Performance analysis of cellular mobile communication networks supporting multimedia services *

M. Ajmone Marsan [a], S. Marano [b], C. Mastroianni [b] and M. Meo [a]

[a] *Dipartimento di Elettronica, Politecnico di Torino, Italy*
[b] *D.E.I.S., Università della Calabria, Rende (CS), Italy*

This paper illustrates the development of an analytical model for a communication network providing integrated services to a population of mobile users, and presents performance results to both validate the analytical approach, and assess the quality of the services offered to the end users. The analytical model is based on continuous-time multidimensional birth–death processes, and is focused on just one of the cells in the network. The cellular system is assumed to provide three classes of service: the basic voice service, a data service with bit rate higher than the voice service, and a multimedia service with one voice and one data component. In order to improve the overall network performance, some channels can be reserved to handovers, and multimedia calls that cannot complete a handover are decoupled, by transferring to the target cell only the voice component and suspending the data connection until a sufficient number of channels become free. Numerical results demonstrate the accuracy of the approximate model, as well as the effectiveness of the newly proposed multimedia call decoupling approach.

1. Introduction

The ever-increasing popularity of cellular and cordless telephone services is making wireless access and mobility two major components in the evolution of telecommunication networks.

In the design of the first and second generation cellular mobile telephony systems (such as ETACS and GSM in Europe), the technical approach mainly focused on increasing the capacity available for voice services, so as to cope with the explosive growth in the number of subscribers.

Today, the need for an increased system capacity is combined with the request for a wider spectrum of telecommunication services, in order to be able to offer data services in addition to plain telephony; this will pave the way to the introduction of wireless multimedia services for mobile users, including voice, data and images.

While the performance of cellular telecommunication networks offering mobile telephony services was investigated by many authors under several different operating conditions, the same cannot be said of networks offering a variety of services to mobile users.

The performance of cellular telecommunication networks can be investigated by using either simulation or analytical models (or a combination of both). Though simulation is often preferred when aiming at the detailed study of the behavior of a specific cellular system covering a given area, in recent years a number of analytical frameworks were developed to obtain more general results. Some examples of analytical approaches to the performance analysis of cellular telecommunication networks are found in [1–9].

In [8] the performance of a personal communication network based on microcells covering city streets is analyzed to determine some important teletraffic parameters, such as the blocking probabilities of new calls and handovers, the carried traffic and the spectrum efficiency. The fluid model in [4] describes a wireless system fed with traffic scenarios based on Poisson time dependent processes. Techniques to reduce forced terminations of calls in progress due to handover failures are proposed and evaluated in [2], where several priority schemes are defined, which reserve channels to handovers. In [1] the performance of a hierarchical cellular system based on microcells and overlaying macrocells is analyzed, and the benefit of introducing “tier handovers”, i.e., handovers between cells belonging to different hierarchical levels, is discussed. This research is extended in [6], where the performance of a more complex network, comprising n hierarchical cell levels, is evaluated. In [9] a model of a circuit switched cellular network with m classes of traffic sources is presented. Each class is defined by different resource and performance requirements, and therefore has a different blocking behavior. New access protocols for the integration of different classes of service in a packet switched environment, mainly based on PRMA (Packet Reservation Multiple Access) are proposed and analyzed in [3,7].

In this paper we develop an analytical model based on continuous-time multidimensional birth–death processes, in order to evaluate the performance of a cellular system designed to provide three classes of service: the basic voice service, a data service with bit rate higher than the voice service, and a multimedia service with one voice and one data component, to be managed together (multimedia communication systems must be able to manage not only isolated connections, possibly belonging to different service classes, but also groups of connections that carry the different components of a multimedia service, such as voice

* This work was supported by the Italian National Research Council (CNR) and by the Italian Ministry for University and Scientific Research.

and data, that in the simpler cases are set up at the same time, and terminate at the same time).

Furthermore, a new mechanism is proposed to decrease the probability that an active multimedia call is interrupted due to a handover failure: when the base station governing a cell cannot serve the handover request of a multimedia call, an attempt is made to serve at least the voice component. If this operation succeeds, the multimedia call is decoupled: the voice call is continued, while the data connection is suspended, and will be resumed as soon as possible.

The paper is organized as follows: section 2 presents the analytical model. Section 3, which contains numerical results, consists of two parts. First, we validate our model by comparing some analytical results with those obtained by a very detailed simulation model. Then we present some performance comparisons among alternate configurations of the system. Conclusions are given in section 4.

2. Model description

The analytical model is based on a continuous-time multidimensional birth–death process, as we already mentioned; the model illustration is organized in four stages:

1. Discussion of the main assumptions and of the model parameters.
2. State definition and identification of the model driving processes.
3. Derivation of the flow balance equations and computation of the equilibrium state probabilities.
4. Evaluation of aggregate performance measures.

2.1. Assumptions and parameters

The cellular telecommunication system comprises a large number of cells, and provides three classes of service to satisfy different kinds of users requests:

- *Class A* service is intended to support basic voice calls which require narrowband connections.
- *Class B* service satisfies the requirements of data calls (or slow video calls) with (moderately) wideband connections.
- *Class C* service is intended to support multimedia calls composed of a data component and a voice component, which are set up and terminated at the same time, and must be managed together.

Denoting with N the number of radio channels available in a given cell of the cellular telecommunication system, we assume that the establishment of connections of class A requires just one channel, that connections of class B require C_d channels and, finally, that connections of class C require $C_\mathrm{d}+1$ channels.

A service request from a mobile user to a base station may be due either to the generation of a new call or to a handover request. Since handover failures force the termination of a call in progress, these events are considered to be worse than new calls blocking, whose effect is just to force the user to repeat his access request at a later time. Therefore, two mechanisms are introduced in the system in order to favor handover requests with respect to new call requests, under heavy traffic conditions. Priority is given to handover requests by reserving a small number of channels, denoted by C_h, for handovers ($C_\mathrm{h} \geqslant 0$). Furthermore, when the handover of a multimedia call is requested, the base station tries to accommodate the entire call, but, if this is not possible, the two components of the call can be decoupled: the voice connection is accepted, if possible, while the data connection is temporarily suspended, waiting to be resumed as soon as enough channels are available in the cell to serve the entire multimedia call.

Decoupling a multimedia call because of a partially served handover can improve the quality of service, mainly because it allows the user to continue at least partly his communication. Furthermore, though the data connection is suspended, the user is given a chance to resume it later from the point it was interrupted; this can be a very useful feature for data applications, such as file transfer, for example. In fact, if the entire handover had been blocked, the whole multimedia connection would have been suddenly terminated, and all transferred data would have been lost.

A decoupled multimedia call occupies only one channel, like a normal voice call; in the following we will refer to decoupled multimedia calls as *class D* connections.

To facilitate recombination of decoupled multimedia calls, new calls are not accepted as long as some active class *D* connections exist in the cell.

The amount of time that a Mobile Terminal (MT) with an ongoing call remains within the area covered by the base station of the cell under investigation is called *dwell time*; if the call is still active after the dwell time, the MT requests a handover toward an adjacent cell. The *unencumbered session duration* of a call is defined as the amount of time that the call would remain in progress if it could continue to completion, without forced termination due to handover failure.

The analysis of the cellular telecommunication system focuses on a single cell, whose behavior is isolated from those of other cells, that are collectively described only through the handover request processes observed by the cell under investigation. Of course, if the network is highly symmetric, and the traffic is homogeneous, the analysis of one cell may suffice for the assessment of the quality of the services offered by the whole system; otherwise, several cells must be separately studied.

The following assumptions render the isolated cell analysis problem amenable to solution using the theory of multidimensional birth–death processes:

- The new call request arrival processes for voice, data and multimedia services are modeled as Poisson point processes. The mean new voice call request arrival rate, denoted by λ_{v}, is the product of the mean new voice call request rate from one MT and the number of MTs in the cell. Mean call rates for data calls and multimedia calls, λ_{d} and λ_{m}, are obtained in a similar way.

- The handover request arrival processes from adjacent cells follow Poisson point processes. The mean handover arrival rates corresponding to voice, data and multimedia calls are denoted by λ_{hv}, λ_{hd} and λ_{hm}, respectively.
- The dwell time of a MT in the cell is a random variable with negative exponential probability density function, with mean $1/\mu_{\mathrm{hv}}$ in the case of a voice call, $1/\mu_{\mathrm{hd}}$ for a data call, and $1/\mu_{\mathrm{hm}}$ for a multimedia call.
- The unencumbered session duration of a call is a random variable with negative exponential pdf whose mean value is $1/\mu_{\mathrm{v}}$, $1/\mu_{\mathrm{d}}$, and $1/\mu_{\mathrm{m}}$, for voice, data and multimedia calls, respectively.

The values of the parameters can be derived from characteristics of the system such as the cell size and the speed of the mobile users.

2.2. State definition and driving processes

The cell state, in any instant, is determined by the number of currently active connections for each class of traffic, and it is, therefore, given by the vector:

$$s = (v, d, m, r),$$

where

- v is the number of active voice calls in the cell;
- d is the number of active data calls;
- m is the number of multimedia calls, with both components (voice and data) active;
- r is the number of active voice components of decoupled multimedia calls; if $r > 0$, some suspended data calls are waiting to be resumed as soon as a sufficient number of channels becomes available.

We denote with $n(s)$ the function giving the total number of channels allocated to active connections when the cell is in state s:

$$n(s) = v + dC_{\mathrm{d}} + m(C_{\mathrm{d}} + 1) + r.$$

We shall simply write n instead of $n(s)$ when no ambiguity arises.

Since the number of channels available in the cell is N, the maximum values of v, d and m are respectively N, $\lfloor N/C_{\mathrm{d}} \rfloor$ and $\lfloor N/(C_{\mathrm{d}} + 1) \rfloor$. Whereas these values are determined by the cell configuration, the maximum number of decoupled calls that can be active at the same time, r, is limited by a threshold $r_{\max}$, that must be fixed by the operator of the cellular system; if no threshold is defined, r can grow up to N. A permissible state s must satisfy the condition $n(s) \leqslant N$.

Let $\mathcal{S}$ be the state space of the model we just described; it is convenient to order and number states from 0 to $S_{\max}$.

The model dynamics is determined by a number of *driving stochastic processes* which cause state transitions at random instants.

Driving processes produce different kinds of events that must be processed by the network. In our case, with reference to the cell under investigation, they are the following:

- requests of new voice, data and multimedia calls,
- incoming handover requests for voice, data and multimedia,
- outgoing handover requests for voice, data, multimedia and decoupled calls,
- completion of voice, data, multimedia and decoupled calls.

Since only one cell is analyzed, outgoing handover requests and call completions have the same effect on the model (i.e., some previously busy channels become free); on the contrary, new call requests and incoming handover requests have to be considered separately, because a handover request of a multimedia call can lead to a call decoupling, while a new call request cannot.

2.3. Flow balance equations and equilibrium probabilities

Since the Markovian model is homogeneous and irreducible, with finite state space, an equilibrium (or steady-state) distribution $\mathbf{p} = \{p(i)\}$, with $i = 0, \ldots, S_{\max}$ exists, and can be computed through the matrix equation $\mathbf{p} \cdot \mathbf{Q} = \mathbf{0}$, where $\mathbf{Q}$ is the infinitesimal generator matrix, together with the normalization condition $\sum_{i=0}^{S_{\max}} p(i) = 1$.

The transition rates, i.e., the elements $q(i, j)$ of matrix $\mathbf{Q}$, are obtained from the analysis of the system driving processes. For each driving process, it is possible to determine what state transitions can happen, i.e., what are the possible *successor states* of a generic state $s = (v, d, m, r)$. This is what we discuss next.

New call requests

A new call is accepted in the cell if the number of free channels, excluding those reserved to handovers, is such that the call can be accommodated. Furthermore, new calls are refused if some class *D* connections are active, (i.e., if $r > 0$), in order to favor recombination of decoupled multimedia calls.

Table 1 shows, for each type of new call, the conditions on the model state for a transition to be possible, the rate associated with the transition, and the successor state.

Table 1
Transitions due to a new call request.

Type	Condition	Successor state	Rate
Voice	$(n \leqslant N - C_{\mathrm{h}} - 1) \wedge r = 0$	$(v + 1, d, m, r)$	λ_{v}
Data	$(n \leqslant N - C_{\mathrm{h}} - C_{\mathrm{d}}) \wedge r = 0$	$(v, d + 1, m, r)$	λ_{d}
Mmedia	$(n \leqslant N - C_{\mathrm{h}} - (C_{\mathrm{d}} + 1)) \wedge r = 0$	$(v, d, m + 1, r)$	λ_{m}

Note that a new multimedia call is accepted only if the system can accommodate both the voice and the data components; decoupling a multimedia call is only permitted when the connection is already active.

Incoming handover requests

The incoming handover request for a voice or data call is accepted if enough free channels to accommodate the request are available in the cell. In the first two rows of table 2 the transitions associated with the acceptance of an incoming voice or data handover are listed, together with their rates, and the conditions on the state s.

In the case of an incoming handover request for a multimedia call (whether decoupled or not), two possible successor states are identified, corresponding to a handover that can either be completely served or produces a decoupled call. In the first case (third row of table 2), both the voice and the data components can be accommodated. Otherwise, the handover results in the transfer of the voice component only, and in the decoupling of the multimedia call.

Completion of calls and outgoing handover requests

Both the completion of calls and the outgoing handover requests have the effect of freeing some channels in the cell; in some cases this leads to the recombination of a decoupled multimedia call. In fact, the data component of a decoupled call is resumed if, after channels are freed, at least C_d channels are available.

The first row of table 3 refers to the case of a voice call termination that is not followed by a recombination of a decoupled multimedia call. This event takes place if either $r = 0$ (no recombination is necessary) or $r > 0$ but $n - 1 > N - C_\mathrm{d}$ (no recombination is possible). The second row of the same table refers to the case in which a recombination follows a voice call completion. Similarly, the third and fourth rows of table 3 refer to a data call completion.

Table 2
Transitions due to an incoming handover request.

Type	Condition	Successor state	Rate
Voice	$n \leqslant N - 1$	$(v+1, d, m, r)$	λ_hv
Data	$n \leqslant N - C_\mathrm{d}$	$(v, d+1, m, r)$	λ_hd
Mmedia	$n \leqslant N - (C_\mathrm{d}+1)$	$(v, d, m+1, r)$	λ_hm
	$N - (C_\mathrm{d}+1) < n \leqslant N - 1$	$(v, d, m, r+1)$	λ_hm

In the case of a multimedia call completion, $C_\mathrm{d} + 1$ channels are freed in the cell, and it is possible that one, or even two decoupled multimedia calls can be recombined. Two suspended data connections are resumed if $r > 1$, and $n - (C_\mathrm{d} + 1) \leqslant N - 2C_\mathrm{d}$, since $2C_\mathrm{d}$ channels are free after the multimedia call completion.

The last two rows of table 3 refer to the cases of completion of a decoupled multimedia call.

For voice, data and multimedia calls, respectively, $M_\mathrm{v} = \mu_\mathrm{v} + \mu_\mathrm{hv}$, $M_\mathrm{d} = \mu_\mathrm{d} + \mu_\mathrm{hd}$ and $M_\mathrm{m} = \mu_\mathrm{m} + \mu_\mathrm{hm}$.

2.4. Model complexity

The transition rate from a state s to a state k is computed by summing the contributions resulting from the driving processes that were just described.

The state space size S (hence, the dimension of matrix **Q**, which is $S \times S$) depends on the values of C_d, r_max, and, most important, on the number of radio channels available in the cell, N.

Let the maximum number of active voice, data and multimedia calls be respectively $V = N$, $D = \lfloor N/C_\mathrm{d} \rfloor$ and $M = \lfloor N/(C_\mathrm{d} + 1) \rfloor$. The state space size S is upper bounded by $(V + 1)(D + 1)(M + 1)(r_\mathrm{max} + 1)$ (due to the restriction that the total number of busy channels is smaller than or equal to N, the expression above is not exact, but gives a close upper bound for S). S can be rather large, so that in the computation of the equilibrium probability distribution it is essential to use a numerical algorithm that exploits the sparseness of the infinitesimal generator; choosing the best representation of **Q**, in order to minimize computation time and memory requirements, is not a minor task. For the results presented in section 3 we used an iterative solution technique which requires for each iteration step a number of multiplications equal to the number of elements of **Q** which are different from 0, that in our case is about $S_\mathrm{max} \cdot 10$.

2.5. Aggregate performance measures

A number of interesting performance measures can be derived from the steady-state probabilities, to characterize the cell behavior in conditions of statistical equilibrium.

Table 3
Transitions due to completion of calls and outgoing handover requests.

Type	Condition	Successor state	Rate
Voice	$(r > 0 \wedge (n-1 > N - C_\mathrm{d})) \vee r = 0$	$(v-1, d, m, r)$	vM_v
	$r > 0 \wedge (n-1 \leqslant N - C_\mathrm{d})$	$(v-1, d, m+1, r-1)$	vM_v
Data	$r = 0$	$(v, d-1, m, r)$	dM_d
	$r > 0$	$(v, d-1, m+1, r-1)$	dM_d
Mmedia	$r = 0$	$(v, d, m-1, r)$	mM_m
	$r = 1 \vee (r > 1 \wedge (N - 2C_\mathrm{d} < n - (C_\mathrm{d}+1)))$	$(v, d, m, r-1)$	mM_m
	$r > 1 \wedge (n - (C_\mathrm{d}+1) \leqslant N - 2C_\mathrm{d})$	$(v, d, m+1, r-2)$	mM_m
Decoup.	$r > 1 \wedge (n-1 \leqslant N - C_\mathrm{d})$	$(v, d, m+1, r-2)$	rM_m
	$r = 1 \vee (r > 1 \wedge (n-1 > N - C_\mathrm{d}))$	$(v, d, m, r-1)$	rM_m

The following performance indices will be used in the presentation of numerical results:

- the average carried traffic,
- the new call blocking probability,
- the handover failure probability,
- the probability of recovering a decoupled multimedia call.

First of all, it is useful to define the *transition rate* or *event rate*, that is the mean frequency of events that cause a state transition, where an event can be a new call attempt, a handover request, or a call termination. The event rate R is given by

$$R = -\sum_{s=0}^{S_{\max}} q(s,s)p(s),$$

where $-q(s,s)$ is the total transition rate out of state s, and $p(s)$ is the steady-state probability of state s, i.e., the average fraction of time that the system spends in state s at steady-state.

The performance indices of interest can be defined as follows, for the cell under consideration.

- The *average carried traffic* AC is the average number of channels in use in the cell, and it is given by

  $$\mathrm{AC} = \sum_{s=0}^{S_{\max}} n(s)p(s),$$

  where $n(s)$ is the number of busy channels when the system is in state s and $p(s)$ is the steady-state probability of state s.

  It is also possible to evaluate how channels are partitioned among the different kinds of service. $\mathrm{AC_v}$, $\mathrm{AC_d}$ and $\mathrm{AC_m}$ are defined as the average number of channels respectively used by voice, data and multimedia connections. These quantities are computed as

  $$\mathrm{AC_v} = \sum_{s=0}^{S_{\max}} v(s)p(s),$$
  $$\mathrm{AC_d} = C_\mathrm{d} \cdot \sum_{s=0}^{S_{\max}} d(s)p(s),$$
  $$\mathrm{AC_m} = (C_\mathrm{d}+1) \cdot \sum_{s=0}^{S_{\max}} m(s)p(s).$$

- The *new call blocking probability* PB is defined as the average fraction of new call requests that cannot be satisfied by the cell base station, due to the lack of free channels or to the presence of some decoupled multimedia connections waiting to be recombined. This probability is computed separately for each class of service.

  A new voice call is blocked if $n > N - C_\mathrm{h} - 1$ or if $r > 0$, that is if the system is in one of the states belonging to the subset B_v defined as $B_\mathrm{v} = \{s\colon (n > N - C_\mathrm{h} - 1) \vee r > 0\}$. Thus, the blocking probability of new voice calls is

  $$\mathrm{PB_v} = \sum_{s\in B_\mathrm{v}} p(s).$$

  In a similar way it is possible to compute the blocking probabilities of new data calls ($\mathrm{PB_d}$), and new multimedia calls ($\mathrm{PB_m}$):

  $$\mathrm{PB_d} = \sum_{s\in B_\mathrm{d}} p(s),$$
  $$\mathrm{PB_m} = \sum_{s\in B_\mathrm{m}} p(s),$$

  where B_d and B_m are the subsets of states in which new data or multimedia calls cannot be accepted:

  $$B_\mathrm{d} = \bigl\{s\colon (n > N - C_\mathrm{h} - C_\mathrm{d}) \vee r > 0\bigr\},$$
  $$B_\mathrm{m} = \bigl\{s\colon (n > N - C_\mathrm{h} - C_\mathrm{d} - 1) \vee r > 0\bigr\}.$$

  Note that $B_\mathrm{v} \subseteq B_\mathrm{d} \subseteq B_\mathrm{m}$ so that $\mathrm{PB_v} \leqslant \mathrm{PB_d} \leqslant \mathrm{PB_m}$.

- The *handover failure probability* PH is defined as the average fraction of incoming handover requests that cannot be satisfied, causing the forced termination of the call. Incoming handover requests for voice, data and multimedia calls fail if the system is in one of the states belonging respectively to subsets H_v, H_d and H_m, defined as follows:

  $$H_\mathrm{v} = \{s\colon n > N-1\},$$
  $$H_\mathrm{d} = \{s\colon n > N-C_\mathrm{d}\},$$
  $$H_\mathrm{m} = \{s\colon n > N-C_\mathrm{d}-1\}.$$

  Therefore, handover failure probabilities are given by

  $$\mathrm{PH_v} = \sum_{s\in H_\mathrm{v}} p(s),$$
  $$\mathrm{PH_d} = \sum_{s\in H_\mathrm{d}} p(s),$$
  $$\mathrm{PH_m} = \sum_{s\in H_\mathrm{m}} p(s).$$

  $\mathrm{PH_m}$ is the probability that the system is not able to accommodate both components of a multimedia connection. However, if at least the voice component can be served, the call is decoupled, and the data component is suspended. This event happens when, at the time of the incoming handover request, the system is in one of the states of the subset $H_\mathrm{dec} = \{s\colon (N - C_\mathrm{d} - 1 < n \leqslant N-1) \wedge r < r_{\max}\}$. We define the *decoupling probability* $\mathrm{PH_{dec}}$ as the probability that a multimedia call is decoupled after a handover request. $\mathrm{PH_{dec}}$ is given by

  $$\mathrm{PH_{dec}} = \sum_{s\in H_\mathrm{dec}} p(s).$$

  Finally, PH_r is defined as the probability that the handover request of the multimedia call completely fails, and both components of the multimedia connection are forced to terminate:

$$\mathrm{PH}_r = \mathrm{PH}_\mathrm{m} - \mathrm{PH}_\mathrm{dec}.$$

- The *probability of recovering a decoupled multimedia call*, P_rec, is defined as the probability that the data component of a decoupled multimedia call is resumed before the call is terminated by the user. P_rec is obtained as follows:

$$P_\mathrm{rec} = \frac{\overline{R}_\mathrm{rec}}{\mathrm{AC}_\mathrm{r}}.$$

In this expression, $\overline{R}_\mathrm{rec}$ is the relative rate of events that lead to the recombination of a multimedia call (normalized to the overall event rate R), and AC_r is the average number of decoupled calls in the cell:

$$\overline{R}_\mathrm{rec} = \frac{1}{R} \cdot \sum_{s=0}^{S_\mathrm{max}} R_\mathrm{rec}(s)p(s),$$

$$\mathrm{AC}_\mathrm{r} = \sum_{s=0}^{S_\mathrm{max}} r(s)p(s),$$

where $R_\mathrm{rec}(s)$ is the recombination rate and $r(s)$ is the number of decoupled calls, when the system is in state s.

3. Results

This section consists of two parts. First we present some comparisons between analytical and simulation results in order to validate our approximate modeling approach; then we explore some alternate system configurations to assess their effectiveness.

As a basic scenario we refer to a configuration similar to the one being considered by ETSI (the European Telecommunications Standards Institute) for the introduction of (moderately) high speed data services within the European wireless telephony network. We, thus, consider two classes of service only: class A (voice call) and class B (data call). Each data call requires the allocation of $C_\mathrm{d} = 4$ channels.

The mean dwell time of voice or data calls is set to 80 s, while the mean unencumbered session duration is taken to be 100 s. The fraction of voice calls is assumed to be 75% of the total, the remaining 25% being data calls. The number of channels in the cell is taken to be 64. The number of users in the cell under investigation is taken to be 500. The performance indices will be presented as curves plotted versus the input load in terms of total call rate.

Table 4 provides the list of the parameter values that refer to the basic scenario (values given in the first position); additional values are given within brackets; when the used values differ from those of the basic scenario, this will be explicitly indicated in the figure caption.

Table 4
Parameter values used in the presentation of numerical results.

Parameter	Values
N	64
C_d	4 (8)
C_h	0 (4,8)
r_max	C_d (0)
$1/\mu_\mathrm{v} = 1/\mu_\mathrm{d} = 1/\mu_\mathrm{m}$	100 s
$1/\mu_\mathrm{hv} = 1/\mu_\mathrm{hd} = 1/\mu_\mathrm{hm}$	80 s

3.1. Model validation

As we already mentioned, validating the analytical model by comparison with the results produced by a very detailed simulation of the wireless network is a necessity, because of the numerous simplifying assumptions adopted in the model development, in order to keep complexity under control. Recall that the major simplification stems from the choice of studying just one cell in isolation, instead of developing a detailed model of the entire network with the description of the interactions among adjacent cells. This also requires assuming that the average incoming handover flow is equal to the average outgoing handover flow.

The simulator provides the description of a whole network comprising several cells. In each cell the stochastic representation of the new call request traffic and of the user mobility is the same as in the model: new calls are generated according to Poisson processes and call durations and dwell times are random variables with negative exponential distributions. The differences between the two approaches are due to the representation of the handover flow. While in the analytical model the handover flow entering a cell is assumed to be Poisson and its rate is derived by balancing the average flows of incoming and outgoing handovers, in the simulator the handovers are described in detail. Once a user issues a handover request towards a neighboring cell, a procedure starts which releases resources in the current cell, checks for available resources in the target cell and possibly allocates resources to the incoming handover. The correlation between the behaviors of two adjacent cells involved in a handover procedure is in this way accurately described. Simulation results will be presented for a network comprising seven hexagonal cells, comparing the performance estimates referring to the central cell against analytical results.

Confidence intervals are obtained for each simulation experiment by using the “batch means” technique, with confidence level equal to 0.95.

The CPU time needed to solve the analytical model depends on the accuracy required in the computation of the Markov chain solution at steady state, and in the handover flow balancing procedure; the CPU time consumed by simulations, instead, depends on the desired confidence level and interval width. It is therefore difficult to provide a fair comparison between the two approaches from the point of view of computational costs. However, as an example, consider the results shown in figures 1 and 2. In this case

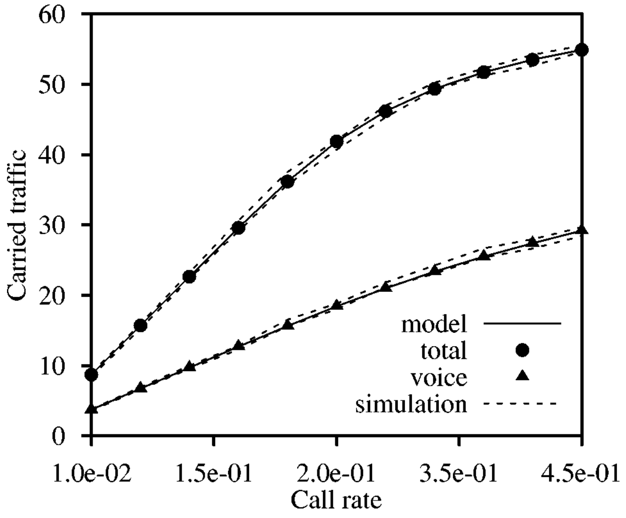

Figure 1. Basic scenario: simulation and analytical results for the average total carried traffic, and average carried voice traffic.

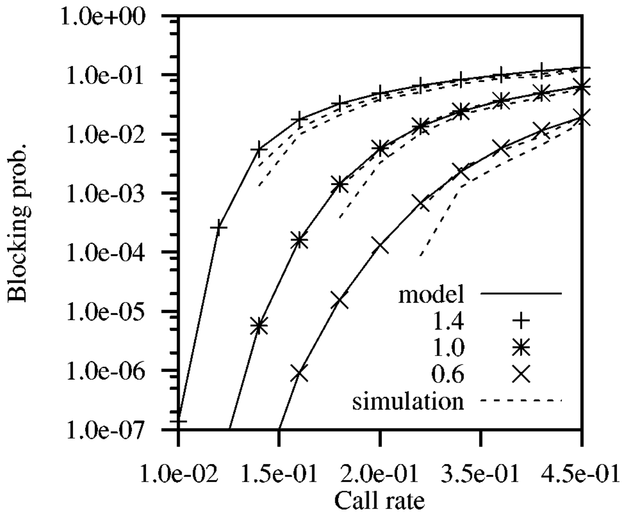

Figure 2. Basic scenario: simulation and analytical results for the average voice blocking probability.

the CPU time needed for the computation of the analytical solution is about 20% less than the cost of simulation experiments.

The first validation results are presented in figure 1, where we plot curves of the average carried traffic (number of occupied channels for both voice and data calls) and of the average voice carried traffic. Obviously, the vertical distance between the two curves gives the average carried data traffic. Analytical results are presented as two solid line curves (one curve for each performance parameter), while simulation results are shown as two pairs of curves (one pair for each performance parameter), giving the upper and lower limits of confidence intervals.

Figure 2 shows simulation and analytical results for the average blocking probability of new voice calls (results for the blocking probabilities of data calls are not shown, but they are quite similar to those presented for voice calls).

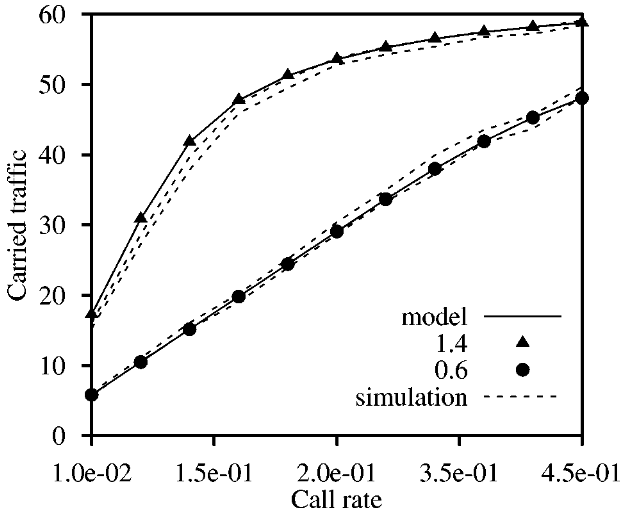

Figure 3. Basic scenario: simulation and analytical results for the average total carried traffic, with $\alpha = 1.4$ and $\alpha = 0.6$.

The curves presented in figures 1 and 2 clearly indicate that the approximations introduced in the model development do not alter significantly the numerical results with respect to those obtained with a detailed simulation of the system. This is a first indication of the accuracy of the analytical model.

However, in order to further expand the model validation, we also consider situations where the incoming and outgoing handover flows are not balanced; this could be expected to induce non marginal perturbations on the accuracy of the analytical model. In particular, we assume that the relation between the average incoming and outgoing handover flows is

$$f_{\text{in}} = \alpha f_{\text{out}},$$

where, of course, f_{in} and f_{out} respectively represent the incoming and outgoing handover flows. This kind of assumption can account for the particular type of traffic that a cell experiences in some period of time. For example, considering a cell covering the business district of a city, during the morning rush hour it can be expected that the incoming handover flow is significantly larger than the outgoing handover flow; thus the value of the parameter α should be set greater than 1. The opposite could be said in the evening rush hour period.

Figure 3 shows simulation and analytical results for the total carried traffic in the cell with $\alpha = 1.4$ and $\alpha = 0.6$. Figure 4 presents curves of the average voice call blocking probability in the same conditions.

It can be observed that the analytical model performance predictions are still quite accurate in comparison with simulation, although some loss of accuracy with respect to the former case can be noted.

As a preliminary conclusion we can thus state that the validation of the analytical results against results produced by quite a detailed simulation program was successful: the model can be considered to be accurate with respect to the experimented sets of values of system parameters.

3.2. Performance evaluation

3.2.1. Class A and B services

In order to show how the approximate model can be used to study the performance of the cellular mobile communication network, we now illustrate some results that are obtained with different configurations of the cell under investigation.

In most of the figures of this section, numerical results explore variations with respect to the basic scenario that was introduced in the previous section, and used for the model validation.

In figures 5 and 6 we compare the average carried load and the voice and data blocking probabilities when the number of channels required for a data call increases from $C_{\rm d} = 4$ to $C_{\rm d} = 8$. All other parameters are kept as in the basic scenario introduced in the previous section.

Increasing the number of channels allocated to each data call from $C_{\rm d} = 4$ to $C_{\rm d} = 8$ while keeping the same call rate significantly increases the data offered traffic, and also the data carried traffic when the channel occupancy is low. Differences in the data carried traffic tend to become smaller for higher call rates, and eventually the data carried traffic with $C_{\rm d} = 8$ becomes smaller than with $C_{\rm d} = 4$ for extremely high call rates (not shown). The carried voice traffic instead is larger when $C_{\rm d} = 4$, due to the larger average number of free channels in the cell. However, the total carried traffic always is larger when $C_{\rm d} = 8$.

Moreover, as expected, the blocking probabilities for voice and data calls (figure 6) with $C_{\rm d} = 8$ are much larger than with $C_{\rm d} = 4$. In particular, the data call blocking probabilities quickly reach unacceptable values.

In order to mitigate the difference in blocking probability between voice and data calls, a threshold $T_{\rm v}$ can be introduced on the number of free channels necessary for the acceptance of a new voice call. By so doing, if the number of busy channels is greater than $N - T_{\rm v}$, no new voice call can be accepted. The performances achievable in this case, with $T_{\rm v} = 0, 4, 8$, can be observed in figures 7 and 8, that report, respectively, the average carried traffic (voice, data, and total) with $C_{\rm d} = 8$, and the voice and data

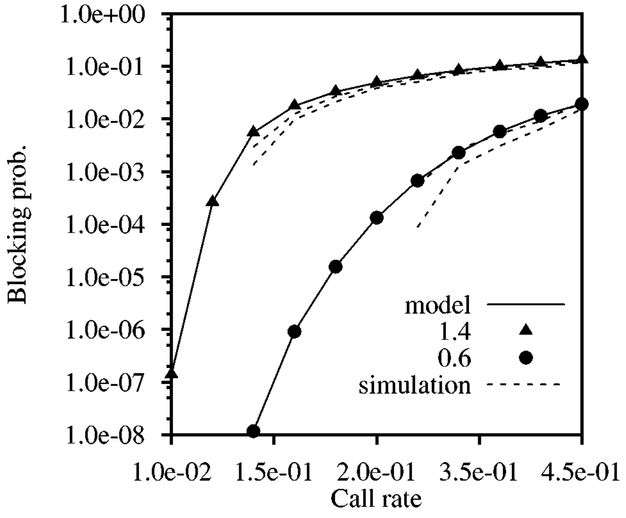

Figure 4. Basic scenario: simulation and analytical results for the average voice blocking probability, with $\alpha = 1.4$ and $\alpha = 0.6$.

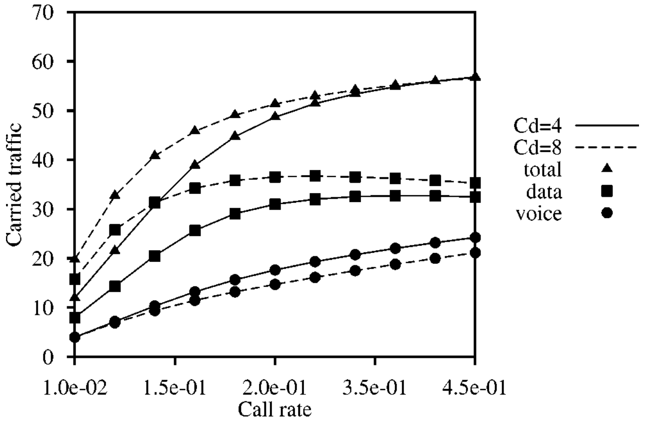

Figure 5. Average carried traffic (voice, data, and total) with $C_{\rm d} = 4$ and $C_{\rm d} = 8$.

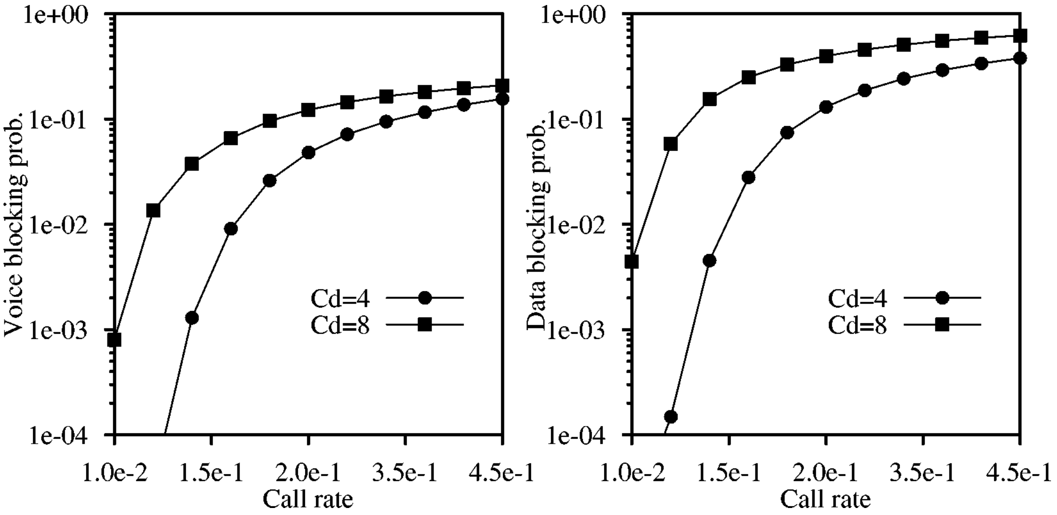

Figure 6. Voice and data call blocking probabilities with $C_{\rm d} = 4$ and $C_{\rm d} = 8$.

call blocking probabilities. The proposed approach leads to better performance for data traffic, at the expense of a performance loss for voice traffic.

3.2.2. Class C services

We now explore the effect induced on the system performance by the introduction of class C traffic, which is carried by multimedia connections comprising one voice and one data component. In the results we show below, the percentage of call requests referring to the three different classes of service is 60% for voice calls, and 20% each for data as well as multimedia calls; a data connection requires $C_\mathrm{d} = 8$ channels.

In particular, we consider in this case the impact on performance indices of the reservation of a number of channels for handover calls (C_h). This is a common approach to favor the successful completion of call handovers, specially under heavy traffic conditions; when the number of busy channels in the cell is greater than C_h, new calls (whether voice, data, or multimedia) are blocked.

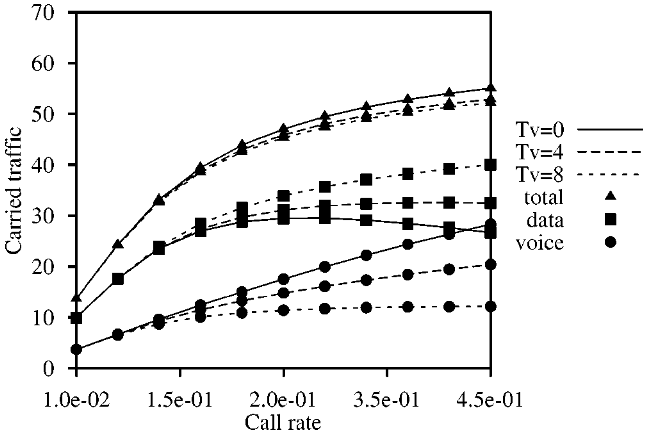

Figure 7. Average carried traffic (voice, data, and total) with $C_\mathrm{d} = 8$ and $T_\mathrm{v} = 0, 4, 8$.

In figure 9 we plot curves of the handover failure probability for voice and data calls, with $C_\mathrm{h} = 0, 4, 8$. As expected, the handover failure probability decreases with the increase of C_h. Note that this decrease is significantly different for voice and data connections: whereas for voice connections a large decrease is observed when C_h is increased from 0 to 4, and only a further minor improvement is obtained with $C_\mathrm{h} = 8$, for data calls the largest gain is achieved by increasing C_h from 4 to 8. This is due to the fact that data connections use 8 channels, and thus they gain little from the availability of 4 channels reserved to handovers.

In the same scenario, figure 10 reports the curves of the probabilities that a multimedia call handover either completely fails, or results in a decoupled multimedia call, with $C_\mathrm{h} = 0, 4, 8$. Recall that, as we noted before, a multimedia call must be decoupled when a handover is requested toward a cell in which the number of free channels is not sufficient for the acceptance of both components ($n > N - C_\mathrm{d} - 1$), but sufficient for the acceptance of just the voice component ($n \leqslant N - 1$). Instead, the handover completely fails if no channel is free ($n = N$). With PH_dec we indicate the probability that a multimedia call is decoupled, while with PH_r we denote the probability that the handover request fails.

Also in this figure we see that the effect of the reservation of channels to handovers is rather different on PH_dec and PH_r; the reduction in the handover failure probability is significant when C_h is increased from 0 to 4, and only a further very minor improvement is obtained with $C_\mathrm{h} = 8$. On the contrary, as regards the decoupling probability, very little is gained by letting $C_\mathrm{h} = 4$, and even the gain obtained with $C_\mathrm{h} = 8$ is not extraordinary. The justification of these results lies again in the difference between the number of channels necessary to serve the data and the voice components.

Decoupled multimedia calls are recombined as soon as enough free channels are found within the cell; the curves

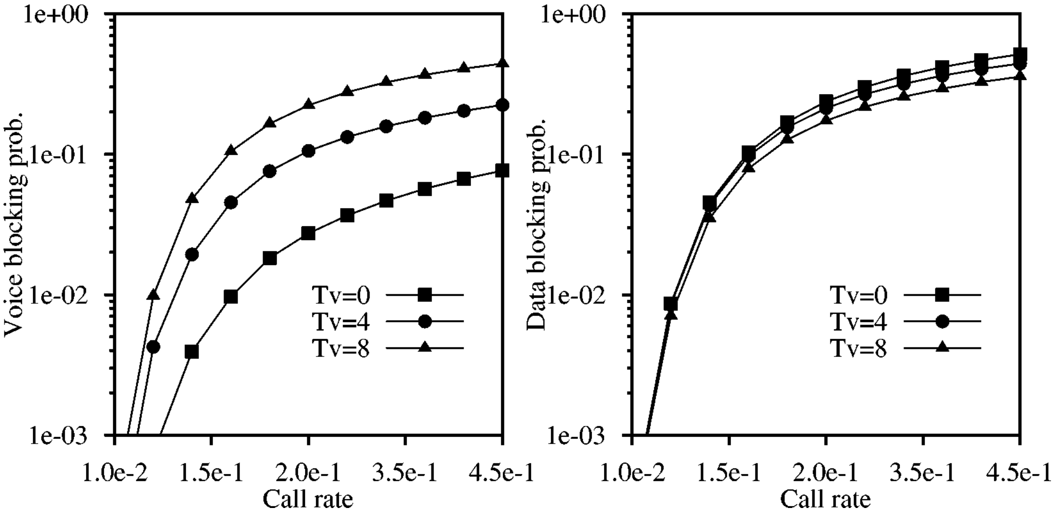

Figure 8. Voice and data call blocking probabilities with $C_\mathrm{d} = 8$ and $T_\mathrm{v} = 0, 4, 8$.

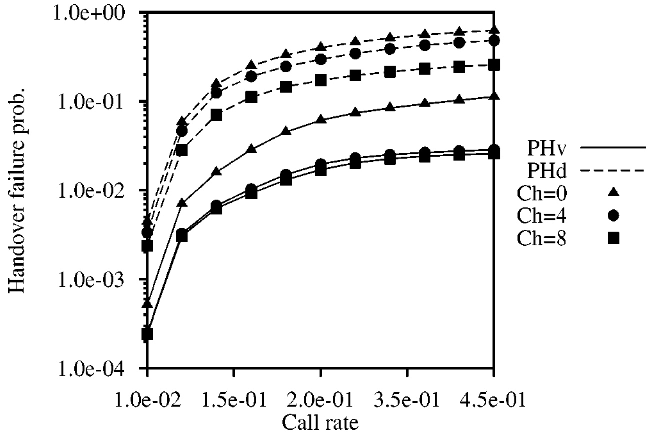

Figure 9. Voice and data handover failure probabilities with three classes of service, $C_{\rm d} = 8$ and $C_{\rm h} = 0, 4, 8$.

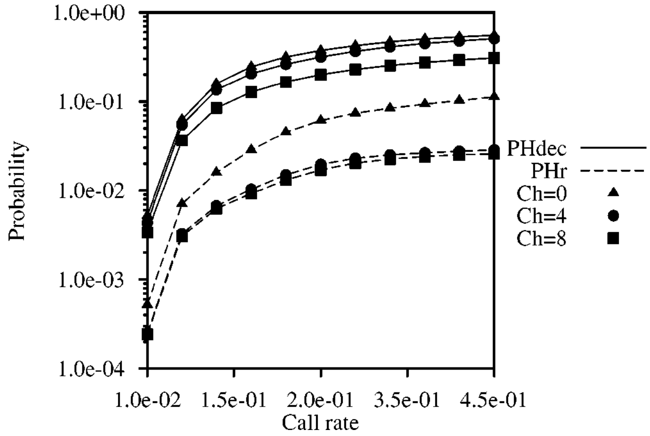

Figure 10. Decoupling and handover failure probabilities for multimedia calls with three classes of service, $C_{\rm d} = 8$ and $C_{\rm h} = 0, 4, 8$.

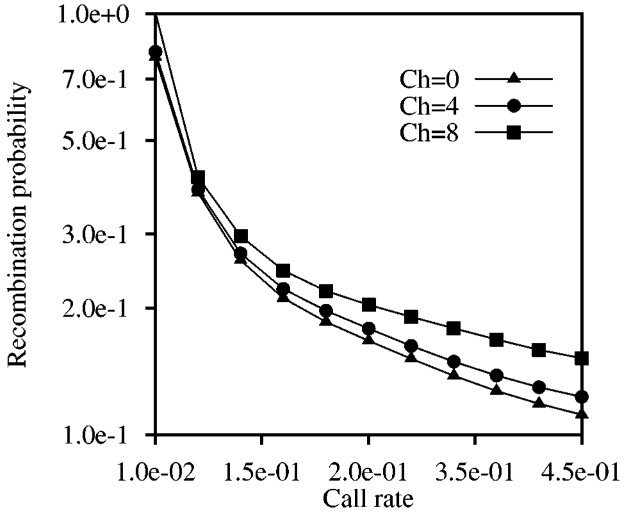

Figure 11. Recombination probability for multimedia calls with three classes of service, $C_{\rm d} = 8$ and $C_{\rm h} = 0, 4, 8$.

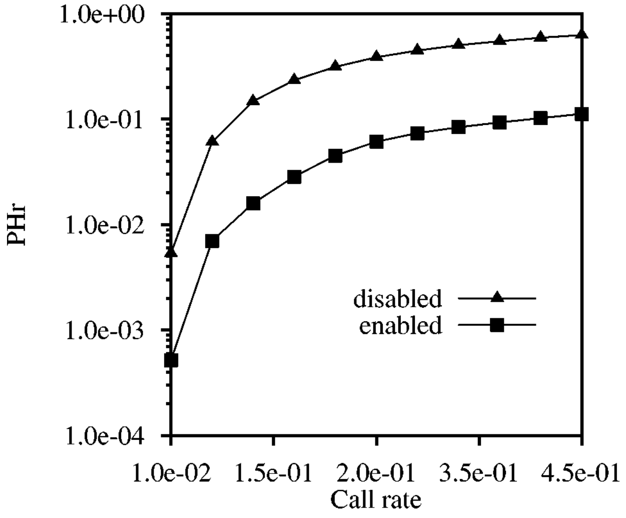

Figure 12. Handover failure probability for multimedia calls with three classes of service, $C_{\rm d} = 8$, $C_{\rm h} = 0$, and $r_{\max} = 0$, $C_{\rm d}$ (decoupling disabled or enabled).

reporting the probability that a decoupled multimedia call is recombined are presented in figure 11. For growing input traffic this probability obviously decreases, because of the smaller probability of finding $C_{\rm d}$ free channels in the cell before the completion of the call dwell time. The reservation of a subset of channels to handovers in this case has rather a beneficial impact, specially at high loads.

In order to assess the effectiveness of the multimedia decoupling approach, we finally show in figure 12, for $C_{\rm d} = 4$, $C_{\rm h} = 0$, and $r_{\max}$ equal to either $C_{\rm d}$ or 0, the probability that a multimedia call handover request is refused in the two cases in which a multimedia call can be decoupled ($r_{\max} = C_{\rm d}$), or is considered as an atomic entity ($r_{\max} = 0$). Results clearly indicate that the decoupling approach allows the reduction of the handover failure probability by a factor between 5 and 10, depending on the cell traffic load.

4. Conclusions

In this paper we have illustrated the development of an approximate Markovian model for a communication network providing integrated services to a population of mobile users, and we have presented performance results to both validate the approximate analytical approach, and assess the quality of the services offered to the end users.

The cellular system is assumed to provide three classes of service: the basic voice service, a data service with bit rate higher than the voice service, and a multimedia service with one voice and one data component.

In order to improve the overall network performance, some channels can be reserved to handovers, and multimedia calls that cannot complete a handover are decoupled, by transferring to the target cell only the voice component and suspending the data connection until a sufficient number of channels becomes free.

The analytical model is based on continuous-time multidimensional birth–death processes, and it is focused on one of the cells in the network only. The model solution is obtained with standard approaches for the solution of large Markovian systems, exploiting the sparseness of the infinitesimal generator.

Numerical results demonstrate the accuracy of the approximate model, as well as the effectiveness of the multimedia call decoupling approach, that is a novel contribution of this work.

References

[1] R. Beraldi, S. Marano and C. Mastroianni, Performance of a reversible hierarchical cellular system, Wireless Networks 4(1) (1997).

[2] D. Hong and S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures, IEEE Transactions on Vehicular Technology 35(3) (August 1986).

[3] S. Kumar and D.R. Vaman, An access protocol for supporting multiple classes of service in a local wireless environment, IEEE Transactions on Vehicular Technology 45(2) (1996).

[4] K.K. Leung, W.A. Massey and W. Whitt, Traffic models for wireless communication networks, IEEE Journal on Selected Areas in Communications 12(8) (October 1994)

[5] Y. Lin, Modeling techniques for large-scale PCS networks, IEEE Communication Magazine 35(2) (February 1997).

[6] S. Marano and C. Mastroianni, A hierarchical network scheme for multilayered cellular systems, in: *Proc. of VTC 1997*, Phoenix, AZ (May 1997).

[7] P. Narasimhan and R.D. Yates, A new protocol for the integration of voice and data over PRMA, in: *Proc. of PIMRC* (1995).

[8] R. Steele and M. Nofal, Teletraffic performance of microcellular personal communication networks, IEE Proceedings I 139(4) (August 1992).

[9] S. Tekinay, B. Jabbari and A. Kakaes, Modeling of cellular communication networks with heterogeneous traffic sources, in: *Proc. of ICUPC* (1993).

Marco Ajmone Marsan is a Full Professor at the Electronics Department, Politecnico di Torino, Italy. He holds a Dr. Ing. degree in electronic engineering from Politecnico di Torino, and a Master of Science from the University of California, Los Angeles. From November 1975 to October 1987 he was at the Electronics Department of Politecnico di Torino, first as a Researcher, then as an Associate Professor. From November 1987 to October 1990 he was a Full Professor at the Computer Science Department, University of Milan, Italy. During the summers of 1980 and 1981 he was with the Research in Distributed Processing Group, Computer Science Department, UCLA. During the summer of 1998 he was an Erskine Fellow at the Computer Science Department of the University of Canterbury in New Zealand. He has coauthored over 200 journal and conference papers in the areas of Communications and Computer Science, as well as the two books “Performance Models of Multiprocessor Systems” published by the MIT Press and “Modelling with Generalized Stochastic Petri Nets” published by John Wiley. He received the best paper award at the Third International Conference on Distributed Computing Systems in Miami, FL, in 1982. His current interests are in the fields of performance evaluation of communication networks and their protocols. M. Ajmone Marsan is a Fellow of IEEE.

Salvatore Marano received his degree in electronics engineering from the University of Rome, in 1973. In 1976 and 1977, he worked on the design of optical digital fiber systems at the ITT laboratory in Leeds, UK. Since 1979 he is Professor of Telecommunications Networks at the University of Calabria. His present research interests involve performance evaluation in mobile communications systems, random access in mobile radio networks and congestion control in ATM networks.

Carlo Mastroianni received his degree in computer engineering in 1995, and his Ph.D. degree in communications engineering in 1999, both from the University of Calabria. His research interests include terrestrial and satellite communications systems and performance evaluation in mobile networks.

Michela Meo received the Dr. Ing. degree in electronic engineering in 1993 and the Ph.D. degree in electronic and telecommunication engineering in 1997, both from Politecnico di Torino, Italy. Since then she has been working in the Telecommunication Networks Group of Politecnico di Torino. During her Ph.D. studies she was for one year at the Computer Science Department of the University of California in Santa Barbara. Her research interests are in the field of performance evaluation of communication networks with a particular focus on wireless systems.