, it is straightforward to see that $\bar{t}_{ph}(\lambda_h, \mu_Q, q, M)$, the mean value of the first passage time from phase $C+M+1$ to phase $C + M$ can be written, after some simple algebra,

$$\begin{aligned} &\bar{t}_{ph}(\lambda_h, \mu_Q, q, M) = \\ &\overline{\tau}_{ph}(\lambda_h, \mu_Q, q, M)\frac{1 - \sum_{k=C}^{C+M} P_k}{\sum_{k=C}^{C+M} P_k} = \\ &\frac{1}{\lambda_h}\sum_{k=1}^{\infty} \frac{A_Q^k}{\prod_{n=1}^{k}(q+M+n)} \end{aligned} \tag{4}$$

where P_k are the steady state probabilities given in (2). In addition, it can be verified that,

$$\bar{t}_{ph}(\lambda_h, \mu_Q, q, M) = \frac{1 + \lambda_h \bar{t}_{ph}(\lambda_h, \mu_Q, q, M+1)}{(q+M+1)\mu_Q} \tag{5}$$

From now, unless ambiguity does not allow it, we will use a short notation, i.e., $\bar{t}_{ph} = \bar{t}_{ph}(\lambda_h, \mu_Q, q, M, T_h)$. Then, for a suitable threshold $T_h \geq 1$, the following lower (lw) and upper (up) bounds can be written as,

$$\bar{t}_{ph,lw} = \frac{1}{\lambda_h}\sum_{k=1}^{T_h} \frac{A_Q^k}{\prod_{n=1}^{k}(q+M+n)} \tag{6}$$

$$\bar{t}_{ph,up} = \bar{t}_{ph,lw} + \frac{1}{\lambda_h}\frac{A_Q^{T_h}}{\prod_{n=1}^{T_h}(q+M+n)}\frac{A_Q}{q+M+T_h-A_Q} \tag{7}$$

Note that $\bar{t}_{ph,lw}$ is obtained by truncating up to the first T_h elements the infinite sum in (4). This approximation defines a lower bound to (4). To obtain $\bar{t}_{ph,up}$ we set μ_k constant at $\mu_k = (C+M+T_h)\mu_Q$ for $k > C+M+T_h$. As $(q+k-C) > (q+M+T_h)$ for $k > C+M+T_h$, this approximations sets an upper bound to (4). The infinite terms $k > C+M+T_h$ in (4) define a geometric progression with ratio $r_{ph} = A_Q/(q+M+T_h)$ that can be added, provided that $r_{ph} < 1$.

B. First passage time from level $L+1$ to level L

As in previous sub-Section IV-A, for a given phase $i \geq C_s = C - C_h$, we analyse the aggregation of states located at level $L+1$, $L > 0$. Let $\bar{t}_{le}(\lambda_f, \mu_R, r, L; i)$ denote the mean value of the first passage time from level $L+1$ to level L, where $r = C_s\mu_H/\mu_R$. Please refer to Table II for details. By inspection of Fig. 2, two different sets of states are identified. First, when the phase is $i = C_s$, left side of Fig. 4. Second, when the phase is $i > C_s$, right side of Fig. 4. For a level $j > L$, the transition rates between levels are given by,

$$\begin{aligned} \lambda_j &= \lambda_f; \quad j \geq 0 \\ \mu_j &= (r\delta_{i,C_s} + j)\mu_R; \quad j \geq 1 \end{aligned} \tag{8}$$

where δ_{i,C_s} is the Kronecker delta. Parallel to (4) we can write, being $A_R = \lambda_f/\mu_R$,

$$\bar{t}_{le}(\lambda_f, \mu_R, r, L; i) = \frac{1}{\lambda_f} \sum_{k=1}^{\infty} \frac{A_R^k}{\prod_{n=1}^{k} (r\delta_{i,C_s} + L + n)} \tag{9}$$

A simple inspection to the state transition diagrams in Fig. 4 reveals that $\bar{t}_{le}(\lambda_f, \mu_R, r, L; i = C_s,) < \bar{t}_{le}(\lambda_f, \mu_R, r, L; i > C_s)$.

For clarity, we use the notation $\bar{t}_{le}(i) = \bar{t}_{le}(\lambda_f, \mu_R, r, L; i)$, unless otherwise specified. Using the same arguments as we did for (6)-(7), and given a suitable threshold $T_f \geq 0$, the lower and upper bounds for $\bar{t}_{le}(i)$ are given by,

$$\bar{t}_{le,lw}(i) = \frac{1}{\lambda_f} \sum_{k=1}^{T_f} \frac{A_R^k}{\prod_{n=1}^{k} (r\delta_{i,C_s} + L + n)} \tag{10}$$

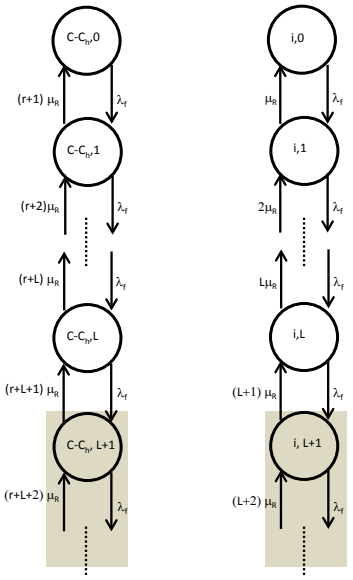

Fig. 4. Part of the CTMC of Fig. 2 for level $j \geq 0$ and phases $i = C_s = C - C_h$ (left); $i > C_s$ (right).

$$\begin{aligned} \bar{t}_{le,up}(i) &= \bar{t}_{le,lw}(i) + \\ &= \frac{1}{\lambda_f} \frac{A_R^{T_f}}{\prod_{n=1}^{T_f} (r\delta_{i,C_s} + L + n)} \frac{A_R}{r\delta_{i,C_s} + L + T_f - A_R} \end{aligned} \tag{11}$$

provided that $r_{le} = A_R/(\delta_{i,C_s} + L + T_f) < 1$. Note that the most restrictive case in the last inequality occurs when $i > C_s$, where $A_R < L + T_f$ must be fulfilled.

C. Steady state probabilities

Following the state aggregation process described in Section IV-A and Section IV-B, the original QBD process is converted into an equivalent finite QBD, for which different efficient solution algorithms have been proposed. The interested reader might refer to [14], or Chapter 10 in [15], for details of the algorithms used to solve the finite QBD process. Let $P_{i,j}$ and $\tilde{P}_{i,j}$ denote the exact and approximate steady state probabilities of the finite QBD, respectively. Clearly, $\tilde{P}_{i,j}$ can be evaluated for the four aggregation approaches shown in Table III.

TABLE III
FOUR APPROACHES FOR SCHEME 4 OF TABLE I

Approaches	Phase (level j independent)	Level (phase i dependent)
1: (6)-(10)	$\bar{t}_{ph,lw}(\lambda_h, \mu_Q, q, M, T_h)$	$\bar{t}_{le,lw}(\lambda_f, \mu_R, r, L, T_f; i)$
2: (6)-(11)	$\bar{t}_{ph,lw}(\lambda_h, \mu_Q, q, M, T_h)$	$\bar{t}_{le,up}(\lambda_f, \mu_R, r, L, T_f; i)$
3: (7)-(10)	$\bar{t}_{ph,up}(\lambda_h, \mu_Q, q, M, T_h)$	$\bar{t}_{le,lw}(\lambda_f, \mu_R, r, L, T_f; i)$
4: (7)-(11)	$\bar{t}_{ph,up}(\lambda_h, \mu_Q, q, M, T_h)$	$\bar{t}_{le,up}(\lambda_f, \mu_R, r, L, T_f; i)$

V. PERFORMANCE PARAMETERS

The performance parameter evaluated in this study are: i) the blocking probability of a new (fresh) call, ii) the handover failure probability, iii) the forced termination probability of an initiated call, and iv) the non-completed call probability. These parameter depend on $P_{i,j}$ through the analytical expressions defined in the next Subsections.

A. The blocking probability of new calls

Upon arrival, if a new (fresh) call finds $C_s = C - C_h$ RUs occupied, or more, it joins the queue. The call is finally blocked when the UE leaves the area of the serving cell. Then,

$$P_B = \frac{1}{\lambda_f} \sum_{i=C_s}^{\infty} \sum_{j=0}^{\infty} j\mu_R P_{i,j} = \frac{1}{A_R} \sum_{i=C_s}^{\infty} \sum_{j=0}^{\infty} j P_{i,j} \tag{12}$$

To evaluate (12) we use the probabilities $\tilde{P}_{i,j}$, obtained in Section IV-C, that is,

$$\begin{aligned} &P_{B,lw} \approx \\ &\sum_{i=C_s}^{C+M+1} \sum_{j=0}^{L} j \frac{\tilde{P}_{i,j}}{A_R} + \sum_{i=C_s}^{C+M+1} \zeta_{i,C_s} \frac{\mu_{le}(i)}{\lambda_f} \tilde{P}_{i,L+1} \end{aligned} \tag{13}$$

where, $\mu_{le}(i) = 1/\bar{t}_{le}(i)$ from (9), $\zeta_{i,C_s} = 1 - \delta_{i,C_s} r/(r + L + 1)$ and $A_R = \lambda_f/\mu_R$.

When the phase is $i = C_s$, $\zeta_{C_s,C_s} = (L+1)/(r+L+1)$ can be interpreted as the fraction of transitions $(C_s, L+1) \rightarrow (C_s, L)$ that represent the blocking of new (fresh) calls. Note that the transition $(C_s, L+1) \rightarrow (C_s, L)$ might also occur when a call terminates successfully and a queued new call occupies the freed RU. When the phase is $i > C_s$, then $\zeta_{i,C_s} = 1$, which means that each transition $(i, L+1) \rightarrow (i, L)$ reflects the lost of one new call.

$P_{B,lw}$ is a low bond as it only considers the blocking of new calls due to UEs abandoning the cell service area. An upper bound can be defined when, in addition, we consider the blocking of new calls that arrive at states $(i, L+1)$, $i \geq C_s$.

$$P_{B,up} \approx P_{B,lw} + \sum_{i=C_s}^{C+M+1} \tilde{P}_{i,L+1}. \tag{14}$$

We remark that, when evaluating $P_{B,lw}$, from (13), we use $\mu_{le,lw}(i) = 1/\bar{t}_{le,up}(i) \leq 1/\bar{t}_{le}(i) = \mu_{le}(i)$ given in (10), and when evaluating $P_{B,up}$, from (14), we use $\mu_{le,up}(i) = 1/\bar{t}_{le,lw}(i) \geq 1/\bar{t}_{le}(i) = \mu_{le}(i)$, given at (11).

B. Probability of a handover attempt failure

The probability of a handover attempt failure, P_{fh} , can be expressed as the quotient between the rate of failure handovers and the rate of handover attempts,

$$P_{fh} = \frac{1}{\lambda_h} \sum_{i=C}^{\infty} \sum_{j=0}^{\infty} (i - C) \mu_F P_{i,j} \tag{15}$$

As before, to evaluate (15) we use the probabilities $\tilde{P}_{i,j}$ of the finite QBD process,

$$P_{fh,lw} \approx \sum_{i=C}^{C+M} \sum_{j=0}^{L+1} (i - C) \frac{\tilde{P}_{i,j}}{A_F} + \sum_{j=0}^{L+1} \frac{\mu_{ph}}{\lambda_h} \tilde{P}_{C+M+1,j} \tag{16}$$

where $\mu_{ph} = 1/\bar{t}_{ph}$ from (4) and $A_F = \lambda_h/\mu_F$.

Using the same arguments as in (13), we consider (16) as the lower bound for P_{fh}. Also, in a parallel way to (14), the upper bound for P_{fh} is define as,

$$P_{fh,up} \approx P_{fh,lw} + \sum_{j=0}^{L+1} \tilde{P}_{C+M+1,j} \tag{17}$$

When evaluating $P_{fh,lw}$, expression (16), we use $\mu_{ph,lw} = 1/\bar{t}_{ph,up} \leq 1/\bar{t}_{ph} = \mu_{ph}$, from (6), and when evaluating $P_{fh,up}$, expression (17), we use $\mu_{ph,up} = 1/\bar{t}_{ph,lw} \geq 1/\bar{t}_{ph} = \mu_{ph}$, from (7).

C. Forced termination probability

Based on previous result for P_{fh}, the forced-termination probability P_{FT} can be evaluated as follows [3],

$$P_{FT} = P_{hd} \sum_{k=1}^{\infty} [(1 - P_{fh}) P_{hd}]^{k-1} P_{fh} = \\ = \frac{P_{hd} P_{fh}}{1 - (1 - P_{fh}) P_{hd}} \tag{18}$$

where P_{hd} is the probability of handover demand or attempt, and it is given by $P_{hd} = \mu_R/(\mu_R + \mu_M)$.

Using previous results of (16) (17), P_{FT} can be approximated as follows,

$$P_{FT,lw} = \frac{P_{hd} P_{fh;lw}}{1 - (1 - P_{fh,up}) P_{hd}} \tag{19}$$

$$P_{FT,up} = \frac{P_{hd} P_{fh;up}}{1 - (1 - P_{fh,lw}) P_{hd}} \tag{20}$$

D. Call non-completion probability

The new call blocking probability P_B, and the forced-termination probability P_{FT} can be combined to define the probability that a call does not terminate successfully, i.e., the non-completion probability,

$$P_{NC} = P_B + (1 - P_B) P_{FT} \tag{21}$$

As before, the lower and upper bounds for P_{NC} are defined as,

$$P_{NC,lw} = P_{B,lw} + (1 - P_{B,up}) P_{FT,lw} \tag{22}$$

$$P_{NC,up} = P_{B,up} + (1 - P_{B,lw}) P_{FT,up} \tag{23}$$

VI. Cellular scenario. Flow equations

The objective of this section is to determine a realistic value for λ_h in a conventional cellular scenario with multiple cells. As in [8], we assume regular tessellation of the 2D cellular area, cells of equal size, and uniform spatial distribution of UEs. We also assume that handover requests arrive following a Poisson process with rate λ_h. Basically, λ_h depends on the mobility of the UE, and must meet the following flow equations. Let $\gamma_{c,in}$ be the rate of calls in progress in a tagged cell. $\gamma_{c,in}$ has two terms, the rate of new calls that are admitted, i.e., $\lambda_{f,in}(1 - P_{B,in})$, and the rate of handover requests arriving from neighboring cells that are admitted, $\gamma_{c,out} P_{hd,out}(1 - P_{fh,in})$. Then,

$$\gamma_{c,in} = \\ \lambda_{f,in}(1 - P_{B,in}) + \gamma_{c,out} P_{hd,out}(1 - P_{fh,in}) \tag{24}$$

In equilibrium we have $\gamma_c = \gamma_{c,in} = \gamma_{c,out}$; $\lambda_f = \lambda_{f,in} = \lambda_{f,out}$; $P_{hd} = P_{hd,in} = P_{hd,out}$; $P_B = P_{B,in} = P_{B,out}$ and $P_{fh} = P_{fh,in} = P_{fh,out}$. Solving (24) for γ_c, hence for λ_h,

$$\lambda_h = \gamma_c P_{hd} = \frac{\lambda_f (1 - P_B)}{1 - P_{hd}(1 - P_{fh})} P_{hd} \tag{25}$$

 ISBN: 978-1-61208-736-8

TABLE IV
BOUNDS FOR $\bar{t}_{ph}(\lambda_h, \mu_Q, q, M, T_h)$, (6) AND (7); NORMALIZED TO $1/\lambda_h = 1/3.5$ WITH $A_Q = 0.70$, $C = 8$, $q = 3.2$,

	$\bar{t}_{ph}$	$T_h = 1$	2	3	4	5
M=	up	0.1555	0.1517	0.1514	0.1514	0.0432
1	lw	0.1346	0.1498	0.1512	0.1514	0.0432
	relative gap	0.1555	0.0129	0.0010	0.0000	0.0000
	r_{ph}	0.1346	0.1129	0.0972	0.0853	0.0432
	up	0.1272	0.1250	0.1249	0.1248	0.1248
2	lw	0.1129	0.1238	0.1248	0.1248	0.1248
	relative gap	0.1272	0.0095	0.0007	0.0000	0.0000
	r_{ph}	0.0853	0.0972	0.0853	0.0760	0.0686
	up	0.1076	0.1062	0.1062	0.1061	0.1061
3	lw	0.0972	0.1055	0.1061	0.1061	0.1061
	relative gap	0.1076	0.0073	0.0004	0.0000	0.0000
	r_{ph}	0.0972	0.0853	0.0760	0.0686	0.0625
	up	0.0933	0.0923	0.0923	0.0923	0.0923
4	lw	0.0853	0.0918	0.0923	0.0923	0.0923
	relative gap	0.0933	0.0058	0.0000	0.0000	0.0000
	r_{ph}	0.0853	0.0760	0.0686	0.0625	0.0573
	up	0.0823	0.0816	0.0816	0.0816	0.0816
5	lw	0.0760	0.0813	0.0816	0.0816	0.0816
	relative gap	0.0823	0.0047	0.0000	0.0000	0.0000
	r_{ph}	0.0760	0.0686	0.0625	0.0573	0.0530

TABLE V
BOUNDS FOR $\bar{t}_{le}(\lambda_f, \mu_R, r, L; i)$, FOR PHASE $i > C_s$, (10) AND (11); NORMALIZED TO $1/\lambda_f = 1/7$, WITH $A_R = 7.00$, $C = 8$, $C_h = 1$,$r =$

	$\bar{t}_{le}$	$T_f = 7$	8	9	10	11
L=	up	2.7023	2.6964	2.6942	2.6935	2.6932
7	lw	2.6547	2.6769	2.6867	2.6907	2.6922
	relative gap	0.0179	0.0058	0.0028	0.0010	0.0003
	r_{le}	0.5000	0.4666	0.4375	0.4117	0.3888
	up	2.0816	2.0791	2.0783	2.0780	2.0779
8	lw	2.0594	0.0705	2.0751	2.0768	2.0775
	relative gap	0.0107	0.0041	0.0015	0.0005	0.0001
	r_{le}	0.4666	0.4375	0.4177	0.3888	0.3684
	up	1.6732	1.6721	1.6717	1.6716	1.6715
9	lw	1.6621	1.6679	0.6702	1.6711	1.6714
	relative gap	0.0066	0.0024	0.0008	0.0002	0.0001
	r_{le}	0.4375	0.4117	0.3888	0.3684	0.3500
	up	1.3887	1.3881	1.3880	1.3879	1.3879
10	lw	1.3828	1.3861	1.3873	1.3877	1.3878
	relative gap	0.0042	0.0015	0.0008	0.0001	0.0000
	r_{le}	0.4117	0.3888	0.3684	0.3500	0.3333
	up	1.1814	1.1811	1.1810	1.1810	1.1810
11	lw	1.1781	1.1800	1.1807	1.1809	1.1810
	relative gap	0.0027	0.0009	0.0003	0.0000	0.0000
	r_{le}	0.3888	0.3684	0.3500	0.3333	0.3181

The rate λ_h in equation (25) (see also (17) in [16]), together with the steady state probabilities, $\tilde{P}_{i,j}$, define a fixed-point equation [17] [18]. To solve it, we set initially $\lambda_h \approx \lambda_f P_{hd}$ and the corresponding $\tilde{P}_{i,j}$ are obtained by solving the QBD process. A new λ_h is obtained from (25) and the iteration process is repeated until the difference between the probabilities $\tilde{P}_{i,j}$ of two consecutive iterations is less than a certain threshold.

The mobility rates μ_R and μ_F can been derived according to the fluid flow model, equations (12), (13) in [19]. Then,

$$\mu_x = \frac{E(v)L_x}{\pi A_x}; \quad x = R, F \tag{26}$$

where $E(v)$ is the expected velocity of the UE and L (A) the perimeter (the area) of the coverage area, R for the cell area and F for the handover area, see Fig. 1. From the geometry of that figure and denoting by R_c the radius of the circle, it can be shown that $\mu_R = 2(3\sqrt{3} - \pi)^{-1}E(v)/R_c \approx 0.9734E(v)/R_c$ and $\mu_F = 2(\pi - 3\sqrt{3}/2)^{-1}E(v)/R_c \approx 3.6797E(v)/R_c$. Then, $\mu_F/\mu_R \approx 3.7801$.

VII. RESULTS

A reference evaluation scenario is define with the following parameters: $C = 8$, $C_h = 1$, $\mu_M = 1$, $\mu_R = 1$ and $\mu_F = 4$.

Table IV shows the accuracy of the proposed lower and upper bounds for $\bar{t}_{ph}$ (phase). Results have been obtained for $\lambda_h = 3.5$, that makes $A_Q = \lambda_h/\mu_Q = 0.7$ Erlangs, and $q = C\mu_H/\mu_Q = 3.2$. The accuracy is measured in terms of the *relative gap* $= (up - lw)/lw$, i.e., the relative difference between both bounds. Note that taking the upper bound as a reference would not change the results. Observe that the *relative gap* decreases faster by increasing T_h than by increasing M.

Table V shows the accuracy of the proposed lower and upper bounds for $\bar{t}_{le}$ (level). Results have been obtained for $\lambda_f = 7$, that makes $A_R = \lambda_f/\mu_R = 7$ Erlangs, and $r = C_s\mu_H/\mu_R = 7$. As with the phase, note that *relative gap* decreases faster by increasing T_f than by increasing L.

Observe that the *relative gap* is below 10^{-4} when $r_{ph} = A_Q/(q + M + T_h) \leq 0.06$ in table IV, and when $r_{le} = A_R/(L + T_f) \leq 0.33$ in table V, approximately. A sensibility study of the impact that M, T_h, L and T_f have on the accuracy of the approximation is left for future work.

Figure 5 and Fig. 6 show the evolution of the main performance parameters studied with the load in a realistic scenario. The scenario is composed of multiple cells, and the cell under study is characterized by $C = 80$ RU. The approximate stationary distributions are obtained for $r_{ph} < 0.01$ and $r_{le} < 0.1$, that are more restrictive than those suggested above by inspection of Table IV and Table V. Note that in a multicell scenario, the fluid flow equation (25) most be solved iteratively using the fixed-point equation. As can be observed, the Guard Channel Algorithm is rather efficient, as the handover attempt failure and the forced termination probabilities decrease quite rapidly when the number of guard RUs changes from $C_h = 1$ to $C_h = 2$, while the non-completion probability keeps approximately invariant. However, the blocking probability of new (fresh) calls increases with C_h, as expected.

VIII. CONCLUSION AND FUTURE WORK

We study a cellular system where new and handover calls that arrive when insufficient free resources are available are queued instead of being lost. The time evolution of the number of new and handover calls in the system is modeled by a double infinite continuous-time Markov chain (CTMC), that has the form of a QBD process. As the solution of the QBD process is computationally expensive, we propose an

 ISBN: 978-1-61208-736-8

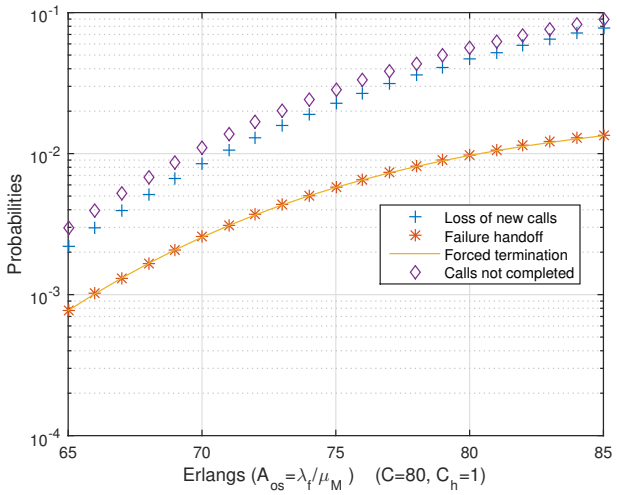

Fig. 5. Main parameters for $C = 80$, $C_h = 1$. P_B (13)-(14); P_{fh}, (16)-(17); P_{FT}, (19)-(20); P_{NC}, (22)-(23).

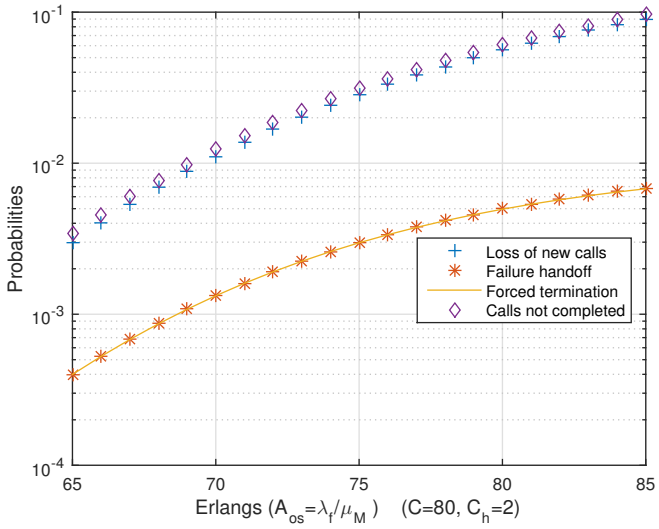

Fig. 6. Main parameters for $C = 80$, $C_h = 2$. P_B (13)-(14); P_{fh}, (16)-(17); P_{FT}, (19)-(20); P_{NC}, (22)-(23).

approximate analytical model based on the aggregation of states of the CTMC.

The approximation shows a very good accuracy and low computational cost. It is applicable to current 4G and forthcoming systems 5G systems. We plan to extend the study to analyze heterogeneous scenarios where femtocells and macrocells coexist, and where the corresponding CTMC that models the system behavior has a huge amount of states. We believe that the state aggregation technique is a powerful tool that makes the analysis of highly complex systems feasible.

ACKNOWLEDGMENT

This work is part of the project PGC2018-094151-B-I00. The authors would like to thank the financial support received from *Ministerio de Ciencia, Innovacin y Universidades (MCIU)*, from *Agencia Estatal de Investigacin (AEI)* and from *Fondo Europeo de Desarrollo Regional (FEDER)* (MCIU/AEI/FEDER.UE).

REFERENCES

[1] R. Guerin, "Queueing-blocking system with two arrival streams and guard channels," *IEEE Transactions on Communications*, vol. 36, no. 2, pp. 153–163, 1988.

[2] V. Casares-Giner, "Integration of dispatch and interconnect traffic in a land mobile trunking system. waiting time distributions," *Telecommunication Systems*, vol. 16, no. 3-4, pp. 539–554, 2001.

[3] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 77–92, 1986.

[4] 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 Transactions on Networking*, vol. 2, no. 2, pp. 166–175, 1994.

[5] J. N. Daigle and N. Jain, "A queueing system with two arrival streams and reserved servers with application to cellular telephone," in *IEEE INFOCOM'92*, 1992, pp. 2161–2167.

[6] M. F. Neuts, *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Johns Hopkins University, Baltimore, 1981.

[7] E. Chlebus and W. Ludwin, "Is handoff traffic really Poissonian?" in *Proceedings of ICUPC'95-4th IEEE International Conference on Universal Personal Communications*, 1995, pp. 348–353.

[8] P. V. Orlik and S. S. Rappaport, "On the handoff arrival process in cellular communications," *Wireless Networks*, vol. 7, no. 2, pp. 147–157, 2001.

[9] V. Casares-Giner, J. Martinez-Bauset, and X. Ge, "Performance model for two-tier mobile wireless networks with macrocells and small cells," *Wireless Networks*, vol. 24, no. 4, pp. 1327–1342, 2018.

[10] C. Jedrzycki and V. C. Leung, "Probability distribution of channel holding time in cellular telephony systems," in *Proceedings of Vehicular Technology Conference-VTC*, vol. 1. IEEE, 1996, pp. 247–251.

[11] F. Barceló and J. Jordán, "Channel holding time distribution in public telephony systems (PAMR and PCS)," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 5, pp. 1615–1625, 2000.

[12] V. Casares-Giner, L. Tello-Oquendo, V. Pla, and J. Martinez-Bauset, "A queueing system with two arrival streams and reserved servers with application to cellular telephone," in *XIII Jornadas de Ingeniera Telemtica, JITEL 2017, Ed. Universitat Politècnica de València*, 2017.

[13] V. C. Giner, "Variable bit rate voice using hysteresis thresholds," *Telecommunication Systems*, vol. 17, no. 1-2, pp. 31–62, 2001.

[14] D. Gaver, P. Jacobs, and G. Latouche, "Finite birth-and-death models in randomly changing environments," *Adv. App. Prob.*, vol. 16, no. 4, pp. 715–731, 1984.

[15] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*. ASA-SIAM, 1999.

[16] D. Hong and S. S. Rappaport, "Priority oriented channel access for cellular systems serving vehicular and portable radio telephones," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 136, no. 5, pp. 339–346, 1989.

[17] D. McMillan, "Traffic modelling and analysis for cellular mobile networks," in *Teletraffic and Datatraffic in a Period of Change*. North-Holland, 1991, pp. 627–632.

[18] F. Kelly, "Fixed point models of loss networks," *J. Austral. Math. Soc. Ser. B*, vol. 31, no. 2, pp. 204–218, 1989.

[19] V. Casares-Giner, V. Pla, and P. Escalle-García, "Mobility models for mobility management," in *Network performance engineering*. Springer-Verlag, 2011, pp. 716–745.

 ISBN: 978-1-61208-736-8