Abstract—The sustained increase of users and the request for advanced multimedia services are some of the key motivations for designing new high capacity cellular telecommunication systems. The proposals which are being pursued by several studies and field implementations consider hierarchical architectures and dynamic resource allocation. In this paper a hierarchical cellular communication network is analyzed, taking into account user mobility and exploiting dynamic channel allocation schemes. In particular, a finite number of users has been considered, moving at different speeds in a geographical region covered by a finite number of cells structured in two hierarchical levels: microcells and macrocells. For such a system, a mobility model and a traffic model have been developed, both based on queueing networks, analyzing Maximum Packin (MP), a dynamic channel allocation scheme. The obtained results, validated by simulation experiments, allow the evaluation of main system performance parameters in terms of new call blocking probability, handoff blocking probability and forced termination probability as a function of load and system parameters.

Index Terms—Cellular Networks, Queueing Model, User Mobility, Hierarchical Network.

I. INTRODUCTION

CELLULAR telephony is one of the consumer electronics fields that has experienced an explosive development in recent years. In fact, the sustained increase of users together with the request for advanced multimedia services are driving the research for the development of second generation systems (for instance, see the GSM phase 2+ [1]) and the definition of new standards for third generation systems (e.g., UMTS [2]). The transmission frequencies of evolved second generation and third generation systems are about 2 GHz, with a microcell coverage, while the second generation systems (for example, the first version of GSM) use frequencies of about 900 MHz with a macrocell coverage.

It should be pointed out that a coverage based only on microcells is not always desirable. In fact, in this kind of cell, the propagation conditions are highly dependent on the environment: width of the streets, moving obstacles and so on. Therefore, when, as an example, users turn a street corner, they may experience a dramatic decrease in signal power. Such rapid variations of the received power level may cause communication interruptions for high-speed users if the network does not have enough time for the handoff process (assuming that there are available resources for this operation). Moreover, such high-speed users may generate lots of handoffs causing a significant increase in signaling traffic in the network. In fact, each handoff process involves various control functions both on wired and wireless networks. Such control procedures, necessary to set up a connection with the new base station and to update state information, introduce delay in the transmission, which can cause catastrophic consequences for multimedia real time traffic. Therefore, it is necessary to preserve the macrocell coverage that may serve high-speed users, decreasing their handoff rate. Moreover, the macrocells may provide coverage in shadow areas and in less dense areas where microcell implementation is not convenient. This leads to the concept of hierarchical or multilayer cellular structure [3]-[12].

The study presented in this paper can be thought of as an attempt at integration, by means of a hierarchical structure, of systems based on microcells and macrocells. In the analyzed cellular systems two classes of users are considered, the slow and the fast ones, normally served by microcell and macrocell level, respectively. In the case of a lack of resources at microcell level, the communication of a slow user is transferred from the microcell level to the macrocell one (overflow). Moreover, the structure is reversible [5], [6]; i.e., an overflowed slow user communication has to go back to microcell level (take-back), when in this level there are available resources. The introduction of reversibility in non-reversible hierarchical systems produces an increase in traffic capacity, with a fixed quality of service (QoS), or an improvement of QoS, with the call load fixed. The cost of this performance improvement is greater system complexity, as the system has to monitor the eventual availability of resources in microcell level. Moreover, the number of call transfers increases, due to the addition of the take-back operations.

In both hierarchical levels we have chosen a dynamic resources allocation (DRA) [13], [14]. This is the trend in new cellular
systems, because with DRA the network is able to adapt itself easily to traffic variations and it allows network designers to add or eliminate cells without a new frequency planning. For instance, particular DRA schemes have been proposed for TD-CDMA UMTS [2], [15], [16].

The system architecture synthetically defined above differentiates explicitly our approach from previous studies on this topic that consider either Fixed Channel Assignment (FCA) [5]-[11], or linear cellular structure [12].

The rest of the paper is organized as follows. In section II system architecture and multiclass mobility modeling are reported. The cell modeling is reported in section III, while the performance parameters have been evaluated in section IV. In section V a numerical example and the related results are reported, section VI concludes the paper.

II. SYSTEM ARCHITECTURE AND MOBILITY MODELING

A. Description of the Hierarchical Structure

A two level hierarchical architecture (microcells and macrocells) has been considered. The microcells have a hexagonal shape and each macrocell covers \( L \) microcells. Therefore, if \( M \) is the number of microcells, the number of macrocells is \( M/L \). The radio resource management criterion exploits the orthogonal sharing, i.e., different parts of the spectrum are assigned to the two hierarchical levels, maximizing, in this way, network capacity [3], [4] and confining the interference problems only to cells of the same level. It is supposed that there is appreciable interference only between adjacent cells. Thus, a cluster (i.e., a set of neighbor cells where it is necessary to use different channels to avoid interference [17]) is composed by three cells. In the network there are \( N \) mobile users divided into two mobility classes: slow and fast users. The new service requests of slow and fast users are assigned to microcells and macrocells, respectively. In the case of resource unavailability at microcell level, slow users can exploit available resources at macrocell level, i.e., there is an overflow. Such slow users, who are being served by a macrocell, try to return to microcell level as soon as they transit the border of a microcell, i.e., there is a take-back. Obviously, the take back is accepted when at microcell level there are available resources; otherwise, the user continues to exploit the macrocell resources. Users can change their mobility class, but, for sake of simplicity, this possibility is not allowed during a call.

In the system there are globally \( C \) available channels: \( C_1 \) for the microcell level and \( C_2 \) for the macrocell one. The term “channel”, refers to a generic logical resource used for communication (e.g., a logical channel could be a double physical channel for communication in the up-link and down-link directions with a generic access method: FDMA, TDMA, CDMA, etc.)

B. Resources Allocation Algorithm

The DRA scheme adopted in each level of the network is the Maximum Packing (MP) [18]-[25]. Such an algorithm is characterized by an indefinite number of possible channel rearrangements in the cells of each level. The generic cell can serve a call request if the number of channels, already occupied in each cluster the cell belongs to, is less than the whole number of available channels.

For each of the two levels, let \( u = (u(1), u(2), \ldots, u(A)) \) be the vector of the active users in the cells, where \( A = M \) for the microcell level and \( A = M/L \) for the macrocell level. A call can be accepted in a generic cell \( i \), according to MP algorithm, if the following expression is satisfied [20]:

\[
u(i) + u(l) + u(j) \leq C_i - 1 \quad \forall \ (t,j) \not\in S_i \tag{1}\]

where \( t \) and \( j \) are two other cells which belong to the same cluster of cell \( i \). \( S_i \) represents the set of clusters the cell \( i \) belongs to. \( C_i \) is the total number of available channels at the considered level (\( k = 1 \), microcells; \( k = 2 \), macrocells). In other words, the condition of the MP algorithm is verified when in each group of three cells, with a common vertex, the active channels are less than \( C_1 \) for microcells and less than \( C_2 \) for macrocells. It is to be noted, however, that as discussed in [21], [25], in some cases the state space determined by (1) is larger than the state space determined by MP, but the difference in terms of performance can be disregarded.

This MP algorithm, although not always completely feasible, is attractive because, apart from being an analytically tractable strategy, it is in some sense optimal. In fact, a new call is blocked only when the number of channels already in use by previous calls plus the number of channels requested to accommodate the new call is larger than the number of available channels. Thus, this algorithm provides an upper limit on system performance of re-use constraint allocation strategies. Moreover, MP represents a reasonable approximation of how well a dynamic channel assignment algorithm with power control performs.

C. Multi-class Mobility Modeling

As regards mobility, a generalization of the mobility model described in [26] has been adopted. It is supposed that a user can pass from microcell \( j \) at class \( r \) to microcell \( j \) at class \( h \) with a transition probability \( p_{rj,h} \). Such a probability can be evaluated, for example, by experimental measures on real systems. It should be noted that a change of mobility class, remaining in the same cell, is allowed by considering \( i = j \) and \( r = h \). In the plausible hypothesis that there are no constraints on the number of users that can sojourn contemporarily in a generic cell, each microcell can be modeled by an infinite server queue, i.e., for each user in the
cell there is a virtual server to consider its sojourn time. As previously mentioned, the considered users belong to two communicating mobility classes. The sojourn time is a random variable with mean $1/\mu_c$ (r = 1, 2) and an arbitrary p.d.f. This formalization allows us to model user mobility, at microcell level, with a closed single chain multi-class queuing network, which admits a product form solution [27], [28].

The microcell transition rate $\mu_c$, measured in transitions per hour [tr/h], represents user mobility and is related to the diameter $d_i$ of a microcell and with the average speed, $v_c$, of users

$$\mu_c \propto v_c / d_i \quad \text{with} \quad r = 1, 2. \quad (2)$$

It can be noticed that, keeping user speed constant, microcell size reduction can be regarded as an increase in user mobility. This allows the extension of model results from pico (few meters) to macro (some kilometers) cellular networks.

Solving the equations of the mobility model (see Appendix A), the average number of users for each class in a generic microcell, $E[n_r]$, is given by

$$E[n_r] = N(e_{r} / \mu_{r}) / \sum_{i=1}^{M} \sum_{r=1}^{2} \left( e_{r} / \mu_{r} \right); i = 1, ..., M; r = 1,2. \quad (3)$$

where $e_{r}$ is the relative visit rate to each microcell also obtained by the mobility model. It can be noted that the average number of users, $E[n_3]$ (fundamental for the model described in the sequel) is directly proportional to $N$. Therefore, changing only the total number of users of the network, $E[n_3]$ can be computed by (3), without re-evaluating all expressions of the mobility model.

### III. MICROCELL AND MACROCELL MODELING

#### A. Service Description in the Two Levels

Each free user of the two classes generates a new call following an exponential distribution with parameter $\lambda$ calls/h. For each class, the call duration, $t_c$, is a general distributed random variable with mean $1/\mu_c$. According to the system description, microcells only attend to slow users. Each new service request is accepted if there is a free channel in the microcell according to the condition (1). Otherwise, there is an overflow, the request is rejected at this level and passed to the umbrella-macrocell, i.e., the macrocell that covers the considered microcell.

As regards the service offered to slow users, i.e., the possibility of making a call, a microcell can be represented by a finite population M/G/m/m/n queue [29]. It is thus considered a loss system where the maximum number of contemporary calls in progress is given by the number of available channels in the microcell. For this queue, the steady state probabilities for the number of calls in progress are given by [30]

$$p(u = k) = \binom{n}{k} \rho^k \sum_{i=k}^{n} \binom{n}{i} \rho^i \quad 0 \leq k \leq m \leq n \quad (4)$$

where $\rho$ is the load factor, defined as the ratio between call rate and service rate.

A macrocell, on the other hand, can attend to either slow or fast users. Thus, the macrocell can be represented by a finite population bidimensional queuing system, i.e., a queuing system with two classes of users. Generalizing the previous monodimensional queue to the bidimensional case, for this system the steady state probabilities for the number of calls in progress of the two classes are given by [31]

$$p(u_1 = k, u_2 = l) = \frac{\binom{n_1}{k} \rho_1^k \binom{n_2}{l} \rho_2^l}{\sum_{i,j} \binom{n_1}{i} \rho_1^i \binom{n_2}{j} \rho_2^j} \quad 0 \leq k + l \leq m \quad (5)$$

The classification of users as fast and slow can be done by exploiting the average number of microcell handoffs experienced during a call ($n_{H}$). In particular, $n_{H} = \mu_c E[t_c] = \mu_c / \mu_c$. The user is classified as slow and fast respectively when $n_{H} \leq \chi$ and $n_{H} > \chi$. The threshold $\chi$ must be chosen specifically for each system, discriminating, as an example, pedestrians from cars.

#### B. Microcell Model Characterization

At this point, the terms in the expressions (4) and (5) have to be characterized. The number of slow users in a generic microcell $z$ is fixed to the average value obtained through the study of mobility and renamed $n_z(z)$ for simplicity, where $r$ refers to first level (i.e., the microcell level) and $s$ means “slow”. Although the users move among the microcells and/or the two classes, all performance indices evaluated in the sequel are obtained considering only the average number of users per cell and per class. The former might seem to be too much approximated, but it leads to good results, as will be shown later, when comparing the analytical results with simulation which is not based on this approximation.
For every microcell, the total number of channels \( C_1 \) is available, but, if \( n_{1s}(z) < C_1 \), the maximum value of the active channels is just \( n_{1s}(z) \). Thus, in (4), the term \( m \) denotes the minimum value between \( C_1 \) and \( n_{1s}(z) \).

The evaluation of the load factor must take into account the occurrences of handoffs between microcells and the occurrences of take-backs, i.e., the returns of slow users with a call in progress from macrocell to microcell level. If there is a handoff when an active slow user, that is a user with a call in progress, passes from one microcell to another one, since the mean number of slow users per microcell does not change, we can consider the effect of handoffs as an increase in the request rate per user. A take-back request can be regarded as the handoff request of a slow user that is served by macrocell level. Thus, similarly to handoffs, we can consider the effect of take-backs as a further increase in the request rate per user. Therefore, if \( \lambda_{H1s}(z) \) is the additional service request rate due to handoffs and \( \lambda_{TB}(z) \) is the additional service request rate due to take-backs, the total request rate in a generic microcell \( z \), \( \lambda(z) \), is given by [5], [14]

\[
\lambda(z) = \lambda + \lambda_{H1s}(z) + \lambda_{TB}(z).
\]

The total service rate is the sum of the call termination rate, \( \mu_s \), and of the handoff attempt rate [32]. From the mobility model and detailing the notation for the two classes of our system, the average sojourn time for each slow user in a microcell is \( \frac{1}{\mu_s} \), and noting for the two classes of our system, the average sojourn time for each slow user in a microcell is \( \frac{1}{\mu_s} \), and of the handoff attempt rate is \( \mu_t \), and we have

\[
\mu_t = \mu_s + \mu_n.
\]

Moreover, in the evaluation of load factor, we have to take into account the blocking probability of a new call in microcell \( P_{ni}(z) \), due to the constraints imposed by expression (1). In fact, in the generic microcell \( z \), a call request might be rejected and passed to the umbrella-macrocell if (1) is not satisfied, even if the number of active calls in the microcell \( z \) is zero. We then have the following expression

\[
\rho(z) = (1 - P_{ni}(z))(\lambda + \lambda_{H1s}(z) + \lambda_{TB}(z))/\mu_c + \mu_n).
\]

The term \( \rho(z) \) will be calculated later because it depends on the state of macrocell. To evaluate the handoff rate \( \lambda_{H1s}(z) \) which is in (8), let \( \mu_{1s}(z) \) be the average number of active slow users served by the generic microcell \( z \)

\[
\mu_{1s}(z) = \min(C_1, n_{1s}(z)), \quad (9)
\]

where \( p(u_{1s}(z) = k) \) is obtained from (4) with the load factor \( \rho(z) \) of the expression (8).

Thus, the average number of slow users that attempt handoff, moving out of the generic microcell \( a \), is \( u_{1s}(a) \). Let \( p_{1s}(a, z) \) be the transition probability of slow users from microcell \( a \) to microcell \( z \). It is simply obtained by transition probabilities \( p_{1s}(a, z) \), introduced in the mobility model, imposing the condition that there are no transitions between mobility classes, as supposed in section II-A. Therefore \( p_{1s}(a, z) \) is equal to \( p_{1s}(a, z) \) (where \( a \) always means slow), normalized so that \( \sum_{z} p_{1s}(a, z) = 1 \). The total rate of handoffs entering the microcell \( z \) is the sum, over all \( M \) microcells, of each mean rate of attempted handoffs from microcell \( a \) to microcell \( z \),

\[
\Lambda_{H1s}(z) = \sum_{a=1}^{M} u_{1s}(a) \cdot \mu_{1s} \cdot p_{1s}(a, z).
\]

To obtain \( \lambda_{H1s}(z) \) it is necessary to normalize \( \Lambda_{H1s}(z) \) to the number of slow users in the cell \( z \), because \( \lambda_{H1s}(z) \) is the call rate per user

\[
\lambda_{H1s}(z) = \frac{\Lambda_{H1s}(z)}{n_{1s}(z)} = \frac{\sum_{a=1}^{M} u_{1s}(a) \cdot \mu_{1s} \cdot p_{1s}(a, z)}{n_{1s}(z)}.
\]

C. Macrocell Model Characterization

Considerations similar to those derived for microcells can be developed at macrocell level. In particular, \( n_1 \) and \( n_2 \) in equation (5), are respectively the number of slow and fast users in the generic macrocell \( w \). These values are assumed equal to the average number of slow and fast users per macrocell, obtained by adding the average number of slow and fast users in each microcell covered by the considered macrocell, and renamed for simplicity \( n_{1s}(w) \) and \( n_{2s}(w) \).

For each macrocell the total number of channels \( C_2 \) is available, but, if \( n_{1s}(w) + n_{2s}(w) < C_2 \), the maximum value of the active channels is just \( n_{1s}(w) + n_{2s}(w) \). Thus, in (5), the term \( m \) denotes the minimum value between \( C_2 \) and \( n_{1s}(w) + n_{2s}(w) \). Similar considerations are valid for the users of each mobility class. Therefore, we have

\[
u_{1s}(w) \leq \min(C_2, n_{1s}(w)) \quad \mu_{TB}(w) \leq \min(C_2, n_{2s}(w))
\]
where \( u_{2s}(w) \) ( \( u_{2f}(w) \)) is the number of active slow (fast) users served by macrocell \( w \).

At this point, it is necessary to characterize the two load factors \( \rho_2 \) and \( \rho_2 \) in (5). At this level there are overflows, i.e., the transfer of calls of slow users from microcell to macrocell level for lack of resources in the former. The total service request rate for slow users is the sum of the overflow request rate to the considered macrocell \( w \), \( \lambda_{2s}(w) \), and the handoff request rate, \( \lambda_{2z}(w) \). Obviously, index 2 refers to the second level, i.e., the macrocell level. The total service rate is the sum of call termination rate, \( \mu_c \), handoff attempt rate, that is equal to the transition rate of slow users between macrocells \( \mu_m \), and take-back requests rate, \( \mu_{2b}(w) \), calculated later. Taking into account the blocking probability of a new call in macrocell \( w \), \( P_{B2}(w) \), due to the constraints imposed by expression (1), the load factor for slow users is

\[
\rho_{2s}(w) = \frac{(1 - P_{B2}(w))(\lambda_{2s}(w) + \lambda_{2z}(w))}{(\mu_c + \mu_m + \mu_{2b}(w))}.
\]  

(12)

In a similar way, the load factor for fast users, who can be served only by macrocell level, is

\[
\rho_{2f}(w) = \frac{(1 - P_{B2}(w))(\bar{\lambda} + \lambda_{2z}(w))}{(\mu_c + \mu_m)}.
\]  

(13)

where \( \lambda_{2z}(w) \) is the handoff request rate to macrocell \( w \) per fast user.

We have to characterize the new rates that we have just introduced. Let \( \overline{u_{2s}(w)} \) be the average number of active slow users served by a generic macrocell \( w \)

\[
\overline{u_{2s}(w)} = \frac{\min_{C_s} u_{2s}(w) \cdot \sum_{i=0}^{\min_{C_s} u_{2s}(w)} \cdot \sum_{j=0}^{\min_{C_s} u_{2s}(w)} \cdot i \cdot p(u_{2s}(w) = i, u_{2f}(w) = j)}{\sum_{i=0}^{\min_{C_s} u_{2s}(w)} \cdot \sum_{j=0}^{\min_{C_s} u_{2s}(w)} \cdot i \cdot j \cdot p(u_{2s}(w) = i, u_{2f}(w) = j)}
\]  

(14)

where \( p(u_{2s}(w) = i, u_{2f}(w) = j) \) is calculated by expression (5), substituting the factors (12) and (13). Let \( p_{2s}(a, w) \) be the transition probability of slow users from the generic macrocell \( a \) to macrocell \( w \). This is obtained by adding up all the probabilities \( p_{2s}(i, j) \), in which microcells \( i \) and \( j \) are adjacent and covered by macrocell \( a \) and \( w \), respectively; therefore it is necessary to normalize so that \( \sum_{a=1}^{M/L} p_{2s}(a, w) = 1 \).

The mean rate of active slow users, served by macrocell level, that pass from macrocell \( a \) to macrocell \( w \) is \( \overline{u_{2s}(a)} \cdot \mu_{2s} \cdot p_{2s}(a, w) \). Naturally, only the slow users that cannot do the take-back ask macrocell \( w \) for a handoff. For a generic microcell \( z \) there is no difference between a new call request and a take-back request, so that the blocking probability of a take-back is equal to the blocking probability of a new call \( P_{B2}(z) \). But, according to the used model, it is impossible to know exactly which microcell, covered by macrocell \( w \) and adjacent to macrocell \( a \), denies the take-back. Therefore we can put this blocking probability of take-back equal to the following average value

\[
P_{B2}(a, w) = \frac{\sum_z n_s(z) \cdot P_{B2}(z)}{\sum_z n_s(z)}
\]  

(15)

where microcell \( z \) is adjacent to macrocell \( a \). Therefore, the mean rate of attempted handoffs of active slow users from macrocell \( a \) to macrocell \( w \) is

\[
\Lambda_{2z}(a, w) = \overline{u_{2s}(a)} \cdot \mu_{2s} \cdot p_{2s}(a, w) \cdot P_{B2}(a, w).
\]  

(16)

The total handoff rate attempted toward a generic macrocell \( w \) is the sum of each \( \Lambda_{2z}(a, w) \), over all the \( M/L \) macrocells

\[
\Lambda_{2z}(w) = \sum_{a=1}^{M/L} \Lambda_{2z}(a, w).
\]  

(17)

Normalizing to the number of slow users sojourning in macrocell \( w \), the handoff request rate to macrocell \( w \) per slow user is

\[
\lambda_{2z}(w) = \frac{\Lambda_{2z}(w)}{n_s(w)}.
\]  

(18)

Instead, the overflow request rate to macrocell \( w \) per slow user \( \lambda_d(w) \) is due to two contributions, the new calls and the handoffs that are blocked in the microcells covered by \( w \). Numbering the microcells covered by \( w \) from \( 1_w \) to \( L_w \), we can write
in which \( P_{B1}(z) \) is the handoff blocking probability in microcell \( z \) for a slow user, calculated later.

The take-back request rate, \( \mu_{TB}(w) \), is the product of the transition rate between microcells \( \mu_i \) and the probability that the take-back request of a slow user served by macrocell \( w \) is accepted. This probability can be approximated by \( 1 - P_{B1}(w) \), where \( P_{B1}(w) \) is the following average value

\[
P_{B1}(w) = \frac{\sum_{z \in C} n_i(z) P_{B1}(z)}{\sum_{z \in C} n_i(z)}
\]

we then have

\[
\mu_{TB}(w) = \mu_i \cdot (1 - P_{B1}(w)).
\]

The handoff request rate to macrocell \( w \) per fast user \( \lambda_{H2}(w) \) is calculated similarly to \( \lambda_{H1}(z) \). If the average number of active fast users is

\[
\lambda_{H2}(w) = \sum_{z \in C} n_i(z) \frac{P_{B1}(z)}{n_i(z)}
\]

we then have

\[
\lambda_{TB}(z) = \lambda_{H2}(w) = \frac{\mu_i \cdot (1 - P_{B1}(w))}{n_i(z)}.
\]

IV. PERFORMANCE PARAMETERS

A. Blocking Probability of a New Call

An important performance parameter is the blocking probability of a new call for each mobility class. In general the event \{new call blocked\} occurs when a user attempts a new call but there are no available channels, according to (1).

For the fast users, who can be served only at macrocell level, this probability is

\[
P_{Bf} = \frac{P_{B2}}{n_i(z)}
\]

where \( P_{B2} \) is a weighted mean of blocking probabilities of new calls in the different macrocells \( P_{B2}(w) \), the weights being the
ratios between the number of users in each macrocell and the number of users in the whole network

$$P_{B_2} = \frac{\sum_{z=1}^{M} (n_z(w)+n_z(w))}{N}. \quad (27)$$

On the other hand, for the slow users, who can be served by microcell and macrocell levels, the blocking probability of new calls is

$$P_{B_1} = P_{B_1} \cdot P_{B_2} \quad (28)$$

where, like $P_{B_2}$, $P_{B_1}$ is

$$P_{B_1} = \sum_{z=1}^{M} n_{1s}(z) \cdot P_{B_1}(z) / \sum_{z=1}^{M} n_{1s}(z). \quad (29)$$

We have to calculate $P_{B_1}(z)$ and $P_{B_2}(w)$. The assumption used to make the problem analytically tractable is the independence of the cells in each level [22]; that is, the number of calls in progress in each cell are assumed independent. This is obviously an approximation of the model, given that, for instance, in each level the handoff rate to cell $z$ depends on the number of active calls in the other cells of the same level. However, such an approximation does not invalidate the results. In fact, as will be shown later, there is a good agreement between analytical and simulation results.

\[ \text{Fig. 1. Reference microcells for probabilities evaluation.} \]

From the previous hypothesis, if $u_1 = (u_{1s}(1), \ldots, u_{1s}(M))$ is the state vector which gives the number of active users (or active calls) in the microcell level, the system state probability $p(u_1)$ is

$$p(u_1) = p(u_{1s}(1), \ldots, u_{1s}(M)) \equiv p(u_{1s}(1)) \ldots p(u_{1s}(M)) \quad (30)$$

where $p(u_{1s}(z))$ is obtained by (4).

To evaluate $P_{B_1}(z)$, the blocking probability of new calls in microcell $z$, let us consider, for sake of simplicity, the non-blocking probability $P_{NB1}(z)$, related to $P_{B1}(z)$ by the following expression

$$P_{B1}(z) = 1 - P_{NB1}(z). \quad (31)$$

Under MP assumptions, $P_{NB1}(z)$ can be expressed as follows

$$P_{NB1}(z) = p\{u(z) + u(t) + u(j) < C\} \quad \forall (t,j) \not\in S_z \quad (32)$$

Referring to Fig. 1.(a), in which the microcells surrounding microcell $z$ are numbered as 1, 2, ..., 6, the expression (32) can be rewritten as

$$p(\{(u_1(z)+u_{1s}(1)+u_{1s}(2)<C_1) \cap (u_1(z)+u_{1s}(3)+u_{1s}(4)<C_1) \cap \ldots \cap (u_1(z)+u_{1s}(5)+u_{1s}(6)<C_1)\}). \quad (33)$$

It is possible to compute $P_{NB1}(z)$ by adding up the state probability $p(u_1(z))$ over all the state spaces that verify the MP condition (1)

$$P_{NB1}(z) = \sum_{u(z),u(t),u(j)} \sum_{u_{1s}(6)} p(u_1(z)) p(u_{1s}(1)) \ldots p(u_{1s}(6)) \quad (34)$$

with $u_{1s}(z) + u_{1s}(t) + u_{1s}(j) < C_1 \quad \forall (t,j) \not\in S_z$.

If we consider the condition (1) directly in (34), $P_{NB1}(z)$ is given by [29]

\[ P_{NB1}(z) = \sum_{u(z),u(t),u(j)=0}^{\min(u_{1s}(6),C_1-u_{1s}(1)-u_{1s}(5))} \sum_{u_{1s}(1)=0}^{\min(u_{1s}(1),C_1-u_{1s}(3)-u_{1s}(4))} \sum_{u_{1s}(2)=0}^{\min(u_{1s}(2),C_1-u_{1s}(4)-u_{1s}(5))} \ldots \sum_{u_{1s}(6)=0}^{\min(u_{1s}(6),C_1-u_{1s}(1)-u_{1s}(5))} p(u_1(z)) \cdot p(u_{1s}(1)) \cdot \ldots \cdot p(u_{1s}(6)) \quad (35) \]

with $u_{1s}(z) + u_{1s}(6) + u_{1s}(1) < C_1$. 

Following a similar way for the macrocell level, we can obtain the non-blocking probability of new calls in macrocell \( w \) [33]

\[
P_{NB2}(w) = \min_{u_{c}(w)+u_{j}(w)} \sum_{u_{c}(w)} \min_{u_{c}(w)} \sum_{u_{j}(w)} \sum_{u_{c}(w)} \ldots \sum_{u_{j}(w)} p(u_{c}(w),u_{j}(w)) \ldots p(u_{c}(6),u_{j}(6))
\]

with the constraints

\[
\begin{align*}
|u_{c}(w) + u_{j}(w)| & \leq \min(u_{c}(w) + u_{j}(w), C_{c} - 1) \\
|u_{c}(l) + u_{j}(l)| & \leq \min(u_{c}(l) + u_{j}(l), C_{c} - 1 - u_{c}(w) - u_{j}(w)) \\
& \ldots \\
|u_{c}(6) + u_{j}(6)| & \leq \min(u_{c}(6) + u_{j}(6), C_{c} - 1 - u_{c}(w) - u_{j}(w) - u_{c}(5) - u_{j}(5)) \\
|u_{c}(z) + u_{j}(z) + u_{c}(6) + u_{j}(6) + u_{c}(l) + u_{j}(l)| & < C_{c}.
\end{align*}
\]

(36)

### B. Handoff Blocking Probability

Another important performance parameter is the handoff blocking probability in each hierarchical level, i.e., the probability that a user, with a call in progress, passing from one cell to another, does not find a free channel in the destination cell of his level.

For microcell level the total rate of handoffs accepted in microcell \( z \), \( \Lambda_{41}(z) \), is

\[
\Lambda_{41}(z) = \sum_{a=1}^{M} \Lambda_{H1s}(a,z) \cdot P_{A4is}(a,z)
\]

(37)

where, as discussed in sec. III-B, \( \Lambda_{H1s}(a,z) = u_{a}(a)\mu_{a}p_{a}(a,z) \) and \( P_{A4is}(a,z) \) is the probability that the handoff request from microcell \( a \) to microcell \( z \) is accepted. The ratio between accepted handoff rate and attempted handoff rate, given by (11), is the probability that a handoff is accepted in microcell \( z \), thus the handoff blocking probability in microcell \( z \), \( PBH_{1}(z) \), is

\[
PBH_{1}(z) = 1 - (\Lambda_{41}(z)/\Lambda_{H1s}(z)).
\]

(38)

To evaluate \( P_{A4is}(a,z) \) we can follow a procedure similar to the one used for evaluating the blocking probability of a new call.

Referring to Fig. 1(b), in which microcell \( a \) and \( z \) have neighbors numbered as 1, 2, ..., 5, \( P_{A4is}(a,z) \) is the probability that in the transition instant in which a user, passing from microcell \( a \) to microcell \( z \), finds a free channel not occupied by the other \( N-1 \) users, i.e., in each cluster there are a maximum of \( C-1 \) calls in progress [34]. Thus, \( P_{A4is}(a,z) \) is

\[
P_{A4is}(a,z) = \sum_{u_{c}(z)+u_{j}(z)+u_{c}(6)+u_{j}(6)+u_{c}(l)+u_{j}(l)<C_{c}} p(u_{c}(z)) \cdot p(u_{c}(a)) \ldots p(u_{c}(5))
\]

(39)

with \( u_{c}(z) + u_{c}(a) + u_{c}(5) < C_{c} \).

The handoff blocking probability in microcell level, \( PBH_{1} \), is a weighted mean of the handoff blocking probability of microcells. The weights are the ratios between the rate of rejected handoffs in a microcell and the total rate of rejected handoffs in all microcells. We have

\[
PBH_{1} = \sum_{a=1}^{M} \Lambda_{H1s}(z) \cdot PBH_{1}(z) / \sum_{a=1}^{M} \Lambda_{H1s}(z)
\]

(40)

Following a similar method for the macrocell level, we can write that the total rate \( \Lambda_{42}(w) \) of handoff accepted in macrocell \( w \) is

\[
\Lambda_{42}(w) = \sum_{a=1}^{M/L} [\Lambda_{H2a}(a,w) + \Lambda_{H2f}(a,w)] \cdot p_{A2}(a,w)
\]

(41)

where \( \Lambda_{H2a}(a,w) \) and \( \Lambda_{H2f}(a,w) \) can be evaluated respectively by expressions (16) and (23).

Therefore, the handoff blocking probability in macrocell \( w \), \( PBH_{2}(w) \), is

\[
PBH_{2}(w) = 1 - (\Lambda_{42}(w)/\Lambda_{H2}(w))
\]

(42)
in which

\[ \Lambda_{H2}(w) = \sum_{k=0}^{M/2} [\Lambda_{H2}(a,w) + \Lambda_{H2}(a,w)]. \]  

(43)

Referring to Fig. 1.(c) and proceeding like the microcell level, the probability that a handoff request from macrocell \( a \) to macrocell \( w \) is accepted \( p_{a,w}(a,w) \) is given by [33]

\[ p_{a,w}(a,w) = \prod_{n_{s}(w)=0}^{\min(C_{a}, C_{w})} p(u_{2s}(w), u_{z}(w)) \cdot \ldots \cdot p(u_{2s}(6), u_{z}(6)) \]

(44)

with the constraints

\[
\begin{align*}
    u_{c}(a) + u_{c}(w) & \leq \min(a_{c}(w) + n_{c}(w), C_{a} - 1) \\
    u_{c}(l) + u_{c}(l) & \leq \min(a_{l}(l) + n_{l}(l), C_{l} - 1 - u_{f}(w) - u_{s}(w)) \\
    \vdots \end{align*}
\]

(45)

Therefore the global handoff blocking probability in macrocell level \( P_{BH2} \) is given by

\[ P_{BH2} = \frac{\sum_{w=1}^{M/2} \Lambda_{H2}(w) \cdot P_{BH2}(w)}{\sum_{w=1}^{M/2} \Lambda_{H2}(w)} . \]

(46)

C. Numerical Method

The previous expressions are mutually dependent. In fact, to calculate \( P_{B}(z) \) we need \( \rho(z) \), which depends on \( P_{B}(z) \). Likewise, to calculate \( P_{BH}(w) \) we need \( \rho_{h}(w) \) and \( \rho_{f}(w) \), which depend on \( P_{BH}(w) \). Furthermore, because of the overflow and take-back mechanisms, there is dependence also between the performance parameters of the two levels. The calculus of these parameters has been done by an iteration based on the value of \( P_{BH} \), because, as verified experimentally, it is the parameter that converges more slowly. In particular, in a preliminary step, after establishing system parameters \( (N, M, L, \lambda, \mu_{c}, \mu_{l}, \mu_{h}, \mu_{f}) \) and the transition probability matrix \( p_{w,wh} \), the average number of users for each cell and class can be evaluated by expression (3). At this point, after initializing the quantities \( P_{B}(z), P_{BH}(w), P_{BH2}(z), u_{s}(z), u_{c}(w) \), and \( u_{z}(w) \) to zero, an iterative procedure starts which allows the evaluation of updated values of the previous quantities. Obviously, the iterative process is stopped when the fixed precision \( \varepsilon \) is reached.

The quantities \( P_{BH2}(w), P_{BH2} \) and the subsequent performance parameters can be evaluated at the end of the iteration exploiting final results.

Similar approaches have been followed by other authors [22] and the common experimental experiences show that convergence is always achieved.

D. Forced Termination Probability

The probability of forced termination, \( P_{FT} \), is very important with regard to the quality of service, since it is the probability that an active call is aborted due to lack of resources in the area where the user moves to. For fast users, who can be served only by the macrocell level, we can evaluate the probability of normal termination \( P_{NFT} = 1 - P_{FT} \). If \( t_{c} \) is the call duration and \( t_{2f} \) is the sojourn time of a fast user in a macrocell, assumed to be independent and exponentially distributed with parameters \( \mu_{c} \) and \( \mu_{f} \) respectively, the probability of one transition between macrocells before the end of the call is [18]

\[ \theta_{f} = p(t_{c} > t_{2f}) = \mu_{f} / \left(\mu_{c} + \mu_{f}\right) . \]

(47)

According to the described system, the probability of normal termination for a fast user is the following sum of mutually exclusive events

\[ P_{NFT} = \sum_{k=0}^{\infty} p\{k \text{ successful handoffs}\} \cdot P\{\text{normal termination without further macrocell change}\} . \]

Considering the memoryless property of exponential distributions, we have
For slow users, forced termination probability evaluation is more complex. For these users the probabilities of one transition between microcells, \( \theta_{1s} \) and between macrocells, \( \theta_{2s} \), before the end of the call have an expression similar to (46). If \( t_{cs} \) is the sojourn time of a slow user in a microcell and \( t_{2s} \) (the sojourn time in a macrocell) are exponentially distributed with parameters \( \mu_{1s} \) and \( \mu_{2s} \), respectively, we have

\[
\theta_{1s} = \mu_{1s}/(\mu_{1s} + \mu_{ts}) \quad \theta_{2s} = \mu_{2s}/(\mu_{2s} + \mu_{ts}).
\] (48)

Considering the following events

\begin{align*}
A_u & = \{ \text{the accepted call began in a microcell} \}, \\
A_d & = \{ \text{the accepted call began in a macrocell} \},
\end{align*}

by the theorem of total probability, we have

\[
P_{FTs} = p\{FT|A_u\}p\{A_u\} + p\{FT|A_d\}p\{A_d\} .
\] (49)

The probability that a call is accepted is

\[
p\{\text{the call is accepted} \} = (1-P_{R1}) + P_{R1}(1-P_{R2}) = 1-P_{R1}P_{R2}
\]

and we have [11]

\[
p\{A_u\} = (1-P_{R1})(1-P_{R2}) ; \\
p\{A_d\} = P_{R1}(1-P_{R2})(1-P_{R1}P_{R2}) .
\] (50)

The two conditional probabilities in (49) (see appendix B) are given by:

\[
p\{FT|A_u\} = \frac{\Phi_1 P_{R2} + \Phi_2 P_{R1}(1-P_{R2})}{1 - \Phi_1 P_{R1}(1-P_{R2})} ,
\] (51)

\[
p\{FT|A_d\} = \frac{\Phi_1 P_{R1}P_{R2} + \Phi_2 P_{R1}(1-P_{R1})P_{R2}}{1 - \Phi_1 P_{R1}(1-P_{R2})} ,
\] (52)

where

\[
\Phi_1 = \frac{\theta_{1s} P_{R11}}{1 - \theta_{1s}(1-P_{R11})} , \quad \Phi_2 = \frac{\theta_{2s}}{1 - \theta_{2s}P_{R12}}.
\] (53)

V. NUMERICAL EXAMPLE

A. Description of the Network Example

To assert the validity of our model, the performance of a reference system will be computed, comparing analytical results with those produced by a suitable simulation program, described later (see section V-C).

A network of \( M = 49 \) microcells and 7 macrocells (i.e., \( L=7 \)) has been considered (see Fig. 2). The macrocell diameter is fixed at \( d_2 = 3d_1 \), in which \( d_1 \) is the microcell diameter. The average speeds of slow and fast users are chosen, as an example, to \( v_s = 5 \) km/h and \( v_f = 30 \) km/h, respectively. It is supposed that call duration and sojourn times of the two user classes in microcell and macrocell are exponentially distributed random variables with parameters \( \mu_{ts} \), \( \mu_{2s} \) and \( \mu_{2s} \). The transition probabilities \( p_{k\ell;jh} \) are chosen in such a way as to create a real non uniform distribution of mobile subscribers among cells. Since there are 49 microcells and two mobility classes, the probabilities \( p_{k\ell;jh} \) make up a 98×98 matrix reported in Table I.

Specifying the total number of users in the system and using expression (3), the average number of users of the two classes in each microcell can be evaluated. Choosing \( N=500 \) and approximating \( E[n_k] \) to the nearest integer, the subsequent average distribution of slow and fast users can be estimated (Table II). Adding the values in the seven microcells covered by each macrocell, the average number of slow users per macrocell can be evaluated. Obviously, the sum of slow and fast users over all microcells gives the total number of users considered in the analysis.

B. Numerical Results

In Figures 3-8 the network performance parameters are reported versus \( A_i \), i.e., the rate of new requests per user measured in calls per hour, considering the average number of users per cell shown in Table II. In the graphs, lines represent analytical results, while simulation results are reported through the symbols ‘x’. The considered number of channels for microcells and macrocells are respectively \( C_i=4 \) and \( C_2=6 \). Cell transition rates are \( \mu_{ts} = 0.83 \) tr/h, \( \mu_{2s} = 0.28 \) tr/h, \( \mu_{2f} = 1.67 \) tr/h; while the value of
It is possible to see that the results of the analytical model follow the simulation results with a high level of accuracy; that is, the model hypotheses are reasonable, although there is a small discrepancy for light load conditions. We can see that, with \( \lambda \) fixed, network performance for slow users is better than that for fast ones: this happens because the former can use the resources of microcell and macrocell levels, while the latter can use only the resources of macrocell level. As foreseeable, system performance worsens as load increases, approaching asymptotic values for high load.

### TABLE I. TRANSITION PROBABILITY MATRIX.

<table>
<thead>
<tr>
<th>State</th>
<th>( i-1 )</th>
<th>( i )</th>
<th>( i+1 )</th>
<th>( i+2 )</th>
<th>( i+3 )</th>
<th>( i+4 )</th>
<th>( i+5 )</th>
<th>( i+6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>microcell</td>
<td></td>
<td>microcell</td>
<td></td>
<td>microcell</td>
<td></td>
<td>microcell</td>
<td></td>
<td>microcell</td>
</tr>
<tr>
<td>slow</td>
<td>0.08</td>
<td>0.60</td>
<td>0.20</td>
<td>0.04</td>
<td>0.43</td>
<td>0.43</td>
<td>0.13</td>
<td>0.05</td>
</tr>
<tr>
<td>fast</td>
<td>0.02</td>
<td>0.08</td>
<td>0.21</td>
<td>0.04</td>
<td>0.41</td>
<td>0.43</td>
<td>0.13</td>
<td>0.05</td>
</tr>
</tbody>
</table>

### TABLE II. AVERAGE DISTRIBUTION OF USERS AMONG CELLS.

<table>
<thead>
<tr>
<th>User Class</th>
<th>( i-6 )</th>
<th>( i-5 )</th>
<th>( i-4 )</th>
<th>( i-3 )</th>
<th>( i-2 )</th>
<th>( i-1 )</th>
<th>( i )</th>
<th>( i+1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>microcell</td>
<td></td>
<td>microcell</td>
<td></td>
<td>microcell</td>
<td></td>
<td>microcell</td>
<td></td>
<td>microcell</td>
</tr>
<tr>
<td>slow</td>
<td>0.03</td>
<td>0.12</td>
<td>0.03</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>fast</td>
<td>0.02</td>
<td>0.08</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Fig. 2. Reference system.
Figures 9-12 show network parameters versus microcell diameter $d_1$. Naturally, if $d_1$ changes, macrocell diameter, $d_2$, changes according to relation $d_2 \approx 3d_1$. If the average speeds of the two user classes do not change, the variation in cell dimension is equivalent to a variation in cell transition rate, according to relation (2). Then, when $d_1$ varies between 1 and 10 km, $d_2$ varies between 3 and 30 km, $\mu_1$, between 5 and 0.5 tr/h, $\mu_2$, between 1.67 and 0.167 tr/h, $\mu_2'$ between 10 and 1 tr/h. Moreover, we have considered $N = 500$ users, $\mu_1 = 30$ calls/h, $C_1 = 4$ and $C_2 = 6$, while two values of $\lambda$ have been considered, equal to 2 calls/h in one case and 3 calls/h in the other one.
As expected, forced termination probabilities decrease as cell dimensions increase, while blocking probabilities of new calls are almost insensitive to cell dimensions. In fact, the duration of a call $t_c$ is dependent on the sojourn time in the cells of the two levels. If cell dimensions increase, the sojourn time in the cells increases too and, during a call, there are fewer handoffs; then, the probability that an active user enters into a cell without available channels with a call interruption decreases. As a direct consequence, fewer channels result free to accept new requests with a slight increase in the blocking probability of a new call, which measures the system capacity of accepting new service requests.
Fig. 8. Forced termination probability vs. $\lambda$ for fast users.

Fig. 9. Blocking probability of a new call vs. microcell diameter for slow users.

Fig. 10. Blocking probability of a new call vs. microcell diameter for fast users.

Fig. 11. Forced termination probability vs. microcell diameter for slow users.
C. Simulation Issues

Careful simulation experiments have been exploited for validating analytical results. The simulator is an ad hoc event oriented simulation program, written in C language using the libraries of SMPL simulation language [35]. In the designed simulation tool, the clock advances when one of these events occurs: new call start, call termination and change of microcell and/or class. Naturally, if the user is slow and active, a microcell change could involve an overflow or a take-back.

Call duration, interarrival times and microcell sojourn time, distinguished for the two classes, are generated randomly by suitable exponential distributions. Microcell and/or class transition process follows the described $98 \times 98$ transition probability matrix. As a matter of fact, since we have supposed that active users cannot change class, active slow users move according to probabilities $p_{1s}(a, z)$, introduced in section III-B, while active fast users move according to probabilities $p_{1f}(a, z)$. Collecting for slow and fast users, tried calls, blocked new calls, successful calls, blocked handoffs, tried handoffs, indicated respectively by $T_s$, $T_f$, $B_s$, $B_f$, $S_s$, $S_f$, $BH_s$, $BH_f$, $TH_s$, $TH_f$, the parameters used to measure system performance can be evaluated by the relations

\[ P_{B_s} = \frac{B_s}{T_s}; \quad P_{B_f} = \frac{B_f}{T_f}; \quad P_{Ff} = \frac{T_s - B_s - S_s}{T_s - B_s}; \]
\[ P_{Ff} = \frac{T_f - B_f - S_f}{T_f - B_f}; \quad P_{BHs} = \frac{B_{BHs}}{T_{BHs}}; \quad P_{BHf} = \frac{B_{BHf}}{T_{BHf}}. \]  

(54)

The events of interest are considered after a suitable warm-up system life, thus discarding the transient period, in order to approximate the asymptotic value of the needed quantities.

VI. CONCLUSIONS

Using queuing theory, we have developed an analytical model for two-level hierarchical cellular communication networks with two user classes and dynamic channel allocation in each level. In an urban area the slow users (pedestrians) are normally more numerous than the fast ones (vehicles); then, the former can use the resources of both levels, while the latter can use only the resources of macrocell level. Each cell of the two levels is considered with its own load condition. Both number of cells and population of users are finite, which allows real networks to be represented more precisely. The performance parameters $P_{BS}$, $P_{BF}$, $P_{BHs}$, $P_{BHf}$, $P_{FTs}$, $P_{FTf}$ have been plotted as functions of load $\lambda$ and microcell diameter $d_1$. Simulation results also have been plotted; they assert model validity in the considered context. For each class, other quantities, such as rate of blocked new calls, rate of accepted handoffs, rate of terminated calls, throughput, etc., can be easily derived.

The developed model can be easily generalized to include other significant effects. As an example, users with different mobility patterns and/or service requirements, e.g., users supporting multimedia traffic, can be analyzed considering more than two classes of users. Only minor straightforward modifications to the model are needed, keeping its structure unaltered. Another possible generalization of the model can be obtained by reserving to overflow traffic only a subset of macrocell channels. In this way more fairness between slow and fast users can be achieved. Such a generalization can be integrated quite easily in the model. It is sufficient to consider, for evaluating expression (12), $P_{NBR}(w)$ in (36) without all channels available at macrocell level ($C_2$) but with the number of channels reserved for overflow. A further generalization can be obtained by differentiating the service rate used for evaluating the new call and handoff blocking probabilities, considering, in this way, a more realistic model [36].

APPENDIX A

In this appendix the mobility modeling of the system as a generalized version with multi-class users of the model presented in [26] is described. The target of this mobility model is to find the average number of users in the various cells. User mobility is
described with a closed single chain multi-class queuing network. There are $M$ cells and $N$ users divided in $Q$ mobility classes. The sojourn time is a random variable with mean $1/\mu_r$ ($r = 1, 2, \ldots, Q$) and an arbitrary p.d.f. A user can pass from microcell $i$ at class $r$ to microcell $j$ at class $h$ with a transition probability $p_{r:jh}$. Thus, in the stated hypothesis the network of queues admits a product form solution \cite{27, 28}.

The first step is the resolution of the following homogeneous linear system to find the relative visit rate to each microcell

$$e_r = \sum_{i=1}^{M} \sum_{j=1}^{Q} e_{ir} P_{jr} \quad i = 1, \ldots, M; \; r = 1, \ldots, Q. \quad (A1)$$

Applying the result for BCMP networks with infinite servers, the evaluation of distribution $n = (n_{i1}, n_{i2}, \ldots, n_{iQ}, n_{Q1}, n_{Q2}, \ldots, n_{MQ}, \ldots, n_{MQ})$ of users at each class in the various microcells \cite{28} can be found

$$P(n) = \frac{1}{G} \prod_{r=1}^{M} \prod_{i=1}^{Q} \frac{1}{n_{ir}!} \left( \frac{e_{ir}}{\mu_r} \right)^{n_{ir}} \quad (A2)$$

where the departure rate $\mu_r$ is equal to $\mu_e$ for each microcell. The normalization constant $G$ is given by

$$G = \sum_{n} \prod_{r=1}^{M} \prod_{i=1}^{Q} \frac{1}{n_{ir}!} \left( \frac{e_{ir}}{\mu_r} \right)^{n_{ir}} = \sum_{n} \prod_{k=1}^{MQ} \frac{e_{nk}}{\mu_k} \quad (A3)$$

where $k = (i-1)Q + r$ and $x_k = e_r/\mu_r$.

Using multinomial expression, $G$ can be computed in closed form

$$G = \frac{1}{N!} \left( \sum_{i=1}^{N} x_i \right)^N \quad (A4)$$

where $N$ is always the total number of users in the network.

At this point, the most significant parameters of the network might be evaluated. In particular, the marginal state probability is given by

$$p(n_r = t) = \sum_{n_{ir} \geq 0} \frac{1}{G} \prod_{k=1}^{Q} \frac{n_{rk}!}{n_{rk}!} \left( \frac{e_{rk}}{\mu_k} \right)^{n_{rk}}.$$

Using the substitutions $k = (i-1)Q + r$, the multinomial formula, we have

$$p(n_r = t) = \frac{x_r^t}{t!} \frac{1}{G} \prod_{k=1}^{Q} \frac{n_{rk}!}{n_{rk}!} \left( \frac{e_{rk}}{\mu_k} \right)^{n_{rk}} = \frac{1}{G} \frac{1}{t!} \frac{1}{(N-t)!} \left( \sum_{k=2}^{M} \frac{x_k}{n_{rk}!} \right)^{N-t}.$$

Coming back to the indexes $i$ and $r$ and using (A4), the marginal state probability is given by

$$p(n_r = t) = \left( \frac{N}{t} \right)^{x_r} \left( \sum_{j=1}^{M} \sum_{h=1}^{Q} X_{rh} \right)^{N-t} \left( \sum_{j=1}^{M} \sum_{h=1}^{Q} X_{rh} \right)^{t}.$$

After a bit of algebra, the average number of users at class $r$ in microcell $i$ is given by

$$E[n_r] = \sum_{i=1}^{M} \left( \frac{N}{t} \right)^{x_r} \left( \sum_{j=1}^{M} \sum_{h=1}^{Q} X_{rh} \right)^{N-t} \left( \sum_{j=1}^{M} \sum_{h=1}^{Q} X_{rh} \right)^{t}.$$

thus, $E[n_r]$ is the average value of a binomial random variable

$$E[n_r] = N \frac{x_r}{\sum_{j=1}^{M} \sum_{h=1}^{Q} X_{rh}} = N \frac{e_r/\mu_r}{\sum_{j=1}^{M} \sum_{h=1}^{Q} e_{rh}/\mu_{rh}} \quad i = 1, \ldots, M \quad (A6)$$

$$r = 1, \ldots, Q \quad (A6)$$

APPENDIX B

In this appendix, the two conditional probabilities considered in (49) will be evaluated. In particular, due to the overflow and take-back mechanisms, the slow user could “bounce” between the two hierarchical levels several times before the forced termi-
nation of the call. Therefore, the event \{forced termination | the accepted call began in a microcell\} is the union of the subsequent infinite disjoint events:
1. after a certain number of accepted handoffs between microcells, \(m-m\), the call is interrupted for the failure of both a handoff attempt toward a microcell and an overflow attempt;
2. after several accepted handoffs between microcells \(m-m\), an accepted overflow \(m-M\) and several handoffs between macrocells \(M-M\) without possibility of take-back, the call is interrupted for the failure of a handoff between macrocells and the failure of a take-back attempt (because a macrocell change always coincides with a microcell change for the correspondence between the cells of the two levels);
3. after several handoffs \(m-m\), an overflow \(m-M\), several handoffs \(M-M\) without possibility of take-back, the call is interrupted for the failure of a handoff between macrocells and the failure of a take-back attempt;
4. after several handoffs \(m-m\), an overflow \(m-M\), several handoffs \(M-M\) without possibility of take-back, a take-back \(M-m\), several handoffs \(m-m\), an overflow \(m-M\) and several handoffs \(M-M\) without possibility of take-back, the call is interrupted for the failure of a handoff between macrocells and the failure of a take-back attempt;
and so on, bouncing between the two levels.

The probability of the first event is

\[
P(1) = \sum_{n=0}^{\infty} \theta_i^{n+1} (1 - P_{BH1})^n P_{BH1} P_{B2} = \frac{\theta_i P_{BH1} P_{B2}}{1 - \theta_i (1 - P_{BH1})} = \Phi_1 P_{B2}
\]

whit \(\Phi_1\) defined in (53).

Now let us suppose that an active slow user passes to the second level, after an overflow, and executes several handoffs \(M-M\) without possibility of take-back. In this situation, the used model is not able to reveal the exact number of take-back attempts effected and failed in each macrocell that the user crosses. In order to evaluate the probabilities of the events 2, 3, 4, etc. we have to make use of an estimate of the mean number of these failed take-back attempts, that is reasonably equal to the subsequent ratio

\[
k = \mu_i / \mu_{s2}.
\]

Since the probability of overflow failure is the same as \(P_{B2}\) and the probability of take-back failure is the same as \(P_{B1}\), the probability of the second event is

\[
P(2) = \sum_{n_1=0}^{\infty} \theta_i^{n_1+1} (1 - P_{BH1})^{n_1} P_{BH1} P_{B2} (1 - P_{B2})^n = \frac{\theta_i P_{BH1} P_{B2} (1 - P_{B2})}{1 - \theta_i (1 - P_{BH1})(1 - P_{B2})} = \Phi_2 P_{B2} (1 - P_{B2}) P_{BH2}
\]

in which there is the product of two geometrical series. Therefore

\[
P(2) = \frac{\theta_i P_{BH1} P_{B2}}{1 - \theta_i (1 - P_{BH1})(1 - P_{B2})} = \Phi_2 P_{B2} (1 - P_{B2}) P_{BH2}
\]

whit \(\Phi_2\) defined in (53). In a similar way

\[
P(3) = \Phi_2^2 P_{B2} (1 - P_{B2}),
P(4) = \Phi_2^3 P_{B2} (1 - P_{B2})^2 P_{BH2},
\]

generalizing, we have

\[
P(2n + 1) = \Phi_2^{n+1} P_{B2} (1 - P_{B2})^{n+1} P_{B2},
P(2n + 2) = \Phi_2^{n+1} P_{B2}^{n+1} (1 - P_{B2})^{n+1} P_{B2} P_{BH2}
\]

with \(n = 0, 1, \ldots\)

\(^1\) \(m-m\) means “handoff between two microcells”; \(m-M\) means “overflow between a microcell and a macrocell”; \(M-m\) means “take-back between a macrocell and a microcell”; \(M-M\) means “handoff between two macrocells".
Summing the probabilities of all events, we have the expression for \( p\{|FT, |A_m\} \) in (51).

In a similar way, the probability of the event \( \{|\text{forced termination} | \text{the accepted call began in a macrocell}\} \) is given by the sum of the probabilities of the subsequent infinite disjoint events:

1. after several accepted handoffs between macrocells M-M without possibility of take-back, the call is interrupted for the failure of a handoff between macrocells and the failure of a take-back attempt;
2. after several handoffs M-M without possibility of take-back, an accepted take-back M-m and several handoffs m-m, the call is interrupted for the failure of both a handoff attempt toward a microcell and an overflow attempt;
3. after several handoffs M-M without possibility of take-back, a take-back M-m, several handoffs m-m, an overflow M-M and several handoffs M-M without possibility of take-back, the call is interrupted for the failure of a handoff between macrocells and the failure of a take-back attempt;
4. after several handoffs M-M without possibility of take-back, a take-back M-m, several handoffs m-m, an overflow m-M and several handoffs M-M without possibility of take-back, an accepted take-back M-m and several handoffs m-m, the call is interrupted for the failure of both a handoff attempt toward a microcell and an overflow attempt;

and so on, bouncing between the two levels.

Following an approach similar to the previous case, we have

\[
P(1) = \Phi_1 \Phi_2 P_{M-M} P_{M-M} P_{M-M},
\]

\[
P(2) = \Phi_1 \Phi_2 (1 - P_{M-M}) P_{M-M},
\]

\[
P(3) = \Phi_1 \Phi_2^2 (1 - P_{M-M}) (1 - P_{M-M}) P_{M-M} P_{M-M} P_{M-M},
\]

\[
P(4) = \Phi_1 \Phi_2^3 (1 - P_{M-M}) (1 - P_{M-M}) P_{M-M} P_{M-M} P_{M-M} P_{M-M},
\]

and so on

\[
P(2n + 1) = \Phi_1^{n+1} \Phi_2^n (1 - P_{M-M})^n (1 - P_{M-M}) P_{M-M} P_{M-M} P_{M-M} P_{M-M},
\]

\[
P(2n + 2) = \Phi_1^{n+1} \Phi_2^n (1 - P_{M-M})^n (1 - P_{M-M})^2 P_{M-M} P_{M-M} P_{M-M} P_{M-M} P_{M-M},
\]

(B7)

with \( n = 0, 1, \ldots \)

Summing the probabilities of all events, we have the expression for \( p\{|FT, |A_M\} \) in (52).

REFERENCES


