**Wireless Networks manuscript No.**
(will be inserted by the editor)

# Performance Model for Two-tier Mobile Wireless Networks with Macrocells and Small Cells

**Vicente Casares-Giner · Jorge Martinez-Bauset · Xiaohu Ge**

Received: date / Accepted: date

**Abstract** A new analytical model is proposed to evaluate the performance of two-tier cellular networks composed of macrocells (MCs) and small cells (SCs), where terminals roam across the service area. Calls being serviced by MCs may retain their channel when entering a SC service area, if no free SC channels are available. Also, newly offered SC calls can overflow to the MC. However, in both situations channels may be repacked to vacate MC channels. The cardinality of the state space of the continuous-time Markov chain (CTMC) that models the system dynamics makes the exact system analysis unfeasible. We propose an approximation based on constructing an equivalent CTMC for which a product-form solution exist that can be obtained with very low computational complexity. We determine performance parameters such as the call blocking probabilities for the MC and SCs, the probability of forced termination, and the carried traffic. We validate the analytical model by simulation. Numerical results show that the proposed analytical model achieves very good precision in scenarios with diverse mobility rates and MCs and SCs loads, as well as when MCs overlay a large number of SCs.

**Keywords** Two-tier mobile wireless networks · overflow and repacking · performance evaluation · blocking probability · forced termination probability.

## 1 Introduction

The traffic of cellular networks has been sustainably increasing over the last decade due to successful business applications that have attracted to over 6 billion subscribers. However, there is a growing consensus that current network features will

Corresponding author: Xiaohu Ge.

Vicente Casares-Giner, Jorge Martinez-Bauset
Universitat Politècnica de València, València, Spain
E-mail: vcasares@upv.es,jmartinez@upv.es

Xiaohu Ge
Huazhong University of Science & Technology, Wuhan, P.R. China
E-mail: xhge@mail.hust.edu.cn

not be able to cope with the required demands in the near future. Subscribers demand for multimedia wireless services is driving the search for a new generation of systems capable of offering the requested services in a cost-efficient manner.

Market surveys have revealed the that majority of mobile traffic is either originating or terminating indoors today, and this trend is growing [1]. In this context, the small cell (SC) concept has emerged as a solution to increase both network capacity and indoor coverage. It combines fixed-line broadband access with cellular networks by deploying low-cost, low-power base stations in the premises or homes of subscribers. SC deployment benefits both users and operators.

When SCs are deployed in addition to an existing macrocell (MC) coverage, we say that they mainly provide *capacity enhancement*. On the other hand, if they are deployed to bring a new service to an area where it was previously unavailable we talk about *coverage enhancement*. Other important considerations are: indoor or outdoor deployment, size of the coverage area, if they are deployed by consumers and enterprises or by operators, or if their access is open or closed [2]. It is envisaged that the vast majority of SCs will be self-deployed by consumers or enterprises indoors, in order to provide them with enhanced QoS cellular connectivity [1].

Several technical challenges have been identified for the deployment of SCs such as [3]: strategies to share the spectrum between MCs and SCs (shared, spliting, etc.), self-organization capabilities of SCs to share the spectrum, time synchronization of SCs with the rest of the network, QoS provided by the backhaul (particularly when shared with WiFi traffic), cross-tier interference in co-channel frequency allocations, open or closed access to SCs, handover schemes that minimize ping-pong effects and other performance degradation effects [4, 5].

As SCs usually share the licensed spectrum of the mobile operator, one of the most crucial factors is the cross-tier interference between MCs and SCs, that might substantially limit the system performance. Different approaches have been proposed in the literature to address this problem such as, for example, power control, or advanced spectrum management techniques [6, 7].

In this study we assume that elaborated interference management schemes are in place, such that the sets of channels allocated to MCs and SCs are perceived (in the long term) as being non-overlapping and with no co-channel interference. The non-overlapping channel perception can be achieved by above physical layer techniques, such as in [8]. In this respect, our setting is similar to the one studied in [9, 10].

In the deployment scenario we have in mind, SCs are disposed at homes or offices indoors with a closed access policy, in addition to operator deployed MCs (*capacity enhancement*). We focus on the radio resource allocation problem from the traffic perspective. Note that the traffic in cellular networks can be broadly categorized in two classes: streaming and elastic [11]. Streaming traffic requires a minimum transfer rate as well as bounds for the packet delay and jitter. It is generated by real-time applications that support voice or video. Elastic traffic is generated by file transfer applications, which transfer rate adapts to the available free resources.

In our study user terminals generate streaming traffic. It is commonly accepted that the tear-down of an ongoing call (forced termination) is more annoying for a subscriber than the blocking of a new setup request. Then, from the traffic perspective, the important QoS parameters are the blocking and forced termination probabilities. These probabilities are required to determine the fraction of calls

that terminate successfully. Typically, calls that terminate successfully are the those that generate revenue for the operator. Also they are a measure of the QoS perceived by subscribers. We aim at obtaining approximate, but accurate enough, expressions for the blocking probability of new calls initiated at MCs or SCs, and the probability of forced termination of accepted calls. Note also that one of the outcomes of our model is the traffic the system can carry (while QoS objectives are met), which is a measure of the system capacity.

### 1.1 Related Work

Our study is in part motivated by the fact that most of recent studies mainly concentrate on interference mitigation in rather static scenarios, where terminals are static, and in most cases connections are always active [12]. This scenario is quite different to ours, where terminals move and connections are established and released dynamically.

A number of works have studied the resource allocation problem from the traffic perspective. An important fraction of them study the blocking probability and signaling load trade off, typically in a two-tier network where terminals move at different speeds. To improve system performance, a call arriving at a SC with no free channels overflows to the MC, instead of being blocked. Exact solutions tend to be computationally intractable when the number of SC and MC channels take practical values [13]. Then, approximate solutions that focus on the characterization of the overflow traffic are the most common proposals. Please, refer to [10, 14], and references therein, for more details about these approximate solutions.

We share the view of the authors of [15] that models for two-tier network with terminals moving at different speeds are more appropriate for a scenario where operators deploy MCs overlaying public SCs deployed outdoors (hierarchical cellular networks). However, in our scenario of study SCs are deployed indoors, where users are static or have only pedestrian-level low mobility. This leads to a deployment scenario completely different to most of the ones studied so far. This scenario was explored in a previous work by the authors, but with no mobility [16].

The repacking of calls that overflowed from a SC to the MC has been shown to improve the system performance, i.e., to reduce the blocking and forced termination probabilities, as a repacked call releases valuable resources at the MC [17]. By repacking we refer to a process by which a call residing in a SC but using a MC channel is requested to handover to a free SC channel. Call repacking might, in addition, reduce system interference and save energy. However, the traffic analysis of two-tier systems with mobility, and call overflow and repacking, has not been sufficiently studied by analytical models.

An approximate analytical model was proposed in [18] to determine the blocking probabilities in a hierarchical cellular network with repacking and two classes of subscribers moving at different speeds, slow and fast. Although our study shares with it some goals, there are substantial differences. The stochastic processes defined and approximations used in [18] are completely different to the ones proposed here. Most importantly, the scale of the problems studied are different. While our aim is to study systems with hundreds of SCs per MC, the system studied in [18] is composed of a few tens of SCs per MC. Also, the forced termination probability is not obtained there.

One of the earlier studies of repacking in wireless networks was presented in [19]. There, the authors resort to the classical fixed-point approximation for loss systems [20]. This approximation has also been used in many other studies [21,18,14]. The great advantage of this approximation is that each cell can be treated independently, by assuming that call arrivals from adjacent cells (handovers) follow a homogeneous Poisson process. This assumption is known to be sufficiently accurate for practical purposes. See for example [10]. Our approach is also based on the same principle.

One of the significant differences of our work with previous ones such as [10], and references therein, is that we approximately solve the multidimensional Markov process that describes the time evolution of the system. On the contrary, the aim in many other works such as [10] is to accurately characterize the traffic that arrives to a tagged base station (BS) of a given tier, composed by the aggregated overflow traffic from the BSs of an immediately lower tier. Also, observe that except [18], previous works do not study analytically the impact of repacking in the system performance.

The system we study suffers from the curse of dimensionality problem. To illustrate the magnitude of the problem we have to solve, we will refer to the simplest system analyzed, which is composed by a single macrocell with 50 channels and 50 SCs with 4 channels each. The number of states of the multidimensional Markov process that models the system is on the order of $10^{29}$ states. To determine this figure we considered the state aggregation approach described in Section 4, which greatly reduces the state space. In addition, the solution of the Markov process has not a product-form, which might drastically reduce the computational complexity required to obtain it.

### 1.2 Main Contributions

In summary, the main contributions of this paper are:

- A cellular network were MCs overlay hundreds of SCs is analyzed from the traffic perspective. In the studied system, new calls generated at SCs are first offered to the SC access point, and if rejected, are offered to the MC BS (*call overflow*). Also, when a terminal with an ongoing call serviced by the MC enters a SC coverage area, it will try to attach to its access point, if free SC channels are available. Otherwise, the call will continue with the MC channel, until a call *repacking* is executed.
- We obtain performance parameters such as the blocking probability of MCs, as well as SCs, the probability of forced termination of accepted calls and the carried traffic. The distribution of the number of handover executed before a forced termination occurs, and the carried traffic in different mobility scenarios are also derived.
- To handle the huge dimensionality of the problem, aggravated by the fact that the original multidimensional Markov process is not reversible, three strategies are combined: i) an equivalent reversible Markov process is constructed for which a product-form stationary distribution exists; ii) the state space is transformed by aggregating states that are not relevant to obtain the performance parameters of interest; and iii) the desired performance parameters are obtained by convolution.

– The performance parameters determined by the proposed approximation are validated by simulation, showing that an excellent accuracy is obtained. Note that in the simulation model the arrival process of calls to a cell from adjacent cells is implicitly defined by the system call dynamics and, therefore, it will most likely do not follow a Poisson process. Also, note that while obtaining the performance parameters by simulation can take on the order of two hours for the largest system studied, it only takes one to two seconds by the proposed analytical model in the same conventional desktop computer.

The rest of the paper is organized as follows. The system model is defined in Section 2. An approximate solution of the CTMC is presented in Section 3. This solution is further elaborated by aggregating states as described in Section 4. In Section 5 we determine the different performance parameters by convolution. The forced termination probability is determined in Section 6. The numerical evaluation of the proposed approximation, along with the computation of the carried traffic, is presented in Section 7. Finally, Section 8 concludes the paper.

## 2 System model

We consider the service area of a given cellular operator covered by a set of MCs, where each MC contains multiple non-overlapping SCs that are fully overlaid by the MC. Here, a moving terminal with an ongoing call that leaves a SC will try to attach to the corresponding MC, before visiting another SC or MC. This scenario might be appropriate for indoor SC deployments with closed access. However, the model can be extended to the overlapping case.

We consider the homogeneous case where all MCs are statistically identical and independent. Consequently, the global performance of the system can be analyzed focusing on a single MC, together with its SCs. Note that the independence assumption is validated by the results of Section 7. The model is more general and can be applied to heterogeneous load scenarios as well. We focus on a single MC labeled as $m$. Denote by $\mathcal{M}$ the set of MCs that are neighbors of MC $m$, by $\mathcal{S}$ the set of SCs that are overlaid by MC $m$, and by $F$ the cardinality of $\mathcal{S}$.

For simplicity, terminals generate calls (sessions) of a single streaming traffic class (service), that require a single channel. Let $C_m$ and $C_s$ , $s \in \mathcal{S}$ , be the number of resource units, or channels, of MC $m$ and SC $s$, respectively. Let the vector

$$\boldsymbol{x} = (x_1, \ldots, x_s, \ldots, x_F, x_m)$$

define the system state, where $x_m$ and $x_s$ are the number of ongoing calls *residing* at MC $m$ and SC $s$, respectively. Then, with the appropriate assumptions, $\boldsymbol{x}$ defines an irreducible ergodic finite state CTMC, with state space

$$\mathcal{X} := \{\boldsymbol{x} : 0 \leq x_s \leq C_s + C_m \,,\; 0 \leq x_m \leq C_m \,,\; \forall s \in \mathcal{S}\} \,.$$

Observe that $x_s$ can be as large as $C_s + C_m$ , because we study the case where calls in a SC might borrow channels from the umbrella MC. This is the reason why we emphasize that in the Markov process of interest, the elements of the state vector $\boldsymbol{x}$ refer to calls *residing* at the corresponding MC or SC. Then, the elements of $\boldsymbol{x}$ do not refer to the number of channels occupied at the MC or SC, although this information can be inferred from $\boldsymbol{x}$.

For the sake of mathematical tractability, we make the common assumptions of Poisson arrival processes for new calls and changes of residence. Also, we assume exponentially distributed session duration and cell residence times. The consideration of non-exponential distributions is left for further study. Note that the assumption about the arrival process of changes of residence is validated by the results provided in Section 7.

Let $\lambda_m^n$ and $\lambda_s^n$ be the new call arrival rate to the MC BS and SC access point (SAP) $s$, respectively. Let $\mu^d$ be the call (session) unencumbered duration rate. Let $\mu_m^r$ and $\eta_m = \mu^d + \mu_m^r$, be the MC residence time and channel holding time rates, respectively. Also, for SC $s$, let $\mu_s^r$ and $\eta_s = \mu^d + \mu_s^r$, be the residence time and channel holding time rates, respectively.

Note that the mobility pattern of users changes according to the cell (location) the terminal is residing at. That is, mobile terminals with an ongoing call attached to the SAP of a SC will exhibit a low mobility pattern ($\mu_s^r \leq \mu^d$), as SCs will mainly support indoor traffic. On the other hand, terminals attached to the BS of a MC will exhibit higher mobility ($\mu_m^r \geq \mu^d$). Also, attaching fast moving terminals to SCs might be a bad practice, due to the high signaling load [22].

Define $H$ as the fraction of ongoing calls that, after ending its residence in the MC, move to any SC of the same MC. Then, $1 - H$ is the fraction of ongoing calls that, after ending its residence in the MC, move to a neighbor MC. Note that $H$ can be derived from historical information. According to the deployment scenario defined, $H$ should take a low value, as with closed access SCs the probability of a terminal attaching to a SC in a MC will be low.

Let $h_s^m$ be the fraction of ongoing calls residing in the MC that perform a *change of residence* (COR) towards SC $s$. For exponentially distributed call duration and cell residence time, $h_s^m = p_s^m H \mu_m^r / \eta_m$, where $p_s^m$ is the probability of selecting SC $s$. Note that when SCs are selected with equal probability, $p_s^m = 1/F$.

Likewise, let $h_r^m$ be the fraction of ongoing calls residing in MC $m$ that perform a handover attempt to MC $r$, $r \in \mathcal{M}$. Then, $h_r^m = p_r^m \left(1 - H\right) \mu_m^r / \eta_m$ , where $p_r^m$ is the probability of selecting MC $r$. If neighbor MCs are selected with equal probability, $p_r^m = 1/n_g$ , where $n_g$ is the number of MCs that are neighbors of MC $m$. Typically, $n_g = 6$ for hexagonal tessellations.

Finally, let $h_m^s = \mu_s^r / \eta_s$ be the fraction of ongoing calls residing in SC $s$, $s \in \mathcal{S}$, that perform a COR to MC $m$. When SC calls perform a COR occurs, it is always towards the umbrella MC, i.e., $p_m^s = 1$. A COR inside a MC might trigger a handover or not, as explained below. However, when a COR is performed, the system state ($\boldsymbol{x}$) changes, regardless of it triggers a handover or not. Clearly, $n_g h_r^m + F h_s^m + \left(\mu^d / \eta_m\right) = 1$ . A list of all model parameters is given in Table 2.

## 2.1 Channel assignment

For the channel assignment, the *SC priority scheme* is considered. A new call generated by a terminal that is covered by a SC is first offered to the corresponding SAP, and if blocked, it is then offered to the MC BS. Note that the call might also be blocked at the MC, and will be finally lost. The second type of call attempts are referred to as call *overflows* or *directed retries* towards the umbrella MC [23]. Also, a terminal that resides in a SC but with an ongoing call serviced by its umbrella

**Table 1** Model notation.

| | | | |
|---|---|---|---|
| $C_m$ | MC channels | $C_s$ | SC $s$ channels |
| $F$ | SCs per MC | $n_g$ | number neighbor MCs |
| $\lambda_m^n$ | MC new call arrival rate | $\lambda_s^n$ | SC $s$ new call arrival rate |
| $\gamma_m^o$ | MC offered call rate | $\gamma_s^o$ | SC $s$ offered call rate |
| $\gamma_m^a$ | MC accepted call rate | $\gamma_s^a$ | SC $s$ accepted call rate |
| $H$ | fraction of CORs to SCs | $\mu^d$ | call duration rate |
| $\mu_m^r$ | MC residence rate | $\mu_s^r$ | SC $s$ residence rate |
| $p_r^m$ | transition prob. MC $m$ to MC $r$ | $p_s^m$ | transition prob. MC to SC $s$ |
| $h_r^m$ | fraction of OCH MC $m$ to MC $r$ | $h_s^m$ | fraction of OCCOR MC to SC $s$ |
| $\eta_m$ | MC ch. holding time $(\mu^d + \mu_m^r)$ | $\eta_s$ | SC $s$ ch. holding time $(\mu^d + \mu_s^r)$ |
| $P_m^B$ | MC blocking prob. | $P_s^B$ | SC $s$ blocking prob. |
| $P_s^L$ | SC $s$ loss prob. | $P_s^R$ | SC $s$ COR failure prob. |
| $P_s^D$ | directed handover prob in SC $s$ | $P^{sc}$ | call success completion prob. |
| $P^{ft}$ | global forced term. prob. | $P_m^{ft}$ | MC forced termination prob. |
| $P_s^{ft}$ | SC $s$ forced termination prob. | $P_s^{ft*}$ | SC $s$ forced term. prob. ch. in MC |

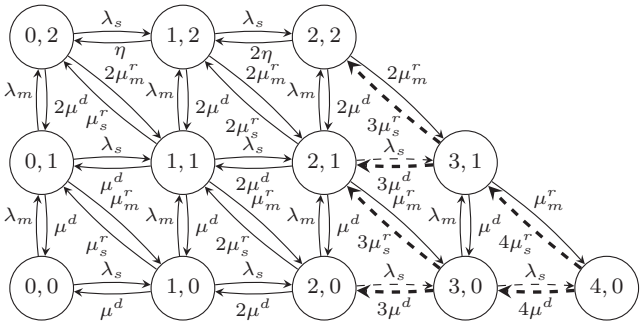


**Fig. 1** State and transition diagram of a system with one MC, a single SC, and with terminals moving along the MC. States defined by $(x_s, x_m)$.

MC, will try to release the MC channel as soon as a free SC channel becomes available. This is known as *directed handover* or *call repacking* [24].

This scheme is motivated by the fact that if the call can be serviced by the SAP, less energy will be used and less interference will be generated to other terminals. On the other hand, when no free channels are available in a SC at a new call arrival, executing a directed retry might be desirable to blocking the call.

When a call residing at a MC moves into a SC (a COR occurs), a handover might not be executed. If the destination SC has all channels busy, then the call continues its service with the umbrella MC. However, if the destination SC has free channels, a handover will be executed to save energy, reduce interference and save valuable resources at the umbrella MC. Also, if a call residing at the SC moves into the MC coverage area (a COR occurs), a handover will not be executed if the call was being serviced by the MC. On the other hand, if a call residing at the SC and using a SC channel moves into the MC, a handover attempt will be issued. The call will first try to swap its channel with the MC channel of another call residing in the same SC (*call repacking*). If such calls are not found, then a handover attempt to the MC is executed. If the MC has all channels busy, the call will be *forced to terminate*.

As an example, let us consider a system with one MC and a single SC, where the number of channels of the MC and SC are $C_m = 2$ and $C_s = 2$, respectively. Clearly, the state vector $(x_s, x_m)$, that describes the number of ongoing calls *residing* at the SC ($x_s$) and the MC ($x_m$), defines an irreducible ergodic finite state CTMC with state space, $\mathcal{X} := \{\boldsymbol{x} = (x_s, x_m)\}$, where $0 \leq x_s \leq C_s + C_m$ , $0 \leq x_m \leq C_m$ , and whose state and transition diagram is shown in Fig. 1. For simplicity, we do not consider handovers across the MC boundary by now. They will be introduced in the model later. At the top of Fig. 1, arrows with weights $\eta$ and $2\eta$, $\eta = \mu^d + \mu_s^r$, correspond to transitions where a SC call either terminates or leaves the SC and performs a handover attempt to the MC. As the MC has no free channels, the handover will be rejected and the call will be forced to terminate.

Note that in Fig. 1, *directed retries* are shown as thin discontinuous arcs, while transitions where *directed handovers* might occur are shown as thick discontinuous arcs. For example, the transition from $(3, 0)$ to $(4, 0)$ occurs when a new call arrives to the SC. As no free SC channels are available, the call is served by the MC (*directed retry*), while it is residing in the SC.

Also, the transition from state $(4, 0)$ to state $(3, 0)$ occurs when a call in the SC successfully terminates. If the call that terminates was using a SC channel, then the new free SC channel is immediately assigned to one of the other two calls using MC channels. Likewise, the transition from state $(4, 0)$ to state $(3, 1)$ occurs when a call in the SC moves to the MC. If the call that did the COR was using a SC channel, then it will swap its SC channel with the MC channel of another call residing in the same SC (*directed handover*). Note that as the COR occurs in state $(4, 0)$ there will be two calls in the SC using MC channels. If the call that did the COR was using a MC channel, then no *directed handover* occurs.

The objective of the study is to determine performance parameters such as the blocking probability of MC and SCs, and the forced termination probability of an accepted call. This performance parameters can be obtained from the stationary distribution of the CTMC that models the system. However, the huge dimensionality of the state space in any realistic scenario makes the exact computation of this distribution an unfeasible task. In the following section we propose an approximate methodology to determine the stationary distribution and the common performance parameters.

## 3 Approximate solution of the CTMC

Figure 2 shows the state and transition diagram of the same system described in Fig. 1, but where terminals do not move (static). This transition diagram meets the Kolmogorov criterion, i.e., for any cycle along the state and transition diagram, the product of the transition rates in the clockwise direction is equal to the product of the transition rates in the anticlockwise direction [25].

Then, the CTMC defined by Fig. 2 is *reversible*, and its stationary distribution has a *product-form solution* [25]. Let $\pi(\boldsymbol{x})$ be the stationary probability of state $\boldsymbol{x}$, i.e., the fraction of time the system spends in $\boldsymbol{x} = (x_s, x_m)$ . Then, $\pi(\boldsymbol{x}) = G^{-1}\pi(x_s)\,\pi(x_m)$ , where $\pi(x_m)$ and $\pi(x_s)$ are the stationary probabilities of finding the the unidimensional birth-and-death (B&D) processes of Fig. 3 in states $x_m$ and $x_s$ respectively, and $G$ is a normalization constant. Note that very

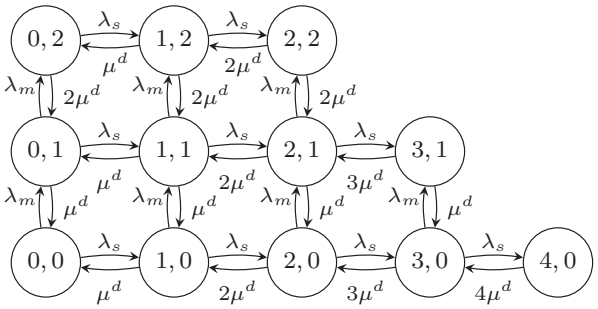


**Fig. 2** State and transition diagram of a system with one MC, a single SC, and with static terminals. States defined by $(x_s, x_m)$.

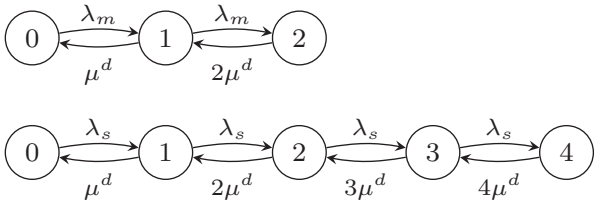


**Fig. 3** Unidimensional state and transition diagrams of the MC (top) and SC $s$ (bottom).

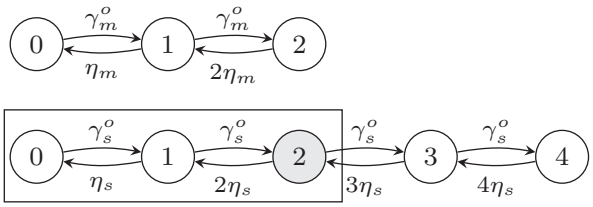


**Fig. 4** Unidimensional state and transition diagrams of the MC (top) and SC $s$ (bottom). Inside the box, states that will be aggregated.

efficient recursions exist to determine $\pi(x_m)$ and $\pi(x_s)$. States in Fig. 3 represent the number of ongoing calls residing at the MC and SC, respectively.

However, we study a system with mobility such as the one described in Fig. 1. To this end, we use the following approximation. We incorporate the impact of mobility to the model of Fig. 2 in such a way that its reversibility property is preserved. Then, we expect that its stationary distribution approximates the one of the system of Fig. 1.

Consider a simple network composed by MCs overlaying SCs. The number of ongoing calls residing at the MC and each of the SCs is modeled by the unidimensional B&D processes of Fig. 4. In this figure, the top process models the MC and the bottom one a SC. Let $\gamma_m^o$ ($\gamma_m^a$) and $\gamma_s^o$ ($\gamma_s^a$) be the offered (accepted) call rates

to (at) MC $m$ and SC $s$, $s \in \mathcal{S}$, respectively. Then,

$$\gamma_s^o = \lambda_s^n + \gamma_m^a \, h_s^m \,, \tag{1}$$

$$\gamma_m^o = \lambda_m^n + \sum_{r \in \mathcal{M}} \gamma_r^a \, h_m^r + \sum_{s \in \mathcal{S}} \gamma_s^a \, h_m^s \,. \tag{2}$$

The terms in $\gamma_s^o$ are the new call arrival rate and the rate of calls accepted at the MC that perform a COR towards SC $s$. The terms in $\gamma_m^o$ are the new call arrival rate, the rate of calls accepted at neighbor MCs that perform a handover to MC $m$, and the rate of calls accepted at SCs of MC $m$ that perform a COR towards MC $m$.

$$\gamma_s^a = \lambda_s^n \left(1 - P_s^L\right) + \gamma_m^a \, h_s^m \,, \tag{3}$$

$$\gamma_m^a = \lambda_m^n \left(1 - P_m^B\right) + \left(1 - P_m^B\right) \sum_{r \in \mathcal{M}} \gamma_r^a \, h_m^r + \sum_{s \in \mathcal{S}} \gamma_s^a \, h_m^s \left(1 - P_s^R\right) \,. \tag{4}$$

The terms of $\gamma_s^a$ are the rate of new call arrivals that are not lost and the rate of calls accepted at the MC that perform a COR towards SC $s$. Note that these last calls are never blocked as they can retain the MC channel if necessary. The terms of $\gamma_m^o$ are the rate of new call arrivals that are not blocked, the rate of calls accepted at neighbor MCs that perform a handover to MC $m$ and are not blocked, and the rate of calls accepted at SCs of MC $m$ that perform a successful COR towards MC $m$.

Note that $P_m^B$ is the MC blocking probability, i.e., the fraction of time all its channels are busy, $P_s^L$ is the SC $s$ loss probability, i.e., the fraction of offered new calls that will be lost, $P_s^R$ is the SC $s$ probability that a COR towards the MC fails, i.e., the fraction of ongoing SC calls with a SC channel that leave the SC and are lost. This COR failure occurs when two simultaneous conditions occur, the MC has no free channels and there are no SC calls serviced by MC channels with which a channel swapping can be executed. They can be determined as,

$$P_m^B = \sum_{\boldsymbol{x}} \pi\left(\boldsymbol{x}\right) \,,\, \boldsymbol{x} : x_m + \beta_{\boldsymbol{x}} = C_m \,, \tag{5}$$

$$P_s^B = \sum_{\boldsymbol{x}} \pi\left(\boldsymbol{x}\right) \,,\, \boldsymbol{x} : x_s \geq C_s \,, \tag{6}$$

$$P_s^L = \sum_{\boldsymbol{x}} \pi\left(\boldsymbol{x}\right) \,,\, \boldsymbol{x} : x_s \geq C_s \,, x_m + \beta_{\boldsymbol{x}} = C_m \,, \tag{7}$$

$$P_s^R = \sum_{\boldsymbol{x}} \pi\left(\boldsymbol{x}\right) \,,\, \boldsymbol{x} : x_s \leq C_s \,, x_m + \beta_{\boldsymbol{x}} = C_m \,, \tag{8}$$

where $\beta_{\boldsymbol{x}} = \sum_\sigma \beta\left(x_\sigma\right)$, $\beta\left(x_\sigma\right) = max\left(0, x_\sigma - C_\sigma\right)$ the number of ongoing calls in SC $\sigma$ using a MC channel, and $\sigma$ is the SC index. For completeness, we also define $P_s^B$, the SC $s$ blocking probability, i.e., the fraction of time all its channels are busy. Observe that $P_s^L \leq P_s^B$. The expression for $P_s^R$ in (8) is an approximation, as it assumes that the PASTA (Poisson Arrivals See Time Averages) property holds.

Note that the multidimensional stationary distribution $\boldsymbol{\pi} = \{\pi\left(\boldsymbol{x}\right)\}$ is required to determine $P_m^B$, $P_s^L$ and $P_s^R$. Also, $P_m^B$, $P_s^L$ and $P_s^R$ are required to determine

$\boldsymbol{\pi}$. Then, equations (1) to (8) define a *fixed-point equation* that can be solved iteratively [20]. Along the iterative process, $\boldsymbol{\pi}$ is assumed to have a product-form as described below.

As an example, Fig. 4 shows the B&D processes for a system with $C_m = C_s = 2$, $s = 1, 2$. We consider that convergence has been achieved when the maximum difference between the elements of the stationary distributions between two consecutive iterations is lower than a certain bound, for example $\epsilon = 10^{-4}$. That is, $max \left| \pi_m^{(t+1)}(x_m) - \pi_m^{(t)}(x_m) \right| \leq \epsilon, \forall x_m$, and $max \left| \pi_s^{(t+1)}(x_s) - \pi_s^{(t)}(x_s) \right| \leq \epsilon, \forall x_s$, where the super-index $t$ indicates the iteration number, and $\boldsymbol{\pi_m}$ and $\boldsymbol{\pi_s}$ are the unidimensional stationary distributions of MC and SC.

Convergence has been always achieved in our numerical experiments. After convergence, the resulting MC B&D process has *incorporated* the impact of COR arrivals from SCs and other MCs. Likewise, the resulting SC B&D process has *incorporated* the impact of COR arrivals from the MC. That is, all B&D process have *incorporated* the impact of mobility. In addition, we might consider that the *converged* B&D *processes are independent*, as further iterations will not change their stationary distributions.

Let $Y' = (X_1, X_2, X_m)$ be a multidimensional B&D process, where $X_m$ ($X_s$) is the unidimensional MC (SC) B&D process. As $X_m$, $X_1$ and $X_2$ are *reversible and independent* processes, then $Y'$ is also reversible, and its stationary distribution has a product-form [25], that is given by

$$\pi'(x_1, x_2, x_m) = \pi_1(x_1) \cdot \pi_2(x_2) \cdot \pi_m(x_m) ,$$

where $\boldsymbol{\pi_m}$ ($\boldsymbol{\pi_s}$) is the stationary distribution of the MC (SC) unidimensional *converged* B&D process.

Note that the state space of $Y'$ is larger than the state space of the system under study, then a *truncation of the state space* is required. As an example, for a system with $F = 2$ and $C_m = C_1 = C_2 = 2$, state $(4, 4, 2)$ is not feasible as it would require that $C_m = 6$. Let $Y$ be the truncated process, then $Y$ is also reversible and its stationary distribution preserves the product-form of the one of $Y'$, except a normalization constant [25]. Its stationary distribution is given by

$$\pi(x_1, x_2, x_m) = G^{-1} \pi_1(x_1) \cdot \pi_2(x_2) \cdot \pi_m(x_m) , \tag{9}$$

where $G$ is the normalization constant.

The normalization constant is found by adding all feasible states, such us those one defined by (9). However, the number of states is prohibitively large in realistic scenarios, due to the huge dimensionality of the state space as the number of SC grows (curse of dimensionality problem). Then, more elaborate approximations are required to reduce the computational complexity.

## 4 CTMC state aggregation

To make the presentation easy to follow, along this section we will refer to the system of Fig. 4, with $F = 2$, defined in previous section. Observe in (6) that to determine $P_s^B$, only part of the distribution $\boldsymbol{\pi_s}$ is required, in particular, $x_s \geq C_s$. That is, for every SC, all states in the rectangle of Fig. 4 can be aggregated into a single state. For SC $s$, let this set of states be denoted by $\mathcal{X}_s^a$. In addition, the

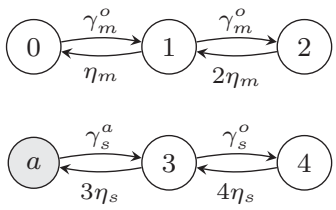


**Fig. 5** Unidimensional state and transition diagrams of the MC (top) and SC $s$ (bottom). State $a$ represents the aggregation of states.

fraction of time spent in state $x_s = C_s$ during a residence in $\mathcal{X}_s^a$ is required. Note that in Fig. 4 this state is $x_s = C_s = 2$, the state colored in gray.

Let $T_s^a$ be the random variable residence time in $\mathcal{X}_s^a$. That is, the time since the system enters the aggregate through state $C_s$, until it leaves the aggregate also through $C_s$. $\overline{T_s^a}$ is given by

$$\overline{T_s^a} = \frac{\sum_{n \in \mathcal{X}_s^a} \pi_s(n)}{\gamma_s^o \pi_s(C_s)} = \frac{1}{\gamma_s^o Er_B(A_s, C_s)}, \tag{10}$$

where $Er_B(A_s, C_s)$ is the Erlang-B formula and $A_s = \gamma_s^o/\eta_s$. Then, the fraction of time spent in state $x_s = C_s$ during a residence in $\mathcal{X}_s^a$ is given by $Er_B(A_s, C_s)$.

## 5 Computing performance parameters by convolution

Due to the state aggregation proposed in Fig. 5, all traffic flows now compete for the same set of $C_m$ channels. Observe in Fig. 5 that $\gamma_s^a = 1/\overline{T_s^a}$, where $\overline{T_s^a}$ was defined in (10). Then, under the assumption that the multidimensional CTMC is reversible, the stationary distribution of the number of channels occupied in the MC can be efficiently obtained by convolution [26,27]. Note that the distribution $\boldsymbol{\pi} = \{\pi(\boldsymbol{x})\}$ cannot be obtained by convolution. Instead we determine the distribution of the number of resource units (channels) occupied at the MC by different subsystems of the system under study, and from them the performance parameters of interest.

Let us define the distribution $\boldsymbol{\pi_s'} = \left[\pi_s'(0), \pi_s'(1), \ldots, \pi_s'(C_m)\right]$ , $\pi_s'(0) = \sum_{n=0}^{C_s} \pi_s(n)$ , $\pi_s'(n) = \pi_s(C_s + n)$, where $1 \leq n \leq C_m$, and $\boldsymbol{\pi_s}$ is the stationary distribution of SC $s$ unidimensional B&D processes. Note that $\pi_s'(0)$ is the unnormalized probability of SC $s$ not borrowing any MC channel.

Let $\boldsymbol{\Omega_2'} = \boldsymbol{\pi_1'} \otimes \boldsymbol{\pi_2'}$ be a distribution of $2(C_m + 1) - 1$ terms, where the operator $\otimes$ denotes the convolution of distributions. Let $\boldsymbol{\Omega_2}$ be a distribution composed of the first $(C_m + 1)$ terms of $\boldsymbol{\Omega_2'}$, i.e., we are truncating the state space, as states larger than $C_m$ are not feasible. Also, note that distribution $\boldsymbol{\Omega_2}$ is not normalized. Let $\boldsymbol{\Omega_3'} = \boldsymbol{\Omega_2} \otimes \boldsymbol{\pi_{m3}'}$ and $\boldsymbol{\Omega_3}$ be a distribution composed of the first $(C_m + 1)$ terms of $\boldsymbol{\Omega_3'}$, and so on. Let $\boldsymbol{G}^F = \boldsymbol{\Omega}_F$ be the unnormalized distribution obtained by selecting the first $(C_m + 1)$ terms of the convolution of all $F$ SC stationary distributions $(\boldsymbol{\pi_s'})$.

Also, let $\boldsymbol{G}^{\overline{s}} = \boldsymbol{\Omega}_{F-1}^{\overline{s}} \otimes \boldsymbol{\pi_m}$ and $\boldsymbol{G} = \boldsymbol{\Omega}_F \otimes \boldsymbol{\pi_m}$, be the unnormalized distributions obtained by selecting the first $(C_m + 1)$ terms of the convolution of all SC stationary distributions except SC $s$ and the MC stationary distribution $(\boldsymbol{\pi_m})$,

and all SC stationary distributions and the MC stationary distribution (complete system), respectively. Then,

$$P_m^B = \frac{1}{G}\sum_{n=0}^{C_m} \pi_m(n)\, G_{C_m-n}^F = \frac{G_{C_m}}{G}\,, \tag{11}$$

$$P_s^B = \frac{\pi_s'(0)}{G} Er_B(A_s, C_s) \sum_{k=0}^{C_m} G_k^{\overline{s}} + \frac{1}{G}\sum_{n=1}^{C_m} \pi_s'(n) \sum_{k=0}^{C_m-n} G_k^{\overline{s}}\,, \tag{12}$$

where $G_k^F$, $G_k^{\overline{s}}$ and $G_k$ are the $k$-th elements of $\boldsymbol{G}^F$, $\boldsymbol{G}^{\overline{s}}$ and $\boldsymbol{G}$, respectively, and $G = \sum_{n=0}^{C_m} G_n$ is the normalization constant. Note that terms in $P_m^B$ corresponds to states where $n$ MC channels are used by calls residing at the MC, and $C_m - n$ MC channels by calls residing at SCs, i.e., states with all MC channels occupied. Also, the two terms in $P_s^B$ corresponds to the faction of time spent in state $x_s = C_s$ (gray state in Fig. 4), and in feasible states where $x_s \geq C_s$, respectively. Also,

$$P_s^L = \frac{\pi_s'(0)}{G} Er_B(A_s, C_s)\, G_{C_m}^{\overline{s}} + \frac{1}{G}\sum_{n=1}^{C_m} \pi_s'(n)\, G_{C_m-n}^{\overline{s}}\,, \tag{13}$$

$$P_s^R = \frac{1}{G}\pi_s'(0)\, G_{C_m}^{\overline{s}}\,. \tag{14}$$

Observe that the two terms in $P_s^L$ correspond to: first, states where SC $s$ has $C_s$ channels occupied and all MC channels are occupied, and second, states where SC $s$ has $C_s + n$ channels occupied and there are $C_m - n$ MC channels are occupied. Clearly, $P_s^L \leq P_s^B$, as the terms in $P_s^L$ are contained in $P_s^B$. Also, the term in $P_s^R$ correspond to states where SC $s$ has $x_s \leq C_s$ channels occupied and all MC channels are occupied. $P_m^B$, $P_s^B$, $P_s^L$ and $P_s^R$ where defined in (5) to (8).

Obtaining blocking probabilities by the convolution method is computationally very efficient. For a large system with $F = 200$ SCs, $C_m = 100$ and $C_s = 6$ (see Section 7 for details), the computation time to obtain the converged unidimensional distributions and the performance parameters is around one to two seconds in a conventional laptop computer.

## 6 Forced termination distribution

In this section we formulate the probability that a call is forced to terminate in its $k$-th handover. For that purpose, it is required to determine the probability that a call residing in SC $s$, but using a MC channel, will eventually perform a *directed handover* in SC $s$.

### 6.1 Probability of directed handover

Let $P_s^D$ be the fraction of calls using a MC channel that reside in SC $s$, $s \in \mathcal{S}$, that will execute a *directed handover* before leaving SC $s$. Recall that a call may leave a SC because it terminates successfully at the SC, or because it performs a COR.

When a SC call using a SC channel terminates, the new free SC channel is assigned to one of the calls using a MC channel that reside in the SC (if any).

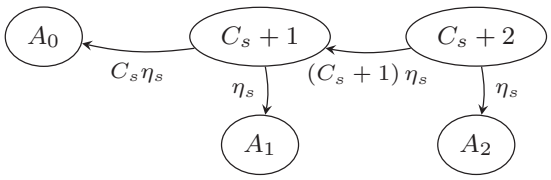


**Fig. 6** Absorption process for a call occupying position 2 at the directed handover FIFO queue in SC $s$.

Different disciplines can be defined to select the call that will execute a *directed handover*. Clearly, $P_s^D$ is not dependent on the channel assignment discipline, but only depends on the SC call dynamics. For formulation simplicity, we determine $P_s^D$ when the call that performs the directed handover is chosen according to a FIFO (*first-in first-out*) discipline, i.e., the call with a MC channel that has been residing in the SC longer. Then, $P_s^D$ can be determined by

$$P_s^D = \sum_{i=0}^{C_m-1} \varphi_s\left(i\right) \prod_{k=0}^{i} \frac{\left(C_s+k\right)\eta_s}{\eta_s+\left(C_s+k\right)\eta_s}\,, \tag{15}$$

$$\varphi_s\left(i\right) = \begin{cases} \dfrac{1}{\Phi_s} Er_B\left(A_s, C_s\right) \pi'_s\left(0\right) \sum_{k=0}^{C_m-1} G_k^{\overline{s}}\,, & i=0\,, \\ \dfrac{1}{\Phi_s} \pi'_s\left(i\right) \sum_{k=0}^{C_m-1-i} G_k^{\overline{s}}\,, & 0<i\leq C_m-1\,. \end{cases} \tag{16}$$

where $\Phi_s = Er_B\left(A_s, C_s\right) \pi'_s\left(0\right) \sum_{k=0}^{C_m-1} G_k^{\overline{s}} + \sum_{i=1}^{C_m-1} \pi'_s\left(i\right) \sum_{k=0}^{C_m-1-i} G_k^{\overline{s}}$ is a normalization constant, and $\varphi_s\left(i\right)$ is the probability that a call that is using a MC channel, upon initiating its residence in SC $s$, finds $i$ other calls in the same SC using MC channels. Clearly, in (16) it is assumed that active mobiles arrive to the SC following a Poisson process, and therefore the PASTA property holds [25].

The term associated to the discrete product operator in (15) defines the probability that a SC call with a MC channel executes a *directed handover* before leaving the SC, conditioned on finding $i$ other calls with MC channels upon arrival to the SC. We describe this term with an example. Assume that a call arriving at SC $s$ finds it in state $C_s+1$, i.e., there are $C_s+1$ calls residing in the SC, $C_s$ with SC channels and 1 with a MC channel. We refer to the newly arrived call as the *tagged call*. Then, the tagged call will initiate its residence in the SC keeping its MC channel, and will occupy the second position in the queue of calls waiting for a SC channel.

Since call durations and residence times are exponentially distributed, and focusing exclusively on the $C_s+1$ calls found upon arrival plus the tagged call, we model the time evolution of the number of calls in the queue as a phase-type process with absorption states as shown in Fig. 6. The absorption state $A_0$ is reached when the tagged call executes a *directed handover*. This happens with probability

$$p = \frac{\left(C_s+1\right)\eta_s}{\eta_s+\left(C_s+1\right)\eta_s} \cdot \frac{C_s\eta_s}{\eta_s+C_s\eta_s}\,, \tag{17}$$

where the first term gives the probability that one of the $(C_s+1)$ calls found by the tagged call upon arrival leaves the SC by successful termination, by executing

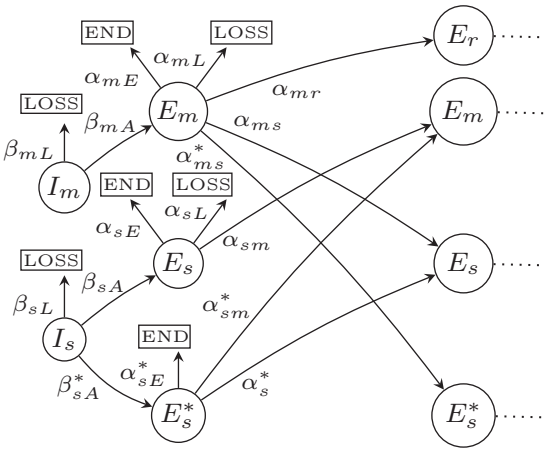


**Fig. 7** State and transition diagram that describes the time evolution of an ongoing call moving along the service area, since initiation to termination.

a handover to the MC or because the one with a MC channel returns to the MC. This moves the tagged call to the head of the queue. The second term gives the probability that one of the $C_s$ calls with channels in the SC leaves the SC, and then, its channel can be used by the tagged call to execute a *directed handover*. Note that the tagged call leaves the SC before executing a *directed handover* with probability $1-p$ (either because it terminates successfully or returns to the MC), where $p$ is given by (17).

## 6.2 Forced termination probability

Figure 7 defines the possible *states* of a call since initiation to termination. A new call initiated at MC $m$ ($I_m$) is accepted (transits to state $E_m$) with probability $\beta_{mA} = 1 - P_m^B$, and blocked with probability $\beta_{mL} = P_m^B$. A new call initiated at SC $s$ ($I_s$), $s \in \mathcal{S}$, is accepted with a SC channel (transits to state $E_s$) with probability $\beta_{sA} = 1 - P_s^B$, accepted with a MC channel (transits to state $E_s^*$) with probability $\beta^*_{sA} = P_s^B - P_s^L$, and lost with probability $\beta_{sL} = P_s^L$.

A call residing at MC $m$ (in state $E_m$) terminates successfully with probability $\alpha_{mE} = \mu^d/\eta_m$, is accepted at MC $r$ (transits to state $E_r$), $r \in \mathcal{M}$, with probability $\alpha_{mr} = h_r^m\left(1 - P_r^B\right)$, is blocked at any neighbor MC with probability $\alpha_{mL} = \sum_{r\in\mathcal{M}} h_r^m P_r^B$, is accepted at SC $s$ with a SC channel (transits to state $E_s$) with probability $\alpha_{ms} = h_s^m\left(1 - P_s^B\right)$, is accepted at SC $s$ with a MC channel (transits to state $E_s^*$) with probability $\alpha^*_{ms} = h_s^m P_s^B$. Note that $\alpha_{mE} + \alpha_{mL} + \sum_{r\in\mathcal{M}} \alpha_{mr} + \sum_{s\in\mathcal{S}} (\alpha_{ms} + \alpha^*_{ms}) = 1$.

A call residing at SC $s$ with a SC channel (in state $E_s$) terminates successfully with probability $\alpha_{sE} = \mu^d/\eta_s$, is accepted at the MC (transits to state $E_m$) with probability $\alpha_{sm} = h_m^s\left(1 - P_s^R\right)$, and is blocked at the MC with probability $\alpha_{sL} = h_m^s P_s^R$. Also, $\alpha_{sE} + \alpha_{sm} + \alpha_{sL} = 1$.

A call residing at SC $s$ with a MC channel (in state $E_s^*$) terminates successfully with probability $\alpha_{sE}^* = \left(1 - P_s^D\right) \mu^d / \eta_s$ , is accepted at the MC (transits to state $E_m$) with probability $\alpha_{sm}^* = \left(1 - P_s^D\right) h_m^s$ , and executes a directed handover (transits to state $E_s$) with probability $\alpha_s^* = P_s^D$ . Clearly, $\alpha_{sE}^* + \alpha_{sm}^* + \alpha_s^* = 1$ .

As observed in Fig. 7, a COR between a MC and their SCs, and between MCs is allowed. However, for simplicity, a COR between SCs of the same MC, or between a SC and other MCs is not supported by the model. Note that they could also be included, but at the expense of a higher model complexity. Additionally, observe that for clarity only SC $s$ is depicted in Fig. 7, while the other $F - 1$ are omitted. Likewise, only neighbor MC $r$ is shown, $r \in \mathcal{M}$. Clearly, once the call transits to a new state (states at the right in Fig. 7) the possible outcomes in the evolution of a call are the same ones described above.

Let $f_m(k)$ be the probability that a call initiated at the MC is forced to terminate at its $k$-th handover attempt. Also, let $f_s(k)$ and $f_s^*(k)$ be the probabilities that a call initiated at SC $s$ is forced to terminate at its $k$-th handover attempt, when it uses a channel from the same SC or from the MC, respectively. Note that in the definition of $f_m(k)$, $f_s(k)$ and $f_s^*(k)$ we exploit the memoryless property of the session duration and cell residence time random variables. Then,

$$f_m(k) = \begin{cases} \alpha_{mL} + \sum_s \alpha_{ms}^* f_s^*(1) \,, & k = 1, \\ \sum_{r \in \mathcal{M}} \alpha_{mr} f_r(k-1) + \sum_{s \in \mathcal{S}} \left(\alpha_{ms} f_s(k-1) + \alpha_{ms}^* f_s^*(k)\right) \,, & k > 1 \,. \end{cases} \tag{18}$$

$$f_s(k) = \begin{cases} \alpha_{sL} \,, & k = 1, \\ \alpha_{sm} f_m(k-1) \,, & k > 1 \,, \end{cases} \tag{19}$$

$$f_s^*(k) = \alpha_s^* f_s(k) + \alpha_{sm}^* f_m(k) \,, \quad k \geq 1 \,. \tag{20}$$

Assuming for simplicity that all MCs, as well as all SCs, have equal characteristics, we proceed to obtain an explicit expression for $P^{ft}$. To this end, let us define the following generating functions, $J_m^{ft}(z) = \sum_{k=1}^{\infty} f_m(k) z^k$ , $J_s^{ft}(z) = \sum_{k=1}^{\infty} f_s(k) z^k$ , $J_s^{ft*}(z) = \sum_{k=1}^{\infty} f_s^*(k) z^k$ . Let $P_m^{ft} = J_m^{ft}(1)$, $P_s^{ft} = J_s^{ft}(1)$ and $P_s^{ft*} = J_s^{ft*}(1)$ be the probabilities that a call initiated at the MC, at SC $s$, and at SC $s$ but using a MC channel, is being forced to terminate. $P_m^{ft}$ , $P_s^{ft}$, and $P_s^{ft*}$ can be obtained by solving the following system of linear equations,

$$J_m^{ft}(z) = \alpha_{mL} z + \sum_{r \neq m} \alpha_{mr} z J_m^{ft}(z) + \sum_s \alpha_{ms} z J_s^{ft}(z) + \sum_s \alpha_{mr}^* J_s^{ft*}(z) \,, \tag{21}$$

$$J_s^{ft}(z) = \alpha_{sL} z + \alpha_{sm} z J_m^{ft}(z) \,, \tag{22}$$

$$J_s^{ft*}(z) = \alpha_s^* J_s^{ft}(z) + \alpha_{sm}^* J_m^{ft}(z) \,. \tag{23}$$

Expressions (21) to (23) follow from expressions (18) to (20).

Then,

$$P_m^{ft} = \frac{(1 - \psi_1) f_m(1) + f_m(2)}{(1 - \psi_1 - \psi_2)} \,, \tag{24}$$

$$P_s^{ft} = \alpha_{sm} P_m^{ft} + \alpha_{sL} \,, \quad P_s^{ft*} = \alpha_s^* P_s^{ft} + \alpha_{sm}^* P_m^{ft} \,. \tag{25}$$

where $\psi_2 = \omega_2/(1-\omega_1)$, $\psi_1 = (\omega_3+\omega_4)/(1-\omega_1)$, $\omega_1 = \sum_s \alpha^*_{ms}\alpha^*_{sm}$, $\omega_2 = \sum_s \alpha_{ms}\alpha_{sm}$, $\omega_3 = \sum_s \alpha^*_{ms}\alpha^*_s\alpha_{sm}$, and $\omega_4 = \sum_{r\neq m}\alpha_{mr}$. Note that $f_m(1)$ and $f_m(2)$ can be obtained from (18).

Let $P^{ft}$ be the probability that a call is forced to terminate before it terminates successfully, regardless where it was initiated. Then,

$$P^{ft} = a_m P^{ft}_m + \sum_s \left( a_s P^{ft}_s + a^*_s P^{ft*}_s \right), \tag{26}$$

where $a_m$, $a_s$ and $a^*_s$ are the fraction of new calls initiated (accepted) at the MC, SC $s$ with a SC channel, and SC $s$ with a MC channel, respectively. They are given by, $a_m = (\lambda^n_m/\lambda^a_t)\,\beta_{mA} = (\lambda^n_m/\lambda^a_t)\left(1-P^B_m\right)$, $a_s = (\lambda^n_s/\lambda^a_t)\,\beta_{sA} = (\lambda^n_s/\lambda^a_t)\left(1-P^B_s\right)$, $a^*_s = (\lambda^n_s/\lambda^a_t)\,\beta^*_{sA} = (\lambda^n_s/\lambda^a_t)\left(P^B_s - P^L_s\right)$, $\lambda^a_t = \lambda^n_m\left(1-P^B_m\right) + \sum_s \lambda^n_s\left(1-P^L_s\right)$, where $\lambda^a_t$ is the total (aggregated) new call acceptance rate for all SCs and the MC.

To evaluate the signaling load it might be of interest to determine the distribution of the random variable number of handovers executed by a call. That is, the probability that a call executes $k$ handovers, for $k \geq 0$, before it terminates, either successfully or being forced to terminate. To determine this distribution, a procedure similar to the one described by equations (18), (19) and (20) might be used.

## 7 Numerical Evaluation

To validate the proposed analytical model we compare the blocking probability of the MC $P^B_m$, the blocking probability of any SC, for example SC $s$, $P^B_s$, and the forced termination probability $P^{ft}$, with results obtained by simulation. The simulation model mimics the physical behavior of the system, and therefore it is completely independent from the proposed CTMC model. Also, it emulates an unbounded service area, so the boundary cell effects can be ignored [28]. We use a custom-based discrete-event simulation program.

We define three reference systems as specified in Table 2. We refer to them as small (SS), medium (MS), and large system (LS), respectively. In each of these systems, any MC is surrounded by other MCs of the same characteristics. For simplicity, we assume that when terminals with ongoing calls perform handover attempts to adjacent MCs, all adjacent destination MCs are selected with equal probability, i.e. $p^m_r = (1-H)/n_g$, $\forall r \in \mathcal{M}$, and $p^m_r = 0$ otherwise, where $n_g$ is the number of MCs that are neighbors of MC $m$.

As described in Section 2, $H$ should take a small value to reflect a closed-access SC scenario. We set $H = 0.25$. Then, 75% of the ongoing calls will move to neighbor MCs. Then, a terminal roaming across the service area will attach to one SC every three MC handovers (on average). We consider hexagonal tessellations where a MC is surrounded by $n_g = 6$ neighbor MCs. For simplicity, we set $\mu^d = 1$. Note that performance parameters depend on the ratios $\gamma^o_m/\eta_m$ and $\gamma^o_s/\eta_s$, and not on the individual values of $\gamma^o_m, \gamma^o_s, \eta_m, \eta_s$.

For each reference system, we evaluate the performance parameters for different mobility profiles. For SCs we consider three mobility rates $\mu^r_s = \{0.1, 0.5, 1.0\}$, while for MCs we consider four mobility rates $\mu^r_m = \{1, 2, 3, 4\}$. These rates have

**Table 2** Definition of reference systems.

| | $F$ | $C_m$ | $C_s$ | $\mu^d$ | $H$ |
|---|---|---|---|---|---|
| SS | 50 | 50 | 4 | 1.0 | 0.25 |
| MS | 100 | 100 | 6 | 1.0 | 0.25 |
| LS | 200 | 100 | 6 | 1.0 | 0.25 |

**Table 3** Probabilities and relative errors (%) in the SS.

| $\mu_s^r$ | $\mu_m^r$ | $P_m^B$ | REr | $P_s^B$ | REr | $P^{ft}$ | REr |
|---|---|---|---|---|---|---|---|
| 0.1 | 1 | 1.00 | 2.52 | 2.41 | 0.00 | 0.36 | 4.99 |
| 0.1 | 2 | 0.98 | 2.02 | 2.99 | 0.27 | 0.62 | 3.93 |
| 0.1 | 3 | 1.82 | 1.43 | 3.07 | −0.02 | 1.69 | 3.08 |
| 0.1 | 4 | 1.88 | 1.41 | 3.80 | 0.25 | 2.19 | 2.81 |
| 0.5 | 1 | 1.03 | 2.28 | 3.27 | 0.47 | 0.59 | 4.82 |
| 0.5 | 2 | 0.98 | 2.06 | 2.70 | −0.91 | 0.85 | 4.19 |
| 0.5 | 3 | 0.96 | 2.10 | 3.13 | −0.51 | 1.09 | 4.07 |
| 0.5 | 4 | 0.95 | 2.41 | 3.62 | −0.74 | 1.32 | 4.18 |
| 1.0 | 1 | 1.03 | 2.43 | 1.57 | 0.25 | 0.89 | 4.38 |
| 1.0 | 2 | 0.98 | 2.95 | 2.85 | −0.41 | 1.13 | 5.16 |
| 1.0 | 3 | 0.98 | 2.80 | 3.20 | −0.58 | 1.39 | 5.04 |
| 1.0 | 4 | 0.96 | 2.37 | 2.72 | 0.15 | 1.66 | 4.14 |

been chose such that $\mu_s^r \leq \mu^d$ and $\mu_m^r \geq \mu^d$, as discusses in Section 2. For each reference system and mobility parameters, we adjust the offered load to achieve $P_m^B \approx 1\%$ and $P_s^B \in [2,4]\,\%$ approximately. According to ITU-T Recommendation E.771 the objective for the radio channel blocking probability in the public land mobile network (PLMN) should be similar to the blocking probability experienced by subscribers in fixed network, i.e., 1 to 2 %. This figure is typically used as a practical dimensioning value [29].

The procedure followed for the adjustment gives priority to the setting of $P_m^B$. That is, a load setting is considered acceptable if it complies first with the objective for $P_m^B$, and, second, the value for $P_s^B$ is also acceptable. We are not imposing any condition on $P^{ft}$ when defining the offered load. Note that in some scenarios it is not possible to set values for $P_s^B$ even close to 1% , particularly for moderate to high SC mobility rates. In these scenarios, an important fraction of SC initiated calls will perform a handover to the MC. Then, to achieve that $P_m^B \approx 1\%$, a low load must be offered to SCs, which induces a low SC blocking probability. In other words, a higher SC load could have been offered, provided that the MC was configured with more channels.

We define the relative error (REr) of a measure as $(x-y)/y$ , where $x$ is the value obtained by the analytical model, and $y$ is the value obtained by the simulation model. The sign of REr indicates overestimation (positive) or underestimation (negative). Note that values obtained by simulation are determined by taking the average of different simulations runs. The 95% confidence intervals are also determined. Their radius (half-lengths) are smaller that $10^{-4}$ for all measures.

The relative errors obtained for the SS, MS and LS are shown in Tables 3, 4, and 5, respectively. Observe that the accuracy of the proposed approximation is very good, i.e., the REr is very small. Note that the relative errors for the worse

**Table 4** Probabilities and relative errors (%) in the MS.

| $\mu_s^r$ | $\mu_m^r$ | $P_m^B$ | REr | $P_s^B$ | REr | $P^{ft}$ | REr |
|---|---|---|---|---|---|---|---|
| 0.1 | 1 | 0.65 | 2.53 | 2.41 | −0.03 | 0.18 | 4.22 |
| 0.1 | 2 | 1.46 | 1.70 | 3.43 | −0.28 | 0.66 | 3.15 |
| 0.1 | 3 | 1.24 | 1.62 | 3.45 | −0.23 | 0.79 | 2.83 |
| 0.1 | 4 | 1.10 | 1.53 | 3.45 | −0.03 | 0.91 | 2.51 |
| 0.5 | 1 | 1.54 | 1.73 | 2.21 | −0.48 | 0.85 | 3.34 |
| 0.5 | 2 | 1.40 | 1.88 | 2.87 | −0.14 | 1.04 | 3.49 |
| 0.5 | 3 | 1.24 | 1.70 | 3.63 | −0.25 | 1.13 | 3.38 |
| 0.5 | 4 | 1.30 | 1.32 | 3.68 | −0.77 | 1.42 | 2.73 |
| 1.0 | 1 | 1.87 | 2.23 | 0.10 | −1.07 | 1.62 | 3.30 |
| 1.0 | 2 | 1.48 | 2.26 | 0.23 | −0.35 | 1.72 | 3.38 |
| 1.0 | 3 | 1.22 | 2.26 | 0.33 | −0.09 | 1.76 | 3.29 |
| 1.0 | 4 | 1.78 | 1.24 | 0.98 | −0.44 | 2.81 | 2.25 |

**Table 5** Probabilities and relative errors (%) in the LS.

| $\mu_s^r$ | $\mu_m^r$ | $P_m^B$ | REr | $P_s^B$ | REr | $P^{ft}$ | REr |
|---|---|---|---|---|---|---|---|
| 0.1 | 1 | 1.06 | 2.43 | 0.86 | 0.72 | 0.24 | 3.82 |
| 0.1 | 2 | 1.12 | 2.26 | 1.10 | 0.48 | 0.40 | 3.55 |
| 0.1 | 3 | 1.11 | 1.77 | 2.41 | 0.60 | 0.48 | 2.92 |
| 0.1 | 4 | 1.06 | 1.69 | 2.89 | −0.62 | 0.66 | 2.75 |
| 0.5 | 1 | 1.18 | 2.31 | 0.09 | −2.89 | 0.66 | 3.10 |
| 0.5 | 2 | 1.13 | 2.34 | 0.16 | −3.05 | 0.84 | 3.31 |
| 0.5 | 3 | 0.98 | 2.34 | 0.33 | 0.73 | 0.87 | 3.16 |
| 0.5 | 4 | 1.07 | 1.85 | 0.63 | −1.40 | 1.06 | 2.68 |
| 1.0 | 1 | 1.08 | 3.28 | 0.00 | −2.49 | 0.94 | 4.19 |
| 1.0 | 2 | 1.03 | 2.82 | 0.00 | 1.51 | 1.21 | 3.89 |
| 1.0 | 3 | 1.00 | 2.70 | 0.00 | 2.45 | 1.42 | 3.64 |
| 1.0 | 4 | 1.07 | 2.23 | 0.02 | 0.15 | 1.75 | 3.08 |

configurations do not seem to be affected by the size of the reference system, i.e., with the number of SCs and MC channels. The largest relative error in the tables is around 5%, which is accurate enough in most practical scenarios. Also, note that in the simulation model, COR arrivals to a cell occur according to the system dynamics, and are not compelled to follow any distribution. Then, the model accurate results would also validate that assuming a Poisson process for COR arrivals has a negligible impact on results.

The mobility configuration that produce larger relative errors for the $P_m^B$ and $P^{ft}$ is $(\mu_m^r, \mu_s^r) = (2, 1)$, approximately. We refer to this configuration as the *worse mobility configuration* (WMC). We now explore the behavior of the proposed approximation in scenarios with the WMC and larger loads.

For Tables 3, 4, and 5, the MC and SC $s$ new call arrival rates at the WMC are $\widehat{\lambda}_m^n = \{0.1, 0.1, 1.0\}$ and $\widehat{\lambda}_s^n = \{1.90, 2.20, 0.95\}$ calls per time unit, respectively. For each reference system and pair of arrival rates at the WMC, we now increase $\widehat{\lambda}_m^n$ while maintaining constant $\widehat{\lambda}_s^n$, and vice versa. Figure 8 shows the evolutions of $P_m^B$, $P_s^B$ and $P^{ft}$ with the arrival rates. Note that figures a) and b) correspond to the SS, c) and d) to the MS, and e) and f) to the LS. In Fig. 8, simulation results are shown by black continuous, discontinuous and dotted lines, and black markers.

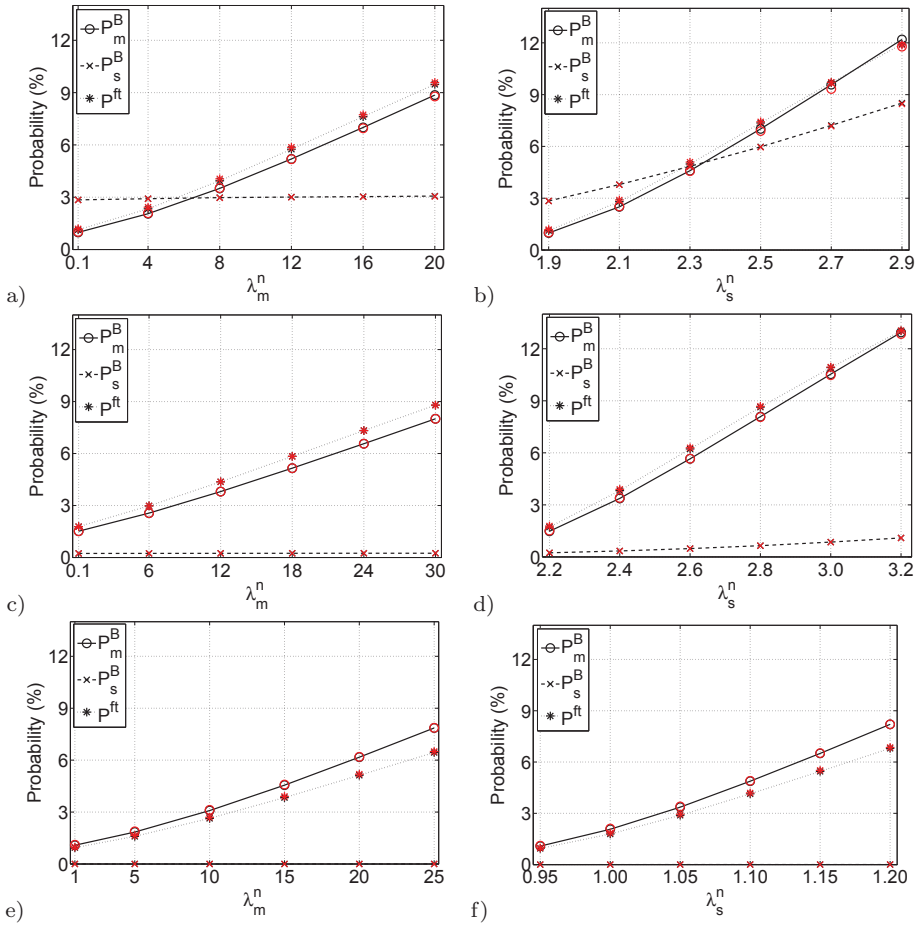


**Fig. 8** Evolution of $P_m^B$, $P_s^B$ and $P^{ft}$ with $\lambda_m^n$ and $\lambda_s^n$ in the LS, MS and LS at the WMC. Analytical results in black, and simulations in red.

That is, $P_m^B$ is shown by a continuous line, $P_s^B$ is shown by a discontinuous line, and $P^{ft}$ is shown by a dotted line. In the same figure, simulation results are shown only by red markers with no lines. Note that 95% confidence intervals have not been shown, as they are negligible.

The relative errors obtained are all smaller than 4%, except for the SS at $\widehat{\lambda}_m^n = 0.1$ and $\widehat{\lambda}_s^n = 1.90$, where it is 5%. We believe that the proposed approximate model provides sufficient accuracy in the majority of realistic scenarios.

### 7.1 Carried traffic

As an example of the versatility of the proposed model, we now determine the traffic carried by the system due to successfully completed calls. We compare the traffic carried by the system in two different network scenarios, one with SCs deployed (WSC) and one with no SCs deployed (NSC). Here by system we mean

a MC overlaying SCs, or just the MC. To make a fair comparison, we maintain the same mobility pattern in WSC and NSC, and in addition we use an equivalent load.

From the mobility perspective, we use the same model to obtain the carried traffic at both scenarios, except that in NSC we set $C_s = 0$ for all SCs. The difference being that when a terminal with an ongoing call serviced by the MC goes indoors, in NSC it always maintains the same MC channel, while in WSC it will try to attach to the SAP, if possible. By going indoors we mean that the mobility rate changes when a terminal enters the coverage area of a SC, and this happens at both, WSC and NSC. Also, in NSC, new calls generated (received) indoors are always serviced by the MC.

From the load perspective, the equivalence is defined in terms of the global QoS perceived by users of both network scenarios. The load at WSC and NSC will be equivalent when they lead to an equal success probability $P^{sc}$, defined as

$$P^{sc} = \left(1 - P^{ir}\right)\left(1 - P^{ft}\right), \tag{27}$$

where $P^{ft}$ is the global forced termination probability defined in (26), $P^{ir} = P_m^{ir} + \sum_{s\in\mathcal{S}} P_s^{ir}$ is the global initial rejection probability, $P_m^{ir} = (\lambda_m^n/\lambda_t^n)\,\beta_{mL} = (\lambda_m^n/\lambda_t^n)\,P_m^B$ , $P_s^{ir} = (\lambda_s^n/\lambda_t^n)\,\beta_{sL} = (\lambda_s^n/\lambda_t^n)\,P_s^L$ , and $\lambda_t^n = \lambda_m^n + \sum_{s\in\mathcal{S}} \lambda_s^n$ . Please refer to Fig. 7. Clearly, $P^{sc}$ is the fraction of calls offered to the system that are accepted and terminate successfully, i.e., are not forced to terminate.

The number of MC channels in NSC is $C_m + kC_s$ , as we assume that in WSC the spectrum is partitioned into two sets of channels: MC channels ($C_m$) and SC channels ($kC_s$). We set $k = 3$. Let $A_m^C$ , $A_s^C$ , $A_s^{C*}$, and $A_C$ be the traffic carried due to calls initiated at the MC, at SC $s$, with either a SC or a MC channel, and the total traffic carried by the system, respectively. Then,

$$\begin{aligned}
A_m^C &= \lambda_m^n \left(1 - P_m^B\right)\left(1 - P_m^{ft}\right)/\mu^d\,,\\
A_s^C &= \lambda_s^n \left(1 - P_s^B\right)\left(1 - P_s^{ft}\right)/\mu^d\,,\\
A_s^{C*} &= \lambda_s^n \left(P_s^B - P_s^L\right)\left(1 - P_s^{ft*}\right)/\mu^d\,,\\
A_C &= A_m^C + F\left(A_s^C + A_s^{C*}\right).
\end{aligned}$$

For the comparative evaluation, we define two extreme mobility scenarios: i) high mobility at MC and low at SCs (HM-LS), and ii) low mobility at MC and high at SCs (LM-HS). The mobility rates for HM-LS and LM-HS are: $(\mu_m^r\,, \mu_s^r) = \{(4, 0.1)\,, (1, 1)\}$, respectively. These mobility rate pairs have been chosen to correspond to entries in Tables 3, 4 and 5. For simplicity, at WSC and for each of the reference systems, we reuse the loads defined to obtain the results displayed at the corresponding entries of the tables.

To obtain the equivalent load at NSC, we define the new call arrival rate to the MC and SC $s$ as $\delta\Lambda_m^n$ and $\delta\Lambda_s^n$, respectively, where $\Lambda_m^n$ and $\Lambda_s^n$ are the arrival rates used for the same mobility configuration at WSC. Then, we adjust $\delta$ to obtain for NSC the same $P^{sc}$ obtained for WSC. Note that in the studied scenarios $0.98 \leq P^{sc} \leq 0.99$, except for SS with HM-LS where it drops to 0.96.

Results in Table 6 show that $A_C$ is larger in the WSC scenario for the mobility configuration HM-LS. This was somehow expected, as in the WSC scenario the

**Table 6** Carried traffic $A_C$ (Erlangs).

| | SS | | MS | | LS | |
|---|---|---|---|---|---|---|
| | WSC | NSC | WSC | NSC | WSC | NSC |
| HM-LS | 101.2 | 50.6 | 320.4 | 99.8 | 536.5 | 98.2 |
| LM-HS | 84.7 | 47.8 | 196.7 | 103.0 | 189.1 | 99.8 |

frequency reuse is much larger and more calls can be carried by the system simultaneously. In addition, in the mobility configuration HM-LS, calls are connected to the MC less time (on average) which leads to more free channels available at the MC (on average). Also, calls are connected to the SCs longer time (on average) which leads to a less frequent generation of changes of residence (COR) towards the MC, to a decrease in the COR failure probability, and to an increase of the call success termination probability.

An interesting observation is that at WSC with LM-HS, $A_C$ is larger for MS than for LS. This might be due to the fact that both reference configurations have equal number of MC channels, however the number of SCs in LS is larger than in MS. Then, when performing SC to MC CORs, a larger fraction of calls are being forced to terminate at LS than at MS. As pointed out before, this penalty at LS might be reduced by adding more channels to the MC or deploying admission control at the MC.

Finally, note the in the NSC scenario, $A_C$ is quite close for MS and for LS. This is due to the fact that in both reference systems the number of MC channels is the same $C_m + 3C_s$, as shown in Table 2.

## 8 Conclusions

A traffic model is proposed to analyze two-tier cellular networks were MCs overlay hundreds of small cells. For the channel assignment, the small cell priority scheme is considered, combined with the system support for *directed retries* and *directed handovers*. Performance parameters such as the blocking probabilities, probability of forced termination of accepted calls, and carried traffic are derived.

To handle the huge dimensionality of CTMC that describes the system behavior, three strategies are combined: i) an equivalent reversible CTMC is constructed for which a product-form stationary distribution exists; ii) the state space is transformed by aggregating states that are not relevant to the performance parameters of interest; and iii) the desired performance parameters are derived by convolution. The results obtained by the proposed approximation are validated by simulation. It is shown that an excellent accuracy is obtained.

The proposed model is used to compare the traffic carried by two different systems, one that deploys small cells and another that does not. In each system, we offer a load that yields an equal probability that a call terminates successfully. Two extreme mobility scenarios are studied, high mobility at macrocells and low mobility at small cells (HM-LS), and the opposite one (LM-HS). As expected, results show that the traffic carried by systems that deploy small cells is larger than in systems that do not, and this difference increases when more small cells are deployed. The traffic carried by systems with small cells is larger for the mo-

bility scenario HM-LS, than for the scenario LM-HS. The scenario HM-LS might resemble a practical deployment scenario where small cells are disposed indoors. However, the traffic carried by systems without small cells depends on the number of resources allocated to the macrocell base stations, and is less dependent on the mobility scenario.

The proposed model can be extended in different ways. For example different call classes could be defined, each of them with different traffic characteristics. Admission control could be added, mainly at macrocells, to reserve resources for the ongoing calls that move from small cells to a macrocell. These calls are particularly susceptible to suffer from forced termination in the studied scenario.

Also the accuracy of the model should be improved in scenarios where time random variables are not exponentially distributed. For example, it has been shown that the residence time in SCs might be better characterized by distributions with coefficient of variation larger than one, particularly when open access policies are deployed.

**Acknowledgements** Authors would like to thank you the anonymous reviewers for the review comments provided to our work, that have decisively contributed to improve the paper. Most of the contribution of V. Casares-Giner was done while visiting the Huazhong University of Science and Technolgy (HUST), Whuhan, China. This visit was supported by the European Commission, 7FP, S2EuNet project. The authors from the Universitat Politècnica de València are partially supported by the Ministry of Economy and Competitiveness of Spain under grant TIN2013-47272-C2-1-R.

## References

1. ABIresearch, "In-Building Mobile Data Traffic Forecast," ABIreseach, Tech. Rep., 2016.
2. NGMN Alliance, "Recommendations for small cell development and deployment," NGMN Alliance, Tech. Rep., 2015.
3. V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59 –67, Jun. 2008.
4. T. Yamamoto and S. Konishi, "Impact of small cell deployments on mobility performance in LTE-Advanced systems," in *IEEE PIMRC Workshops*, 2013, pp. 189–193.
5. R. Balakrishnan and I. Akyildiz, "Local anchor schemes for seamless and low-cost handover in coordinated small cells," *IEEE Transactions on Mobile Computing*, vol. 15, no. 5, pp. 1182–1196, 2016.
6. T. Zahir, K. Arshad, A. Nakata, and K. Moessner, "Interference management in femtocells," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 293–311, 2013.
7. M. Yassin, M. A. AboulHassan, S. Lahoud, M. Ibrahim, D. Mezher, B. Cousin, and E. A. Sourour, "Survey of ICIC techniques in LTE networks under various mobile environment parameters," *Wireless Networks*, pp. 1–16, 2015.
8. M. Andrews and L. Zhang, "Utility optimization in heterogeneous networks via CSMA-based algorithms," *Wireless Networks*, pp. 1–14, 2015.
9. S. M. A. El-atty and Z. M. Gharsseldien, "Performance analysis of an advanced heterogeneous mobile network architecture with multiple small cell layers," *Wireless Networks*, pp. 1–22, 2016.
10. Q. Huang, Y.-C. Huang, K.-T. Ko, and V. B. Iversen, "Loss performance modeling for hierarchical heterogeneous wireless networks with speed-sensitive call admission control," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2209–2223, 2011.
11. T. Bonald and J. W. Roberts, "Congestion at flow level and the impact of user behaviour," *Computer Networks*, vol. 42, pp. 521–536, 2003.
12. Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent advances in radio resource management for heterogeneous LTE/LTE-A networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2142–2180, 2014.

13. S. S. Rappaport and L.-R. Hu, "Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis," *Proceedings of the IEEE*, vol. 82, no. 9, pp. 1383–1397, 1994.
14. X. Ge, T. Han, Y. Zhang, G. Mao, C.-X. Wang, J. Zhang, B. Yang, and S. Pan, "Spectrum and energy efficiency evaluation of two-tier femtocell networks with partially open channels," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1306–1319, 2014.
15. W. Song, H. Jiang, and W. Zhuang, "Performance analysis of the WLAN-first scheme in cellular/WLAN interworking," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1932–1952, 2007.
16. X. Ge, J. Martinez-Bauset, V. Gasares-Giner, B. Yang, J. Ye, and M. Chen, "Modeling and performance analysis of different access schemes in two-tier wireless networks," in *IEEE Globecom*, 2013, pp. 4402–4407.
17. H.-M. Tsai, A.-C. Pang, Y.-C. Lin, and Y.-B. Lin, "Repacking on demand for hierarchical cellular networks," *Wireless Networks*, vol. 11, no. 6, pp. 719–728, 2005.
18. K. Maheshwari and A. Kumar, "Performance analysis of microcellization for supporting two mobility classes in cellular wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 2, pp. 321–333, Mar. 2000.
19. P. Whiting and D. McMillan, "Modeling for repacking in cellular radio," in *7th UK Teletraffic Symp, IEE*, Durham, 1990.
20. F. Kelly, "Fixed point models of loss networks," *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, vol. 31, no. 02, pp. 204–218, 1989.
21. D. McMillan, "Traffic modelling and analysis for cellular mobile networks," in *Proceedigs of ITC-13*, A. Jensen and V. Iversen, Eds., IAC. Copenhaguen: Elsevier Science, Jun. 1991, pp. 627–632.
22. H.-L. Fu, P. Lin, and Y.-B. Lin, "Reducing signaling overhead for femtocell/macrocell networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 8, pp. 1587–1597, Jun. 2012.
23. B. Eklundh, "Channel utilization and blocking probability in a cellular mobile telephone system with directed retry," *IEEE Transactions on Communications*, vol. 37, pp. 329–337, Apr. 1986.
24. J. Karlsson and B. Eklundh, "A cellular telephone system with load sharing – An enhancement of directed retry," *IEEE Transactions on Communications*, vol. 37, no. 5, pp. 530–535, May 1989.
25. R. Nelson, *Probability, Stochastic Processes and Queueing Theory*. Springer-Verlag, 1995.
26. V. B. Iversen, "The exact evaluation of multi-service loss systems with access control," in *Proc. of the Seventh Nordic Teletraffic Seminar (NTS-7)*, vol. 31, Lund, (Sweden), Aug. 1987, pp. 56–61.
27. K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer Verlag, 1995.
28. Y.-B. Lin and V. W. Mak, "Eliminating the boundary effect of a large-scale personal communication service network simulation," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 4, no. 2, pp. 165–190, Apr. 1994.
29. M. K. Karray, "Evaluation of the blocking probability and the throughput in the uplink of wireless cellular networks," in *IEEE ComNet*, 2010, pp. 1–8.