WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2001; **1**:257–268 (DOI: 10.1002/wcm.17)

Improving call admission control procedures by using hand-off rate information

Novella Bartolini*,†

Dip. di Informatica, Sistemi e Produzione
Università di Roma "Tor Vergata"
Rome, Italy

Imrich Chlamtac

Center for Advanced Telecommunications
Systems and Services (CATSS)
University of Texas at Dallas
Dallas, Texas
U.S.A.

Summary

This paper introduces a general decision model, in the shape of a Markov Decision Process, as an instrument to analytically compare the behavior of call admission control policies. This approach allows the study of a wide class of policies, including well-known pure stationary as well as randomized policies, in a way that explicitly incorporates the dependency between the hand-off rate and the system state, assuming that the hand-off rate arriving to a cell is proportional to the occupancy level of the adjacent cells. In particular, some well-known non-preemptive prioritization schemes are analyzed, including the Cutoff Priority Policy (CPP), which consists of reserving a number of channels for the high priority requests stream. Using our analytical approach, we prove the optimality of CPP within the analyzed class.

KEY WORDS

call admission control policies
semi-Markov decision models
QoS optimization

1. Introduction

The current trend in cellular networks is exhibited in a reduced cell size to accommodate more mobile users in a given geographical area. The reduction leads to increased spectrum-utilization efficiency, but also results in more frequent hand-offs and makes guaranteed connection level QoS more difficult to achieve. Since it is impractical to completely eliminate hand-off drops, the next best alternative is a *probabilistic* guarantee on quality.

The rate of hand-off calls in a given cell depends on the number of calls in progress in the adjacent cells. Assuming that the subscriber mobility remains unchanged, an increase in the number of calls in the adjacent cells is naturally likely to increase the rate at which calls are handed off to the given cell, or in other words the hand-off rate is a function of the system

*Correspondence to: Novella Bartolini, Dip. di Informatica, Sistemi e Produzione, Università di Roma "Tor Vergata", Via Orazio Raimondo 18, 00173 Roma, Italy.
†E-mail: novella@uniroma2.it

state. On the other hand, the outgoing hand-off rate from a cell varies as a linear function of the number of calls in progress in the cell.

Several call admission policies have already been proposed and analytical formulas for the most important QoS parameters have been given. A comparison between the behavior of a few different schemes has occasionally been introduced by means of simulative results [1, 2], while analytical comparisons are made only in very few works [3] and with a very narrow class of call admission schemes.

In this paper we want to introduce a general decision model as an instrument to analytically compare the behavior of such schemes.

One of the novelties of this model is that it explicitly incorporates the dependency between the hand-off rate and the system state (in terms of number of calls in progress), and therefore can be expected to be more accurate than models based on average behavior.

Further, this model allows the analysis of call admission policies that enable queueing of hand-off requests when there is no available channel.

By means of this decision model, we search for an access control policy that gives high priority service to the hand-off requests without running the risk of compromising the whole traffic because of an insufficient consideration of initial attempts of connections.

Most of the recently proposed call admission control schemes can be studied through the decision model introduced in this paper.

An optimization analysis, using an objective function in the form of a linear combination of the loss probabilities of the two streams of arriving requests, is carried out.

The main contribution of this paper is the analytical proof of the optimality of a *cutoff priority policy* (CPP) [1, 3–7] when the objective function gives higher priority to the hand-off stream when queueing of requests is not allowed.

Under CPP, priority to hand-off calls is ensured by reserving a certain number of channels, also known as *guard channels*. According to CPP, an initial attempt request is accepted only if the total number of calls in progress, regardless of their type, is below a cutoff value and a free channel is available.

This result has an immediate practical application because the optimal cutoff value can be easily computed once known few statistic parameters defining the traffic of requests. These parameters are used to formulate the analytical models that can be solved by means of very commonly used methods of operations research. The originality of the results comes from the observation that in literature, other comparisons between access policies are either based on simulations [1, 2] or, when analytical, they are limited to few policies [3].

The paper is organized as follows. In Section 2 the continuous-time Markov decision model is described. The main procedure for its uniformization and discretization is introduced and the optimality of CPP is proved when queueing of requests is not allowed. In Section 3 the formulas of the most important quality of service parameters are illustrated and some numerical results that confirm the analytical results that were achieved in the previous sections are introduced. Section 4 concludes the paper with some final remarks.

2. Analytical Model

Since, in the most common admission policies considered in literature works, the decision of whether to accept or refuse a certain call is based on the number of ongoing calls in the given cell, it seems natural that the state of our model represents this measure of the occupancy level. Our traffic model consists of a Markov decision process in which a single cell is modeled as a service center with C servers corresponding to the available frequency channels. Arriving users, representing requests of connection to the base station, belong to two priority classes: high priority for hand-off calls and low priority for initial access requests.

Knowing the number of calls in the neighbor cells gives us some idea of how many calls we can expect will be handed off in the next unit of time. On the other hand, keeping track of this information can significantly increase the size of the state space. If we assume some uniformity in the system, for what mainly concerns the geographic environment of the cells and the mobility of the subscribers, then the number of calls in the current cell gives a good indication of the number of calls in neighbor cells.

As often happens in literature works, arrivals are assumed to be generated according to Poisson processes.

The arrival rate of new requests of connection, that will be served with low priority, will be λ_L while we can assume a hand-off rate proportional to the number i of busy channels, i.e., $i\lambda_H$ that will be treated with high priority, where λ_H is a measure proportional to the hand-off rate per ongoing call from an adjacent cell to the considered cell.

Blocked initial requests are lost, while a blocked hand-off call can wait in the hand-off queue for a channel of the new cell by continuing to use a channel of the previous one. The queueing scheme is briefly described as follows. No initial access request is granted a channel before the hand-off requests in the queue are served. When an MS reaches the overlapping region between two adjacent cells, also called *hand-off region* (HR), and no free channels are available in the destination cell, the call remains queued until either an available channel in the new cell is found, or the MS abandons the HR before a channel becomes available, thus causing the forced termination of the hand-off call and its departure from the queue. In the case of high demand for hand-off, hand-off calls will be denied queueing due to the limited size of the hand-off queue. The queueing device has a finite number of places M_H.

In this model a call may exit from the control of the base station in different ways:

1. The conversation is completed (it may happen even with a queued hand-off request, which thus abandons the queue).
2. The MS goes out of cell.
3. A waiting hand-off call is terminated because it is not served before passing the HR, thus it abandons the queue.

The distribution of these events is supposed to be exponential with parameters μ_1, μ_2 and μ_3 respectively. Figure 1 shows our model configuration.

The switches represent the two actions (accept or refuse) that can be chosen by the access control policy when a call arrives. A refused call is definitively lost, regardless of its priority class.

The evolution of a call as a consequence of the control policy and possible movements of the MS is represented by the state model of Figure 2.

The double rounded states are the decision steps during the lifetime of a call.

A call generated within a cell can be accepted or not, according to a certain control policy. If accepted, it can be completed before the MS goes out of the

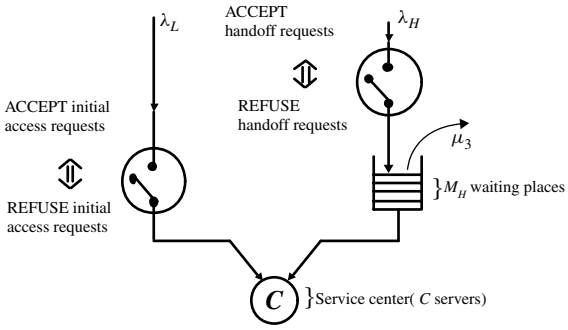

Fig. 1. System configuration.

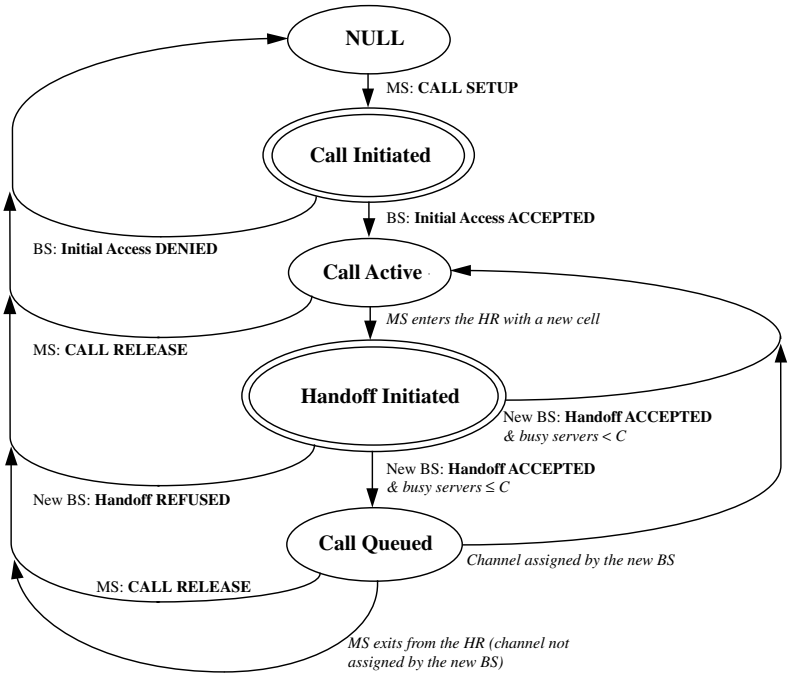

Fig. 2. Call state model.

cell, otherwise the hand-off procedure is initiated. The base station of the destination cell may decide to refuse the hand-off request, to provide it with a new channel, if available, or to put it into the hand-off queue while waiting for a new available channel. While in the hand-off queue, the call continues to use a channel of the old cell. During the time the call spends in the queue, the user may exit from the HR before obtaining a new channel, terminating the hand-off procedure unsuccessfully, or may also decide to conclude the call.

2.1. A General Decision Model

The memoryless property of all probability distributions in a Markov process makes it impossible to represent policies for which the behavior of the system strictly depends on its past history, unless we use several different states to represent the same occupancy level. Each state $\mathbf{s}$, belonging to the finite state space E of the Markov decision process, can be defined through a couple of indexes (i, t), where i represents the number of busy servers, while t is a *state tag*, with $t \in \{1, 2, \ldots, n\}$ introduced to allow different decisions in correspondence with the same occupancy level i. Let us consider the following set of possible actions that can be undertaken at each state of the process:

a_1: accept requests belonging to both streams.
a_2: deny access to initial attempts.
a_3: deny access to hand-off calls.
a_4: deny access to both streams of requests.

We define the function $n(\mathbf{s})$ as follows. Given $\mathbf{s} \in E$, $n(\mathbf{s})$ is the occupancy level characterizing the state $\mathbf{s}$. Thus $n(\mathbf{s})$ is the sum of the number of busy channels and the number of busy places in the queueing device. If $\mathbf{s} = (i, t)$, then $n(\mathbf{s}) = i$. If $C \le n(\mathbf{s}) < C + M_H$, the set of feasible actions reduces to a_2, a_4 because we have no queueing device for initial access requests and if $n(\mathbf{s}) = C + M_H$ the only feasible action is a_4.

Consider now a partition of the set E into classes E_i with the following properties: $E = \bigcup_{i=0}^{C} E_i$, where $\{E_i = \mathbf{s} \in E, n(\mathbf{s}) = i\}$.

From any state $\mathbf{s} \in E_i$, a new request acceptance leads the system to any state $\mathbf{q}$ of the class E_{i+1}, denoted by $Succ(\mathbf{s})$. The choice of the next state among the members of this class follows a certain probability distribution $\pi^{+}_{\mathbf{sq}}$, where $\mathbf{q} \in Succ(\mathbf{s})$, with $\sum_{\mathbf{q} \in Succ(\mathbf{s})} \pi^{+}_{\mathbf{sq}} = 1$.

The transition rate from $\mathbf{s}$ to any state $\mathbf{q}$ of the class $Succ(\mathbf{s})$ is $\lambda(\mathbf{s}, a)\pi^{+}_{\mathbf{sq}}$, where

$$\lambda(\mathbf{s}, a) = \begin{cases} n(\mathbf{s})\lambda_H + \lambda_L & \text{if } a = a_1 \\ n(\mathbf{s})\lambda_H & \text{if } a = a_2 \\ \lambda_L & \text{if } a = a_3 \\ 0 & \text{if } a = a_4 \end{cases} \tag{1}$$

On the other hand, from any state $\mathbf{s} \in E_i$, the termination of a service, either due to call completion or to the MS movements outside the cell, brings the system to any state $\mathbf{k}$ of the class E_{i-1}, denoted by $Prec(\mathbf{s})$ with rate $n(\mathbf{s})(\mu_1 + \mu_2)\pi^{-}_{\mathbf{sk}}$, if $n(\mathbf{s}) \le C$ and $\{n(\mathbf{s})\mu_1 + C\mu_2 + [n(\mathbf{s}) - C]\mu_3\}\pi^{-}_{\mathbf{sk}}$ if $n(\mathbf{s}) > C$, with $\sum_{\mathbf{k} \in Prec(\mathbf{s})} \pi^{-}_{\mathbf{sk}} = 1$.

The transition diagram of the process is represented in Figure 3 which illustrates all the possible outgoing transitions from one state.

The transition probabilities matrix is decision dependent. It can be written as follows:

$$\begin{cases} p^{a}_{\mathbf{sk}} = \dfrac{\lambda(\mathbf{s}, a)\pi^{+}_{\mathbf{sk}}}{\lambda(\mathbf{s}, a) + n(\mathbf{s})(\mu_1 + \mu_2)} \\ \qquad \text{if } \mathbf{k} \in Succ(\mathbf{s}) \text{ and } 0 \le n(\mathbf{s}) \le C \\ p^{a}_{\mathbf{sk}} = \dfrac{\lambda(\mathbf{s}, a)\pi^{+}_{\mathbf{sk}}}{\lambda(\mathbf{s}, a) + n(\mathbf{s})\mu_1 + C\mu_2 + [n(\mathbf{s}) - C]\mu_3} \\ \qquad \text{if } \mathbf{k} \in Succ(\mathbf{s}) \text{ and } n(\mathbf{s}) > C \\ p^{a}_{\mathbf{sk}} = \dfrac{n(\mathbf{s})(\mu_1 + \mu_2)\pi^{-}_{\mathbf{sk}}}{\lambda(\mathbf{s}, a) + n(\mathbf{s})(\mu_1 + \mu_2)} \\ \qquad \text{if } \mathbf{k} \in Prec(\mathbf{s}) \text{ and } 0 \le n(\mathbf{s}) \le C \\ p^{a}_{\mathbf{sk}} = \dfrac{n(\mathbf{s})\mu_1 + C\mu_2 + [n(\mathbf{s}) - C]\mu_3\pi^{-}_{\mathbf{sk}}}{\lambda(\mathbf{s}, a) + n(\mathbf{s})\mu_1 + C\mu_2 + [n(\mathbf{s}) - C]\mu_3} \\ \qquad \text{if } \mathbf{k} \in Prec(\mathbf{s}) \text{ and } n(\mathbf{s}) > C \\ p^{a}_{\mathbf{sk}} = 0 \qquad \text{otherwise} \end{cases} \tag{2}$$

The parameters $\pi^{+}_{\mathbf{sq}}$, $\pi^{-}_{\mathbf{sq}}$ and the stationary state-decision associations can be adequately set to turn our general decision model into the models of CPP and of most of the well-known policies.

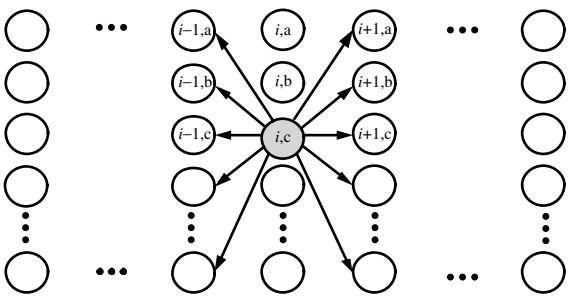

Fig. 3. Possible transitions from state (i, K).

CPP can be obtained by selecting $\pi^{+}_{\mathbf{sq}}$ and $\pi^{-}_{\mathbf{sq}}$ with $\mathbf{s} = (i_s, j_s)$ and $\mathbf{k} = (i_k, j_k)$, in the following way: $\pi^{+}_{\mathbf{sq}} = \pi^{-}_{\mathbf{sq}} = 1$ if $j_s = j_k = fixed\_tag$, for any $fixed\_tag$ else $\pi^{+}_{\mathbf{sq}} = \pi^{-}_{\mathbf{sq}} = 0$, and taking the decision a_1 for all the state $\mathbf{s}$ with i_s lower than the cutoff value T, the decision a_2 if $T \leq i_s < C$ and the decision a_4 if $i_s = C$.

The related model is shown in Figure 4. Access control policies of a randomized kind can also be obtained by allowing a non-deterministic decision in one or more states. In Reference [7] we show how common policies proposed by other authors [1–3, 8] can be viewed as particular instances of the general decision model we are analyzing.

2.2. Optimization within the General Class

The optimization procedure can be summed up as follows.

- The continuous-time process introduced in the previous section is uniformized and discretized in order to apply discrete-time optimization methods. The objective function is introduced with direct application to this discrete-time model.
  Then the analysis of the discretized model follows.
- We analytically prove that there exists an optimal deterministic stationary policy, i.e. not randomized, for which the decision chosen in correspondence to each state is always the same, independently of the particular instant of time.
- Moreover, we prove the existence of an optimal policy for which the optimal decision does not depend on the state tag, but on its occupancy level only.
- The optimality of CPP is proved through the analysis of the structural properties of the optimal cost function.

From now on, we will refer to X_n as to the state of the process at the moment of the nth transition, and with $u_n(X_n)$ to the particular decision chosen in the set $\{a_1, a_2, a_3, a_4\}$.

The Markov chain $\{X(t)\}$ related to the process described above is continuous-time. The dwell time of the process in each state is exponentially distributed with density $\phi(\mathbf{s}, a)e^{-\phi(\mathbf{s},a)t}$. The parameter $\phi(\mathbf{s}, a)$ is the total outgoing rate from a state in which the decision a has been chosen, and depends both on the decision a and on the state $\mathbf{s}$. The set of rates that characterizes the process is bounded by the maximum outgoing rate which is less than $C(\mu_1 + \mu_2) + M_H(\mu_1 + \mu_3) + C\lambda_H + \lambda_N$. Hence, we can conclude that the process is uniformizable.

Adding dummy transitions from states to themselves, a uniform Poisson process can be constructed which governs the epochs at which transitions take place. The uniformization technique transforms the original continuous-time Markov chain with not identical transition times into an equivalent continuous-time Markov process in which the transition epochs are generated by a Poisson process at uniform rate. The transitions from state to state are described by a (discrete time) Markov chain that allows for fictitious transitions from a state to itself. The uniformized Markov process $\{\widehat{X}(t)\}$ is probabilistically identical to the not uniform $\{X(t)\}$ [9–11].

The theory of discrete Markov processes can be used to analyze the discrete-time embedded Markov chain of the uniformized model. Let us assume uniform rate $\Lambda = C(\mu_1 + \mu_2) + M_H(\mu_1 + \mu_3) + C\lambda_H + \lambda_N$.

The transition probabilities of the uniformized process are:

$$
p^{a}_{\mathbf{sk}} = \begin{cases}
\dfrac{\lambda(\mathbf{s}, a)\pi^{+}_{\mathbf{sk}}}{\Lambda} & \text{if } \mathbf{k} \in Succ(\mathbf{s}) \\
\dfrac{n(\mathbf{s})(\mu_1 + \mu_2)\pi^{-}_{\mathbf{sk}}}{\Lambda} & \text{if } \mathbf{k} \in Prec(\mathbf{s}) \text{ and } 0 \leq n(\mathbf{s}) \leq C \\
\dfrac{\{n(\mathbf{s})\mu_1 + C\mu_2 + [n(\mathbf{s}) - C]\mu_3\}\pi^{-}_{\mathbf{sk}}}{\Lambda} & \text{if } \mathbf{k} \in Prec(\mathbf{s}) \text{ and } n(\mathbf{s}) > C \\
\dfrac{\{\Lambda - [n(\mathbf{s})(\mu_1 + \mu_2) + \lambda(\mathbf{s}, a)]\}}{\Lambda} & \text{if } \mathbf{k} = n(\mathbf{s}) \text{ and } n(\mathbf{s}) \leq C \\
\dfrac{\{\Lambda - [n(\mathbf{s})\mu_1 + C\mu_2 + (n(\mathbf{s}) - C)\mu_3 + \lambda(\mathbf{s}, a)]\}}{\Lambda} & \text{if } \mathbf{k} = n(\mathbf{s}) \text{ and } n(\mathbf{s}) > C \\
0 & \text{if otherwise}
\end{cases} \tag{3}
$$

In order to give higher priority to the hand-off stream, rather than to the initial access stream, we introduce a cost function which assigns different penalties to the loss of the two kinds of requests. The system is forced to pay a high penalty H if a hand-off call is refused or if it is firstly queued but no channel is assigned before the MS exits from the HR. If service is denied to an initial attempt of access, the system pays a lower

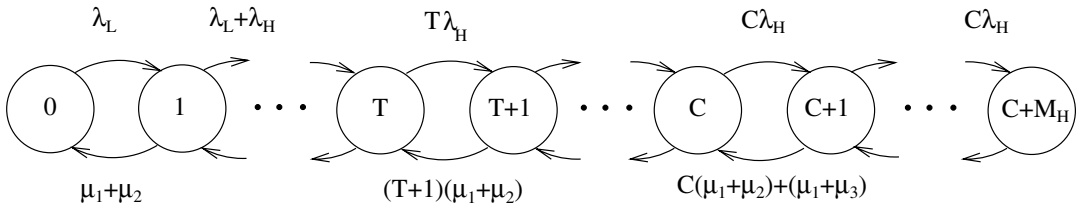

Fig. 4. State diagram of CPP.

penalty $L < H$. The penalty is not paid all the times the system enters a state in which the chosen decision is to refuse a request. In the uniformized process, all the penalties must be weighted with the probabilities that the event which causes a penalty actually occurs. We can define the cost function in the following way:

$$\hat{r}(\mathbf{s}, a) = \frac{H \max\{0, [n(\mathbf{s}) - C]\mu_3\}}{\Lambda} + \hat{r}'(\mathbf{s}, a) \tag{4}$$

$$\hat{r}'(\mathbf{s}, a) = \begin{cases} 0 & \text{if } a = a_1 \\ \dfrac{L\lambda_L}{\Lambda} & \text{if } a = a_2 \\ \dfrac{H\lambda_H \min\{n(\mathbf{s}), C\}}{\Lambda} & \text{if } a = a_3 \\ \dfrac{L\lambda_L + H\lambda_H \min\{n(\mathbf{s}), C\}}{\Lambda} & \text{if } a = a_4 \end{cases} \tag{5}$$

where the decisions a_1 and a_3 are not feasible if $n(\mathbf{s}) \geq C$.

The objective is to determine an optimal policy for admitting customers so as to minimize the expected long run average cost. Using the previous notation and denoting with $N(T)$ the number of transitions being completed at time T, the long-run average cost function can be written as

$$\lim_{T\to\infty} \frac{E\left\{\sum_{n=0}^{N(T)} r[X_n, u_n(X_n)|X_0 = i]\right\}}{T} \tag{6}$$

We refer to Reference [12] for the proof that the optimization procedures can be applied directly to the discrete-time Markov process described by the embedded Markov chain of the uniformized one. The optimal policy is the same for the initial, the uniformized and the discretized process, while the optimal values of the objective functions only differ in a constant factor.

The most important results of the theory of discrete-time Markov decision processes can be applied to the discretized model. In particular, observing the shape of the transition diagram of Figure 3, it can be affirmed, without loss of generality, that the decision model can be restricted to include the only processes with no transient states and with only one communicating class, that is to the only unichain processes. Refer to S as to the finite set of all feasible couples of the kind (*state, decision*). The unichain assumption, together with the finiteness of S implies the existence of a unique stationary state probability distribution which is independent of the initial state of the process. The existence of a stationary optimal policy allows us to conclude that an optimal solution can be expressed through a vector $\mathbf{D}^*$ whose generic component $D^*_{\mathbf{sa}}$ represents the stationary probability that, in correspondence to the state $\mathbf{s}$, the system takes the decision a. We can write

$$D_{\mathbf{sa}} \geq 0 \text{ and}$$

$$\sum_{a\in A_{\mathbf{s}}} D_{\mathbf{sa}} = 1, \quad \mathbf{s} \in E$$

where $A_{\mathbf{s}}$ is the set of all actions that can be taken in state $\mathbf{s}$. The expected value of the cost function can now be expressed in the form

$$z = \sum_{(\mathbf{s},\mathrm{a})\in S} D_{\mathbf{s},\mathrm{a}}\, p_{\mathbf{s}} \hat{r}(\mathbf{s}, a) \tag{7}$$

where $p_{\mathbf{s}}$ denotes the stationary probability that the system is in the state $\mathbf{s}$, and the product $D_{\mathbf{sa}}\, p_{\mathbf{s}}$ represents the joined probability for the system to be in state $\mathbf{s}$ and contemporaneously to take the decision a. Substituting the expression of $\hat{r}(\mathbf{s}, a)$ given by Equation (4) into Equation (7) the following expression for the objective function can be obtained.

$$\begin{aligned} z = {} & H\frac{\lambda_H}{\Lambda} \sum_{\substack{(\mathbf{s},\mathrm{a})\in S \\ \wedge(\mathrm{a}=a_3|\mathrm{a}=a_4)}} p(\mathbf{s}, a) \min\{n(\mathbf{s}, C)\} \\ & + L\frac{\lambda_L}{\Lambda} \sum_{\substack{(\mathbf{s},\mathrm{a})\in S \\ \wedge(\mathrm{a}=a_2|\mathrm{a}=a_4)}} p(\mathbf{s}, a) \\ & + H \sum_{\substack{(\mathbf{s},\mathrm{a})\in S \\ \wedge(\mathrm{n}(\mathbf{s})>\mathrm{C})}} [n(\mathbf{s}) - C]\frac{\mu_3}{\Lambda} p(\mathbf{s}, a) \end{aligned} \tag{8}$$

Analyzing the topology of our transition diagram, we can notice the total absence of transient states that, together with the unichain assumption, gives a particular shape to the set of constraints of the linear programming problem related to our optimization procedure.

Denoting $x_{\mathbf{sa}} \triangleq D_{\mathbf{sa}} p_{\mathbf{s}}$, $(\mathbf{s}, a) \in S$, and recalling that $D_{\mathbf{sa}} = \dfrac{x_{\mathbf{sa}}}{p_{\mathbf{s}}} = \dfrac{x_{\mathbf{sa}}}{\sum_{\mathbf{j}\in A_{\mathbf{s}}} x_{\mathbf{ja}}}$, $a \in A_{\mathbf{s}}$, the linear programming problem becomes:

$$\text{Maximize}$$

$$\sum_{(\mathbf{s},\mathrm{a})\in\mathrm{S}} r(\mathbf{s}, a) x_{\mathbf{sa}}$$

$$\text{constrained to}$$

$$x_{\mathbf{sa}} \geq 0 \qquad (\mathbf{s}, a) \in S$$

$$\sum_{(\mathbf{s},\mathrm{a})\in\mathrm{S}} x_{\mathbf{sa}} = 1$$

$$\sum_{a\in A_{\mathrm{j}}} x_{\mathbf{ja}} = \sum_{(\mathbf{s},\mathrm{a})\in\mathrm{S}} p^{a}_{\mathbf{sj}} x_{\mathbf{sa}} \quad j \in E \tag{9}$$

Proposition 1. *The linear programming problem [Equation (9)] has an optimal deterministic solution.*

Proof. Thanks to the absence of transient states we conclude that the optimal solution $\mathbf{x}^0$ has the following property: $\sum_{a\in A_{\mathbf{s}}} x^0_{\mathbf{sa}} > 0$, $\forall \mathbf{s} \in E$. Thence $\mathbf{x}^0$ has at least $|E|$ strictly positive variables. Summing up all their related equations deriving from the set of positiveness constraints, we again find the normalization equation. We conclude the redundancy of one among the $|E| + 1$ remaining constraints. The operation research applied to linear programming problems proves the existence of an optimal base solution containing a number of positive variables at most equal to the number of non-redundant constraints. Without loss of generality we can suppose that $\mathbf{x}^0$ has this property. So we conclude that $\mathbf{x}^0$ contains at most $|E|$ positive variables. Having already stated that the number of positive variables is at least $|E|$ and at most $|E|$, and that $\sum_{a\in A_{\mathbf{s}}} x^0_{\mathbf{sa}} > 0$, $\forall \mathbf{s} \in E$, we conclude that for all $\mathbf{s} \in E$ there will be exactly a decision a for which $x^0_{\mathbf{sa}} > 0$. This leads to conclude a very important result which is the existence of a pure, not randomized, stationary optimal policy. This result gives us the possibility to further restrict our consideration to policies for which $D_{\mathbf{sa}} \in \{0, 1\}$.

The proof that CPP is optimal among all the policies described by the general model, when queueing of requests is not allowed, can be summed up as follows:

- The existence of an optimal policy for which the optimal decision does not depend on the state tag, but on its occupancy level only, is proved by means of the dynamic programming equation.
- The optimality of CPP is proved through the analysis of structural properties of the optimal cost function.

A first step towards the optimization of the average cost for the infinite horizon problem is the evaluation of the N-step optimal total discounted cost $V^{\alpha}_{N}(\mathbf{s})$. The discrete-time discount factor $\alpha < 1$, which corresponds to the continuous-time discount coefficient $\eta > 0$, related to the not uniformized process, is

$$\alpha = \frac{\Lambda}{\eta + \Lambda} \tag{10}$$

The optimal discounted cost function can be calculated with the following dynamic programming equation [13, 14]

$$V^{\alpha}_{K}(\mathbf{s}) = \min_{a\in A_{\mathbf{s}}} \left\{ \hat{r}(\mathbf{s}, a) + \sum_{\mathbf{z}\in\mathbf{E}} \alpha \hat{p}^{a}_{\mathbf{sz}} V^{\alpha}_{K-1}(\mathbf{z}) \right\} \tag{11}$$

$V^{\alpha}_{K}(\mathbf{s})$ is the minimum expected discounted cost that can be paid in K periods if the system starts with $n(\mathbf{s})$ customers, and a discount factor of α. Naming the arguments of the min function of Equation (11) with A_1, A_2, A_3 and A_4 when in correspondence of decisions a_1, a_2, a_3 and a_4 respectively and substituting the known expression of the cost function [Equation (4)], of the discount factor [Equation (10)] and of the transition probabilities [Equation (3)], the following equation is obtained for the total discounted cost. No queueing of request is now considered.

If $n(\mathbf{s}) \leq C$,

$$V^{\alpha}_{K}(\mathbf{s}) = \frac{1}{\Lambda + \eta} \min\{A_1, A_2, A_3, A_4\} \tag{12}$$

where A_1, A_2, A_3 and A_4 can be defined as follows

$$\begin{aligned} A_1 = & \sum_{\mathbf{l}\in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1 + \mu_2)\pi^{-}_{\mathbf{sl}} V^{\alpha}_{K-1}(\mathbf{l}) \\ & + \sum_{\mathbf{j}\in Succ(\mathbf{s})} (\lambda_L + \lambda_H n(\mathbf{s}))\pi^{+}_{\mathbf{sj}} V^{\alpha}_{K-1}(\mathbf{j}) \\ & + (C - n(\mathbf{s}))(\mu_1 + \mu_2 + \lambda_H) V^{\alpha}_{K-1}(\mathbf{s}) \end{aligned} \tag{13}$$

$$A_2 = \lambda_L L + \sum_{\mathbf{j}\in Succ(\mathbf{s})} n(\mathbf{s})\lambda_H \pi^+_{\mathbf{sj}} V^\alpha_{K-1}(\mathbf{j}) + \sum_{\mathbf{l}\in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1+\mu_2)\pi^-_{\mathbf{sl}} V^\alpha_{K-1}(\mathbf{l}) + \{[C-n(\mathbf{s})](\mu_1+\mu_2+\lambda_H)+\lambda_L\}V^\alpha_{K-1}(\mathbf{s}) \tag{14}$$

$$A_3 = n(\mathbf{s})\lambda_H H + \sum_{\mathbf{j}\in Succ(\mathbf{s})} \lambda_L \pi^+_{\mathbf{sj}} V^\alpha_{K-1}(\mathbf{j}) + \sum_{\mathbf{l}\in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1+\mu_2)\pi^-_{\mathbf{sl}} V^\alpha_{K-1}(\mathbf{l}) + \{[C-n(\mathbf{s})](\mu_1+\mu_2)+C\lambda_H\}V^\alpha_{K-1}(\mathbf{s}) \tag{15}$$

$$A_4 = n(\mathbf{s})\lambda_H H + \lambda_L L + \sum_{\mathbf{l}\in Prec(\mathbf{s})} n(\mathbf{s})(\mu_1+\mu_2)\pi^-_{\mathbf{sl}} V^\alpha_{K-1}(\mathbf{l}) + \{[C-n(\mathbf{s})](\mu_1+\mu_2)+C\lambda_H+\lambda_L\}V^\alpha_{K-1}(\mathbf{s}) \tag{16}$$

Proposition 2. *$\forall \boldsymbol{s}$ and $\boldsymbol{z}$, such that $n(\boldsymbol{s}) = n(\boldsymbol{z})$, $V^\alpha_K(\boldsymbol{s}) = V^\alpha_K(\boldsymbol{z})$, $\forall K \in \mathcal{N}$.*

Proof. By induction on the number of steps K (see Reference [7] for details).

Proposition 2 proves that each time the system has to choose one among the feasible decisions, the choice does not depend on the particular state in which the system is, but on its occupancy level only.

For this reason it can be defined the function $W(\cdot,\cdot)$ on the domain $\{0,1,\ldots,C+M_H\}\times\mathcal{N}$, with the following property: $W^\alpha(n(\mathbf{s}),K) = V^\alpha_K(\mathbf{s}) = V^\alpha_K(\mathbf{z})$. Using the now stated property, the expression of $W^\alpha(i,K)$ can be written as follows:

$$\begin{aligned}
&\text{If } i \le C,\\
&W^\alpha(i,K)\\
&= \frac{1}{\Lambda+\eta}\big[(C-i)(\mu_1+\mu_2+\lambda_H)W^\alpha(i,K-1)\\
&+ i(\mu_1+\mu_2)W^\alpha(i-1,K-1)\\
&+ (\lambda_L + i\lambda_H)W^\alpha(i,K-1)\\
&+ \lambda_L \min\{L, W^\alpha(i+1,K-1)-W^\alpha(i,K-1)\}\\
&+ i\lambda_H \min\{H, W^\alpha(i+1,K-1)-W^\alpha(i,K-1)\}\big]
\end{aligned} \tag{17}$$

Equation (17) shows that the choice of whether to accept or not a given request depends on the value of the increment of the cost function:

$$\Delta W^\alpha_k(i) = W^\alpha(i+1,K) - W^\alpha(i,K) \tag{18}$$

Proposition 3. *$W^\alpha(i,K)$ is not decreasing in i thus $0 \le \Delta W^\alpha_K(i)$.*

Proof. By induction on the number of steps K (see Reference [7] for details).

Proposition 4. *If $M_H = 0$, $W^\alpha(i,K)$ is also concave in the number of busy servers i.*

Proof. By induction on the number of steps K (see Reference [7] for details).

Since H and L are positive and $W^\alpha(i,K)$ is bounded above the geometric series

$$\frac{1}{\Lambda}\sum_{j=0}^{K}\alpha^j \max\{H,L\} \tag{19}$$

the sequence $\{W^\alpha(i,K)\}_{K=0}^{\infty}$ increases monotonically to a finite limiting value for each i and α. Hence, the limit $\lim_{K\to\infty} W^\alpha(i,K)$ exists. We let $W^\alpha(i) = \lim_{K\to\infty} W^\alpha(i,K)$. From Reference [12], it can be verified that $W^\alpha(i)$ is the *minimum infinite horizon discounted cost*.

The structural properties of monotony and concavity of $W^\alpha(i,K)$ are inherited by $W^\alpha(i)$ and imply the following proposition.

Proposition 5. *CPP is optimal under the total discounted cost criterion, for the infinite horizon problem, when $M_H = 0$.*

Proof. The optimal policy chooses the best action to take in each state with the following rule:

- High priority customers are accepted only if $\Delta W^\alpha(i) = W^\alpha(i+1) - W^\alpha(i) \le H$.
- Low priority customers are accepted only if $\Delta W^\alpha(i) = W^\alpha(i+1) - W^\alpha(i) \le L$.

Since $W^\alpha(i)$ is monotone and concave, the term $\Delta W^\alpha(i)$ is not decreasing, thence we can find integer values i_L and i_H such that

$$i_L = \arg\ \min\{\Delta W^\alpha(i) > L\}$$

$$i_H = \arg\ \min\{\Delta W^\alpha(i) > H\}$$

Therefore, the optimal policy regarding the decision to accept or refuse to serve requests of the initial access stream is

- Initial access admitted for $i < i_L$
- Initial access denied for $i \geq i_L$, (that is the already described CPP),

while since $i_H \geq i_L$ the decision of refusing the high priority calls is obviously discarded.

Theorems of equivalence [12] between a continuous-time Markov decision process and its discretization allows us to conclude that CPP based on the parameter i_L is optimal also for the initial continuous-time problem with discount factor η.

The result for the average cost criterion is obtained by referring to the following Derman's theorem [15].

THEOREM *Derman* *If a policy $\mathcal{P}^*$ is optimal among the class of policies Π for all discounted problems with discount factor α close to 1, then $\mathcal{P}^*$ is also optimal among all policies in the class Π under the average reward criterion.*

Thus, the following proposition can be formulated:

Proposition 6. *CPP is optimal under the average cost criterion, for the infinite horizon problem, when $M_H = 0$.*

The proof that, if $M_H = 0$, the optimal policy is CPP, dramatically decreases the feasible region of the optimization problem which is reduced to the only search for the optimal cutoff value T. This allows a relevant reduction of the number of iterations for the solution with the most common algorithms like the simplex or the policy improvement.

3. QoS Parameters and Numerical Results

From the customer point of view, it may be interesting to calculate some QoS parameters such as the probability that a new call attempt is blocked or the probability that a call, once accepted, is terminated before completion. We evaluate these probabilities and other QoS parameters for an arbitrary call in the network, on the basis of the traffic model described in Section 2, under the application of CPP. If v denotes the average speed of a mobile, and D is the diameter of the cell, the mean time that the mobile spends in the cell is $1/(\mu_2) = D/v$, while if L denotes the diameter of the HR, the mean residence time is given by $1/(\mu_3) = L/v$.

The steady state probability $P(k)$ of the Markov process under CPP with cutoff value T, can be derived by the solution of the linear programming problem formulated through Equation (9), or may be simply calculated through the following balance equations.

$$\begin{cases} k(\mu_1+\mu_2)P(k) = [\lambda_L + (k-1)\lambda_H]P(k-1) \\ \qquad\qquad \text{if } 1 \leq k \leq T \\ k(\mu_1+\mu_2)P(k) = \lambda_H(k-1)P(k-1) \\ \qquad\qquad \text{if } T < k \leq C \\ [C(\mu_1+\mu_2) + (k-C)(\mu_1+\mu_3)]P(k) \\ \qquad\qquad = C\lambda_H P(k-1) \\ \qquad\qquad \text{if } C < k \leq C + M_H \\ \displaystyle\sum_{k=0}^{C+M_H} P(k) = 1 \end{cases} \tag{20}$$

The solution is

- if $1 \leq k \leq T$

$$P(k) = \frac{\Pi_{i=0}^{k-1}(\lambda_L + i\lambda_H)}{(\mu_1+\mu_2)^k k!}P(0)$$

- if $T < k \leq C$

$$P(k) = \frac{\lambda_H^{k-T}\Pi_{i=0}^{k-1}(\lambda_L + i\lambda_H)}{k(T-1)!(\mu_1+\mu_2)^k}P(0)$$

- if $C < k \leq C + M_H$

$$P(k) = \frac{C^{k-C-1}\lambda_H^{k-T}\Pi_{i=0}^{T-1}(\lambda_L + i\lambda_H)}{(\mu_1+\mu_2)^C \; \Pi_{j=1}^{k-C}[C(\mu_1+\mu_2)j(\mu_1+\mu_3)](T-1)!}$$

where $P(0)$ is determined from the normalization condition $\sum_{k=0}^{C+M_H} P(k) = 1$.

Once the steady state probabilities have been calculated, the most important QoS parameters can be computed. For example, the probability B_L that an initial access is blocked is

$$B_L = \sum_{i=T}^{C+M_H} P(i) \tag{21}$$

while the probability B_H that a hand-off call is blocked is equal to the probability that no place is available in the queueing device, that is, $B_H = P(C + M_H)$.

The mean queue length is

$$L_H = \sum_{i=C+1}^{C+M_H} (i-C)P(i) \tag{22}$$

We next find the conditioned probability $B_{H\ \overline{out}}$ that a queued hand-off call escapes from the queue before being served. It is given by the fraction of the hand-off calls that cannot get channels while waiting in the hand-off area:

$$B_{H\ \overline{out}} = \mu_3 \frac{L_H}{(1 - B_H)C\lambda_H} \tag{23}$$

while the term $E_H = \mu_3 L_H$ is the unconditioned probability to see a hand-off call escaping from the system before obtaining a channel.

CPP represents a trade-off solution between the optimization of the loss probabilities of the two streams of arriving requests. Another policy may have a better behavior towards the single class of customers, but at the expense of the quality of service of the requests of the other class. Numerical results confirm the optimality of CPP even when hand-off queueing is allowed.

The behavior of CPP with variable cutoff value T is now compared with that of the *hysteresis policy* (HysP) [2, 8]. Under the hysteresis policy (HysP), a hand-off call is accepted as a channel is free, but the decision to accept or not an initial access request is taken on the basis of the number of free channels following a cycle of hysteresis with thresholds M and M', where $M \leq M'$. HysP can be studied as a particular instance of the general model we proposed, and the numerical results confirm the best behavior of CPP. In Figures 5 and 6, we set $M = T$, and $M' = C$, for a system with $C = 10$ available channels, $\lambda_L = 10$, $C\lambda_H = 10$, $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 10$, $M_H = 5$, $L = 5$ and $H = 1500$.

The trend of the loss probability with the number of guard channels $(C - T)$ is shown.

For the above described system, the average cost function is optimal under the application of CPP with cutoff value $T = 8$, that is with two guard channels. The graphs show that increasing the number of guard channels, the advantage of a lower loss probability of the high priority stream corresponds to the disadvantage of a greater loss probability for the low priority stream. If the cutoff value is $T = 8$, a tradeoff solution is achieved, and the loss of new calls is compensated by the gain in terms of successful high priority hand-offs.

Figure 7 shows how the optimal number of guard channels increases with the average rate of high priority requests λ_H for a system with $C = 10$ available channels, $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 10$, $M_H = 5$ and where the penalties are $L = 5$ and $H = 150$. This shows how the system becomes more selective in accepting potentially unprofitable customers, if the arrival rate of requests grows. In Figure 8 we can see that the more unprofitable the new call stream is considered (i.e. the higher priority is given to the hand-off stream), the higher the number of guard channels there will be.

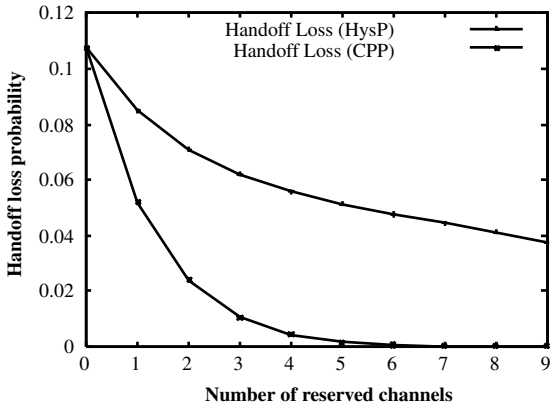

Fig. 5. Hand-off loss probability.

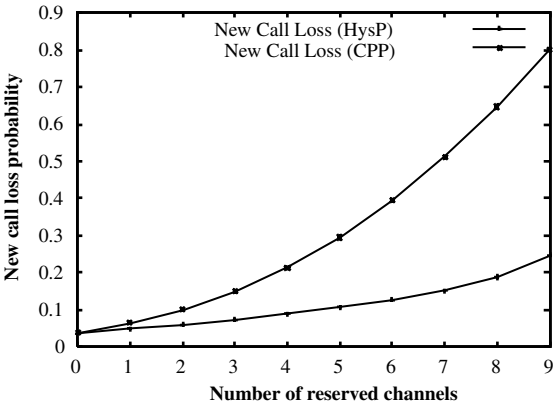

Fig. 6. Initial access loss probability.

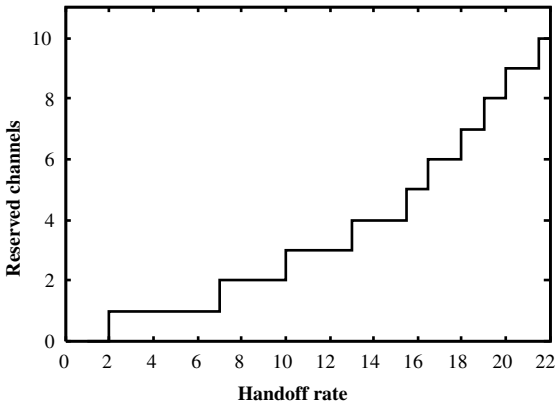

Fig. 7. Guard channel trend with hand-off rate.

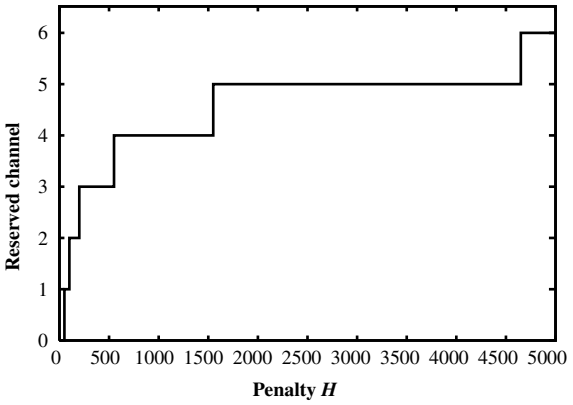

Fig. 8. Guard channel trend with penalty H.

4. Conclusions

This paper proposes an optimization method of call admission control which takes into account the dependency of the hand-off rates on the average occupancy level of the cells. The model is based on a cost function which gives higher priority to hand-off requests than to originating calls.

This cost function has been studied and optimized through a Markov decision model characterized by a great generality. The proposed model is shown to be able to represent both not stationary policies and randomized fractional policies. Moreover, owing to the particular shape of its transition diagram, it becomes possible to study key policies such as the threshold policy and algorithms with one or more cycles of hysteresis. It is analytically proven that if the objective function is the total discounted cost function, or the average cost function applied to the infinite horizon problem, the policy CPP is optimal, when no queueing of requests is allowed. Numerical results confirm the optimality of CPP even when hand-off queueing is allowed.

References

1. Gavish B, Shridar S. Threshold priority policy for channels assignment in cellular networks. *IEEE Transactions on Computers* 1997; **46**(3).
2. McMillan D. Delay analysis of a cellular mobile priority queueing system. *IEEE/ACM Transactions on Networking* 1995; **3**(3).
3. Ramjeee R, Nagarajan R, Towsley D. On optimal call admission control in cellular networks. *Proceedings of IEEE INFOCOM '96 Conference*, March 1996.
4. Guerin R. Queueing-blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications* 1988; **36**(2).
5. Hong D, Rappaport SS. Priority oriented channel access for cellular systems serving vehicular and portable radio telephones. *IEEE Proceedings*, Vol. 136, No. 5, October 1989.
6. Zeng Q, Mukumoto K, Fukuda A. Performance analysis of mobile cellular radio systems with priority reservation hand-off procedures. *IEEE 0-7803-1927-3/94*.
7. Bartolini N, Chlamtac I. Call admission control: solution of a general decision model with state related hand-off rate. *Proceedings of IEEE Wireless Communications and Networking Conference*, WCNC'00, Chicago, 23–28 September 2000.
8. Scherer R. On a cutoff priority queueing system with hysteresis and unlimited waiting room. *Computer Networks and ISDN Systems* 1990; **20**.
9. Bharucha-Reid A. *Elements of the Theory of Markov Processes and their Applications*. McGraw-Hill: 1960.
10. Keilson J. *Markov Chain Models. Rarity and Exponentiality*. Springer-Verlag: New York, 1979.
11. Ross S. *Applied Probability Models with Optimization Applications*. Holden–Day: 1970.
12. Heyman DP, Sobel MJ. *Stochastic Models in Operations Research*. Vols 1 and 2. McGraw-Hill: 1984.
13. Hastings K. *Introduction to the Mathematics of Operations Research*. Dekker: 1989.
14. Tijms HC. *Stochastic Modeling and Analysis. A Computational Approach*. John Wiley & Sons: 1986.
15. Derman C. *Finite State Markovian Decision Processes*. Academic Press: 1970.

Authors' Biographies

Novella Bartolini graduated with honors in 1997 and received her Ph.D. degree in computer engineering in 2001 from the University of Rome, Italy. She is now a researcher at the University of Rome. She worked as a researcher at Fondazione Ugo Bordoni, and in 1999 and 2000 was a visiting scholar at the CATSS center, University of Texas at Dallas, U.S.A. Her research interests lie in the area of computer networks and parallel computing.

Imrich Chlamtac received the B. Sc. and M. Sc. degrees in mathematics with the highest distinction, and the Ph. D. degree in computer science from the University of Minnesota in 1979. He is a Professor of Electrical Engineering and holds the Distinguished Chair in Telecommunications at the University of Texas at Dallas, U.S.A. He is also a member of the Photonics Center at Boston University, and President of BCN Inc., a company dealing with network design, integration, and technology transfer in wireless data, and high-speed communications jointly with Boston University. His research interests include research and implementation aspects of mobile networks and wireless communication systems, optical networks, ATM, and multimedia communication. He has published close to 200 papers in refereed journals and conferences. He is the author of more than 30 invited papers, multiple book chapters, and encyclopedias. In 1981, he coauthored the first textbook on LANs entitled 'Local Networks: Motivation, Technology and Performance'.

Dr. Chlamtac serves as the founding Editor in Chief of the *ACM/Baltzer Wireless Networks (WINET)* and the

Journal on Special Topics in Mobile Networks an Application (MONET). He served on the Editorial Boards of *IEEE Transactions on Communications, Computer Networks*, and *ISDN Systems, High Speed Networks Journal, Telecommunications Systems* and the *Photonic Network Communications Journal*. He was Guest Editor for the *Proceeding of the IEEE*, the *IEEE Journal on Selected Areas on Communications, IEEE Transactions on Computers*, and other journals.

He served as the General Chair of leading ACM and IEEE conferences and workshops, including ACM Sigcomm, ACM/IEEE MOBICOM, and IEE CCW, and acts as the ACM/IEEE MOBICOM Steering Committee Chair. He is the Chairman of ACM SIGMOBILE, the Special Interest Group on Mobile Computing and Networking. He is a fellow of the ACM, winner of the Society of Computer Simulation Award, and ACM Best Paper Award. In the past, he was an IEEE, Northern Telecom, and BNR Distinguished Lecturer, and a plenary and keynote speaker at leading conferences. He was a Fulbright Scholar and is honorary Member of the Senate at the Technical University of Budapest.